Analyse des génomes à la recherche de répétitions en tandem polymorphes : outils d?épidémiologie bactérienne et locus hypermutables humains

France Denoeud

UNIVERSITÉ PARIS XI
UFR SCIENTIFIQUE D'ORSAY

THÈSE

Présentée pour obtenir

Le GRADE de DOCTEUR EN SCIENCES
DE L'UNIVERSITÉ PARIS XI ORSAY

PAR

France DENOEUD

Sujet :

Analyse des génomes à la recherche de répétitions en tandem polymorphes : outils d'épidémiologie bactérienne et locus hypermutables humains

Soutenue le 1er décembre 2003 devant la commission d'examen :

Me Marie-France Sagot	Rapporteur
M Jean Weissenbach	Rapporteur
M Patrick Forterre	Président du jury
M Gilles Vergnaud	Examinateur
M Alain Nicolas	Directeur de thèse

Le chemin le plus court d'un point à un autre c'est de ne pas y aller.
Le Chat (Philippe Geluck).

Je remercie Madame Marie-France Sagot et Monsieur Jean Weissenbach, mes rapporteurs, ainsi que Monsieur Patrick Forterre, président du jury, de m'avoir fait l'honneur de juger ce travail.

Je remercie Alain Nicolas, mon directeur de thèse, de m'avoir permis d'effectuer cette thèse dans les meilleures conditions.

Je tiens à remercier Gilles Vergnaud de m'avoir accueillie dans son laboratoire et encadrée pendant ces quatre années. Merci surtout pour la confiance qu'il m'a accordée.

J'adresse mes remerciements à Gary Benson, qui a été un collaborateur disponible et attentif.

Je remercie l'équipe GPMS pour son accueil, et en particulier les utilisateurs de la base de données : rien n'est plus valorisant que de développer un outil « qui sert vraiment ». Merci de m'avoir permis de contribuer à tous ces projets, et d'avoir veillé avec moi à la maintenance et à l'amélioration du site web.

Merci à tous ceux que j'ai côtoyés à la cantine et à la cafet' : Lucie, Philippe, Sophie, et les autres. En effet, « c'est bien d'avoir des collègues », mais surtout pour parler d'autre chose que de travail ! Vous allez me manquer.

Enfin, en dehors du labo, mais jamais vraiment loin grâce à la magie de l'e-mail, je remercie mes amis (Hélène, Charlotte, Etienne et « la bande »…) et ma famille (Jean-Maxime, Lucile, Lise…) pour avoir grandement contribué à ma bonne humeur au quotidien.

Et encore un grand merci pour votre aide précieuse lors de la relecture et de l'impression du manuscrit comme de la préparation de la soutenance : Lucie, Hélène, Lise, JM… et à Lucile pour le pot de thèse !

Table des matières

Liste des abréviations

ADN : acide désoxyribonucléique

AFLP : amplified fragment length polymorphism

ARN : acide ribonucléique

ARNm : ARN messager

ASP: active server pages

BAC: bacterial artificial chromosome

BLAST: basic local alignment search tool

CDS : coding sequence

CEPH : centre d'étude du polymorphisme humain

CRISPR : clustered regularly interspaced short palindromic repeat

EST : expressed sequence tag

GOLD : genomes online database

HGP : human genome project

HS : hierarchical shotgun

HSP : high scoring segment pair

IP : Internet protocol

IS : insertion sequence

kb : kilobase

LINE : long interspersed nucleotide element

LPS : lipopolysaccharide

LTR : long terminal repeat

Mb : mégabase

MIRUs: mycobacterial interspersed repetitive units

MLST : multilocus sequence typing

MLVA: multilocus VNTR analysis

MMR : mismatch repair

MPTR : multi-period tandem repeat

MSI : microsatellite instability

MSP: maximal segment pair

MVR-PCR : minisatellite variant repeat - polymerase chain reaction

NCBI : national center for biotechnology information

ORF : open reading frame

pb : paire de bases

PCR : polymerase chain reaction

PIC : polymorphism information content

RAPD : random amplified polymorphic DNA

RFLP : restriction fragment length polymorphism

SINE : short interspersed nucleotide element

SNP : single nucleotide polymorphism

SSM : slipped-strand mispairing

SSR : short sequence repeat

STR : short tandem repeat

TIGR : the institute for genomic research

TRDB: tandem repeats database

TRF : tandem repeats finder

UTR : untranslated region

VLTR : variable length tandem repeat

VNTR : variable number of tandem repeats

WGS : whole genome shotgun

Avant-propos

Le séquençage des génomes, qui a connu un essor très important depuis quelques années, est à l'origine d'une nouvelle discipline : la génomique. Le séquençage du génome humain, dont l'achèvement a été annoncé dans les médias au printemps 2001 (même s'il ne s'agissait que de la séquence brouillon), ouvre la voie à de nouvelles stratégies pour la recherche bio-médicale. Cependant, ce « décryptage » n'est pas suffisant pour comprendre le fonctionnement des cellules humaines : d'une part, 40% des gènes prédits ont une fonction inconnue, et d'autre part, même si la fonction de tous les gènes était inférée, il faudrait encore être en mesure de prévoir leur expression différentielle dans les tissus et dans le temps. Moins médiatisé, le séquençage de génomes bactériens (en particulier des pathogènes humains) a énormément bénéficié des progrès technologiques induits par le projet « génome humain ». Actuellement, plus de 120 génomes bactériens sont disponibles et près de 400 sont en cours de séquençage. Chez ces organismes, la comparaison avec les banques de séquences permet d'inférer une fonction pour seulement 50% des gènes prédits. Afin d'attribuer une fonction aux gènes inconnus, on peut faire appel à la génomique fonctionnelle. En effet, chez les organismes unicellulaires, on peut espérer induire un phénotype mutant après l'inactivation ciblée d'un gène, ce qui est beaucoup plus délicat pour les métazoaires. Des projets de génomique fonctionnelle à grande échelle, impliquant la collaboration internationale entre de nombreux laboratoires, ont été initiés depuis quelques années, par exemple pour la bactérie modèle *Bacillus subtilis.*

Les séquences codantes, qui représentent une proportion très variable des génomes (moins de 5% pour les eucaryotes à plus de 95% pour certains procaryotes), pouvaient déjà être étudiées avant l'avènement des projets de séquençage systématique. Par exemple, des banques d'EST (« Expressed Sequence Tags » : séquences d'ADNs complémentaires, correspondant à des gènes exprimés) avaient été constituées pour de nombreux organismes. Le séquençage de génomes complets permet dorénavant d'accéder à d'autres éléments essentiels, comme les locus polymorphes, points chauds d'évolution. Les répétitions en tandem sont l'une des sources les plus importantes de variabilité dans les génomes. Ces séquences particulières, constituées de répétitions successives de motifs nucléiques, sont présentes dans l'ensemble du monde vivant, et peuvent être situées dans les gènes comme dans les régions intergéniques. Même si leur fonction biologique reste à élucider, les répétitions en tandem ont, grâce à leur polymorphisme, des applications dans de nombreux domaines. Tout d'abord chez les bactéries, les répétitions en tandem polymorphes, dont le nombre d'unités varie d'une souche à l'autre, se révèlent un outil puissant pour le génotypage de souches à des fins épidémiologiques. D'autre part, certaines répétitions en tandem humaines, notamment les minisatellites (classe de taille particulière), ont la propriété d'être extrêmement instables

d'une génération à l'autre : les minisatellites hypermutables sont les éléments les plus instables du génome humain. Ils peuvent être utilisés comme biomarqueurs d'exposition à des agents potentiellement mutagènes tels que les radiations ionisantes. Par ailleurs, d'un point de vue plus fondamental, ils sont un modèle d'étude de certains mécanismes d'instabilité des génomes.

A partir des nombreuses séquences génomiques disponibles, il faut être en mesure d'identifier les répétitions en tandem d'intérêt : marqueurs polymorphes chez les bactéries, ou minisatellites hypermutables chez l'homme. Cette tâche peut être accomplie grâce à une discipline fortement associée à la génomique : la bioinformatique. En effet, il est impensable d'espérer traiter les énormes quantités de données de séquençage générées quotidiennement sans outils informatiques efficaces. Ainsi, des programmes performants permettent l'identification des répétitions en tandem dans les séquences d'ADN. Cependant, lorsque j'ai commencé cette thèse, aucun utilitaire ne facilitait l'accès aux données sur les répétitions en tandem de génomes entiers. C'est pourquoi nous avons décidé de développer une base de données des répétitions en tandem, d'accès public (http://minisatellites.u-psud.fr), mise à jour régulièrement afin de mettre à disposition les répétitions en tandem de tous les génomes séquencés. Je présenterai dans cette thèse l'élaboration de la base de données, ainsi que différentes applications que nous en avons faites au laboratoire « Génomes Polymorphisme et Minisatellites », pour le génotypage de souches de bactéries pathogènes, comme pour l'identification de minisatellites hypermutables humains. Nous traiterons notamment d'un point qui n'avait encore jamais été abordé : la prédiction du polymorphisme des répétitions en tandem, qui constituerait une avancée majeure, tant à des fins pratiques (économies de typages) que théoriques (meilleure compréhension des mécanismes de mutation).

1 Introduction

1.1 Le séquençage des génomes

1.1.1 Les génomes procaryotes

1.1.1.1 Génomes de bactéries : la situation actuelle

La publication de la première séquence complète d'un génome bactérien, celui d'*Haemophilus influenzae*, en 1995 (Fleischmann 1995), a démontré l'efficacité de l'approche « WGS » (whole genome shotgun) pour le séquençage de génomes complets (voir Figure 4), et les progrès spectaculaires qui ont été faits au niveau des techniques de séquençage, des stratégies d'assemblage et de finition, et des méthodes d'annotation. La microbiologie est certainement parmi les premiers bénéficiaires de cette évolution. A l'heure actuelle, plus de 100 génomes bactériens ont été entièrement séquencés et trois fois plus sont en cours de séquençage (au 5 août 2003, d'après le site GOLD, « Genome Online Database » [http://ergo.integratedgenomics.com/GOLD/] (Bernal 2001), 114 génomes bactériens étaient achevés et 355 en cours). La Figure 1 montre l'évolution du nombre de génomes procaryotes séquencés chaque année depuis 1995, ce qui témoigne des progrès effectués ces dernières années pour le séquençage systématique des génomes.

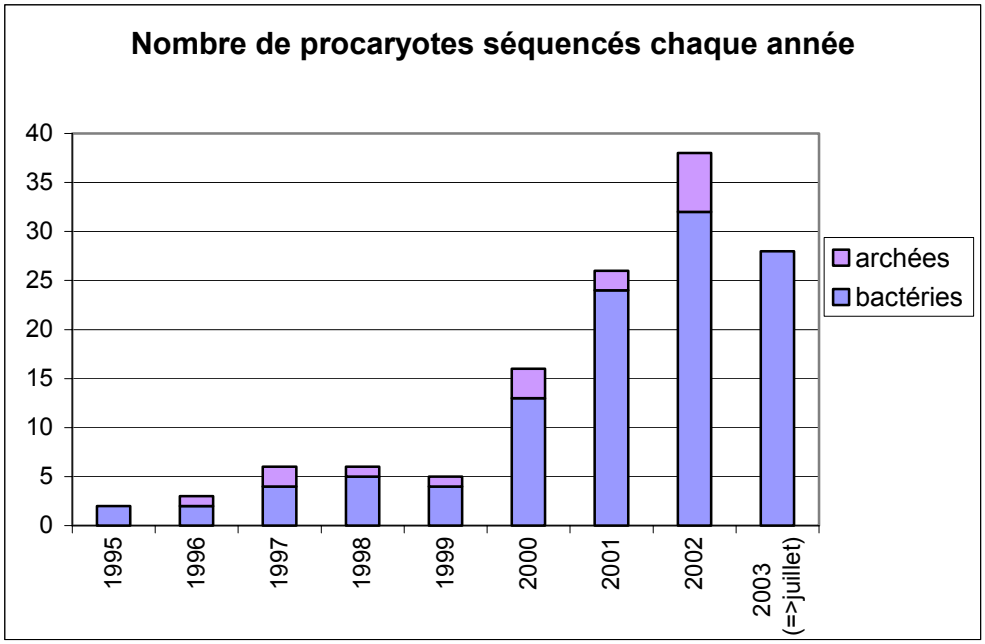


Figure 1 : Evolution du nombre de génomes séquencés chaque année depuis 1995.

Actuellement, le séquençage complet d'un génome bactérien peut être achevé en quelques mois, avec un taux d'erreur de l'ordre de 1/100000 seulement, et à un coût de 7 à 8 centimes d'euros par paire de bases (Fraser 2002), ce qui correspond, pour un génome de quelques mégabases à un coût de quelques centaines de milliers d'euros.

La Figure 2 décrit les propriétés des différentes bactéries séquencées. On peut les classer dans trois grandes catégories, correspondant aux motivations de leur séquençage :

- Les bactéries d'intérêt médical : pathogènes humains (*Yersinia pestis :* peste, deux souches ; *Mycobacterium tuberculosis* : tuberculose, 3 souches ; *Staphylococcus aureus* : infections nosocomiales, 6 souches) et bactéries présentant un intérêt pharmaceutique (*Streptomyces coelicolor :* production d'antibiotiques).
- Les bactéries d'intérêt économique : pathogènes agroalimentaires (*Xylella fastidiosa* : phytopathogène), bactéries présentant un intérêt pour l'industrie agroalimentaire (*Lactococcus lactis*), bactéries utilisées pour la synthèse d'acides aminés (*Corynebacterium efficiens*) ou de solvants (*Clostridium acetobutylicum*), ou encore pour la dépollution (*Pseudomonas putida*, *Shewanella oneidensis*).
- Les bactéries d'intérêt pour la recherche en microbiologie : organismes modèles (*Escherichia coli*, *Bacillus subtilis*) ou ayant des propriétés biologiques intéressantes (résistance à des conditions de vie en milieux « extrêmes » : par exemple, *Deinococcus radiodurans* qui survit dans des milieux fortement irradiés, ou *Thermotoga maritima*, organisme thermophile).

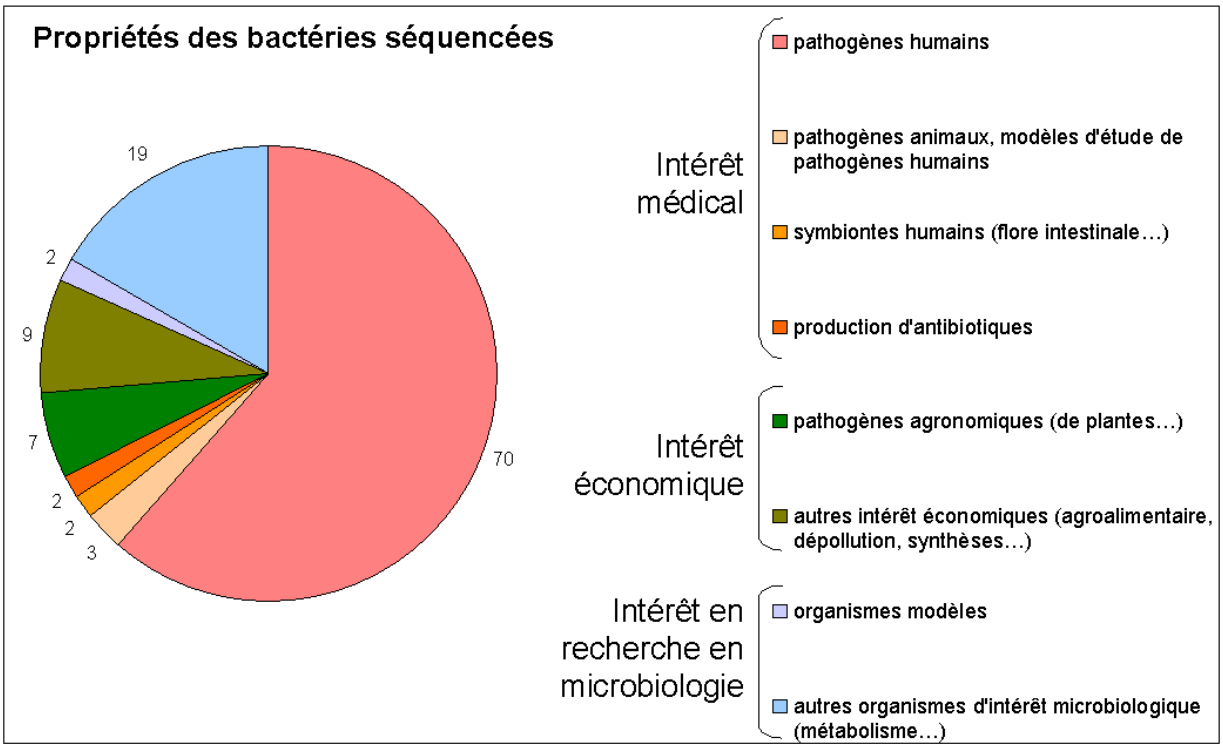


Figure 2 : Motivations du séquençage de génomes bactériens.

On constate que la grande majorité des bactéries séquencées à ce jour est d'intérêt médical : entre 1995 et 2000, les organismes séquencés étaient essentiellement des bactéries pathogènes pour l'homme auxquelles s'ajoutaient quelques organismes d'intérêt plus fondamental (organismes modèles). Depuis 2001, un nombre croissant de projets de séquençage implique

des bactéries d'intérêt économique (Nelson 2000). Le biais en faveur des bactéries d'intérêt médical devrait donc s'atténuer dans les prochaines années.

Le règne bactérien est d'une grande hétérogénéité, du point de vue du pourcentage en GC (de 22% pour *Wigglesworthia glossinidia* à 72% pour *Streptomyces coelicolor*) comme de la taille du génome (de 580 kb pour *Mycoplasma genitalium*, pathogène intracellulaire obligatoire, à 9105 kb pour *Bradyrhizobium japonicum*) ou du nombre de gènes. La Figure 3 représente le nombre de gènes en fonction de la taille du génome pour les 114 bactéries séquencées au 5 août 2003. La densité en gènes des bactéries est relativement constante et voisine de 1 gène par kilobase (elle varie entre 0,49 pour *Mycobacterium leprae*, particulièrement peu dense : cette espèce présente une fraction importante d'ADN non codant et de pseudogènes –non transcrits ou non traduits- (Cole 2001), et 1,29 pour *Escherichia coli* O157:H7 EDL933). La fraction codante est généralement de l'ordre de 90%. Près de la moitié des ORFs (pour « Open Reading Frames », ou phases ouvertes de lecture) de chaque espèce est de fonction inconnue. De plus, environ un quart des ORFs n'a aucune homologie avec des protéines existantes dans les bases de données de séquences. Ce pourcentage devrait diminuer avec le séquençage de davantage de génomes bactériens, mais il témoigne de la grande diversité biologique au sein des organismes procaryotes (Fraser 2000).

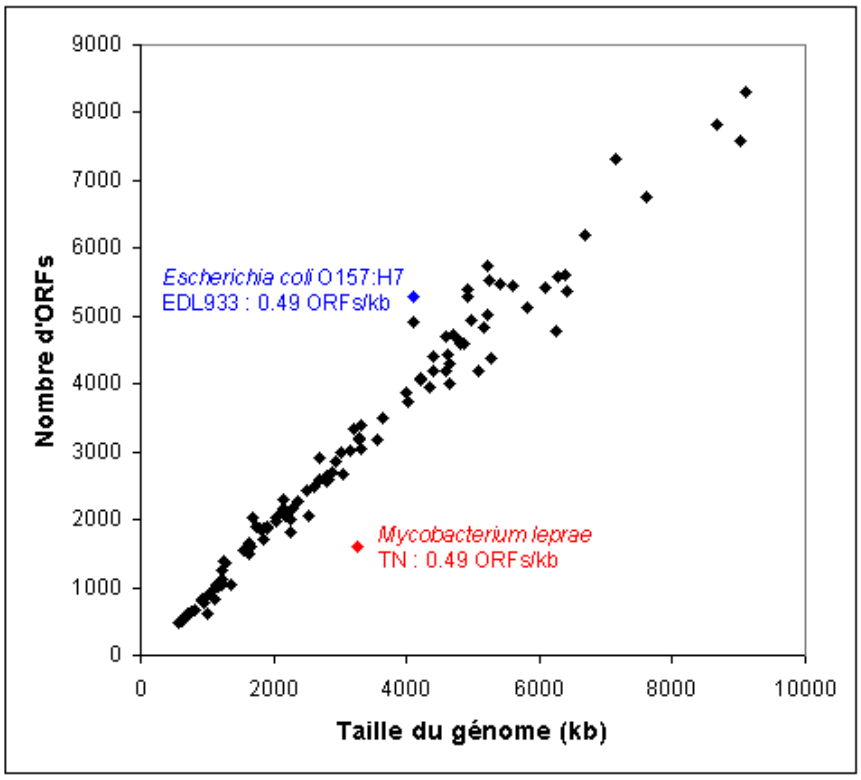


Figure 3 : Nombre de séquences codantes en fonction de la taille des génomes bactériens.

1.1.1.2 Applications du séquençage des génomes bactériens

Le séquençage de pathogènes humains devrait permettre des avancées dans le domaine médical, pour le diagnostic (développement de tests plus rapides) ainsi que l'élaboration de nouveaux vaccins et de nouveaux agents anti-microbiens, dont le besoin est grandissant compte-tenu de la propagation des résistances aux antibiotiques (Fraser 2000). Par exemple, Pizza *et al.* ont identifié de nouveaux candidats de vaccins contre les souches de *Neisseria meningitidis* du sérogroupe B, par une approche basée sur la génomique (Pizza 2000). Les protéines putatives de la surface cellulaire ou sécrétées ont été identifiées à partir de la séquence complète d'une souche du sérogroupe B (MC58). La majorité a pu être exprimée dans *E. coli* et utilisée pour immuniser des souris : parmi 7 protéines à l'origine d'une liaison des anticorps à la surface des méningocoques et d'une activité bactéricide, deux se sont révélées très bien conservées entre différentes souches de *N. meningitidis.* Ces deux protéines sont des vaccins potentiels. Cette étude est la première à avoir tiré parti du séquençage des génomes afin d'accélérer le développement de vaccins contre des organismes pathogènes. D'autres applications de l'analyse des génomes pour l'identification de vaccins sont présentées dans la revue (Zagursky 2003).

D'autre part, le séquençage de génomes entiers de pathogènes a révélé l'existence de mécanismes permettant de générer une variation antigénique au niveau des protéines de la surface cellulaire (Fraser 2000). Ces mécanismes sont de trois types :

- Glissement de la réplication au niveau de répétitions en tandem situées au voisinage de gènes ou dans les régions codantes : ce phénomène a été décrit pour *H. influenzae* (Fleischmann 1995), *H. pylori* (Tomb 1997) et *M. tuberculosis* (Cole 1998). Il sera évoqué plus en détail au paragraphe 1.2.2.1.
- Recombinaison entre des gènes homologues codant pour des protéines de surface : ce phénomène a été décrit chez les mycoplasmes *M. genitalium* et *M. pneumoniae* (Fraser 1995 ; Himmelreich 1996) et chez *Treponema pallidum* (Fraser 1998).
- Variabilité clonale au niveau des protéines de surface, semblable à celle qui est observée chez le parasite responsable de la malaria, *Plasmodium falciparum* (Gardner 1998) et qui semble également s'appliquer au spirochète *Borrelia burgdorferi* (Fraser 1997).

Enfin, le séquençage de génomes est l'approche la plus puissante pour identifier une variabilité génomique au sein d'espèces bactériennes, jusqu'au niveau de la souche ou même de l'isolat. A la suite de l'attaque bioterroriste d'octobre 2001 qui a disséminé par voie postale une poudre de Bacille du charbon (*Bacillus anthracis*), la séquence de la souche des enveloppes (dite « Floride ») a été comparée à la souche d'origine (dite « Ames ») afin de mettre en évidence les locus polymorphes entre ces souches et de tenter d'identifier le laboratoire d'où provient la souche « Floride » (Read 2002). De telles comparaisons de génomes sont grandement facilitées lorsque au moins l'un des génomes est entièrement séquencé (et non-pas sous forme de fragments), ce qui est le cas pour un nombre croissant de

bactéries d'intérêt médical ou économique. Par exemple, 6 souches de *Staphylococcus aureus* sont actuellement disponibles. Par ailleurs, la comparaison de génomes d'espèces proches mais causant des maladies très différentes, telles que *M. leprae* et *M. tuberculosis*, ou *N. meningitidis* et *N. gonorrhoeae*, devrait aider à l'identification des gènes responsables de tel ou tel effet pathogène.

1.1.1.3 Génomes d'archées

Tableau 1 : Description des génomes d'archées séquencés.

Espèce	souche	propriétés de l'archée	numéro(s) d'accession *	Date de publication *	taille du génome (kb)*	nombre d'ORFs (phases ouvertes de lecture)*	densité en ORFs (nombre d'ORFs par kb)	Pourcentage d'ORFS inconnues **	%GC**	fraction codante**
Aeropyrum pernix	K1	hyperthermophile aérobie	BA000002	30/04/1999	1669	2620	1,57	57%	56%	89%
Archaeoglobus fulgidus	DSM4304	hyperthermophile, réduit le sulfate	AE000782	27/11/1997	2178	2493	1,14	53%	48,5%	92%
Halobacterium sp	NRC-1	halophile	AE004437 AE004438 NC_001869	24/10/2000	2014	2058	1,02	60%	68%	nd
Methano-bacterium thermo-autotrophicum	delta H	méthanogène thermophile	AE000666	10/11/1997	1751	1918	1,10			
Methanococcus jannaschii	DSM 2661	méthanogène thermophile	L77117 NC_001732 NC_001733	28/09/1996	1664	1750	1,05	62%	31%	88%
Methanopyrus kandleri	AV19	méthanogène hyperthermophile	AE009439	02/04/2002	1694	1691	1,00			
Methanosarcina mazei	Go1 (DSMZ 3647)	méthanogène thermophile	AE008384	10/07/2002	4096	3371	0,82			
Methanosarcina acetivorans	C2A	méthanogène	AE010299	10/04/2002	5751	4540	0,79	51%	42%	74%
Pyrobaculum aerophilum	IM2	hyperthermophile aérophile	AE009441	22/01/2002	2222	2587	1,16			
Pyrococcus abyssi	GE5	hyperthermophile	AL096836	13/02/2002	1765	1765	1,00	49%	45%	91%
Pyrococcus furiosus	DSM 3638	hyperthermophile	AE009950	12/02/2002	1908	2065	1,08			
Pyrococcus horikoshii	*shinkaj OT3*	hyperthermophile	BA000001	30/04/1998	1738	1979	1,14	42%	42%	91%
Sulfolobus solfataricus	P2	thermoacidophile aérobie	AE006641	03/07/2001	2992	2977	0,99			
Sulfolobus tokodaii	7	thermoacidophile aérobie	BA000023	31/08/2001	2694	2826	1,05			
Thermoplasma acidophilum	DSM 1728	thermophile, acidophile	AL139299	28/09/2000	1564	1478	0,95	45%	46%	87%
Thermoplasma volcanium	GSS1	thermophile, hétérotrophe	BA000011	19/12/2000	1584	1524	0,96			

* D'après le site « GOLD » [http://ergo.integratedgenomics.com/GOLD/]

** D'après le site du Génoscope [http://www.genoscope.cns.fr/externe/Francais/Sequencage/]

Les 16 génomes d'archées dont le séquençage était achevé au 5 août 2003 sont présentés dans le Tableau 1. Comme les bactéries, les archées sont hétérogènes (en ce qui concerne le pourcentage en GC, la taille des génomes, le nombre de gènes…). Plus de 50% des ORFs n'ont pas de fonction connue, ce qui peut refléter le fait que « seulement » 16 archées ont été séquencées à ce jour. Certaines espèces proches ont été séquencées, par exemple les *Pyrococcus* : la comparaison de ces trois génomes a mis en évidence des mécanismes d'instabilité chromosomique (Zivanovic 2002).

1.1.2 Les génomes eucaryotes

1.1.2.1 Le génome humain

1.1.2.1.1 Le Projet Génome Humain

Le Projet Génome Humain (HGP : « Human Genome Project ») est né de deux idées qui ont émergé au début des années 80 : d'une part, qu'une vue d'ensemble sur les génomes pourrait accélérer grandement la recherche biomédicale en permettant aux chercheurs d'appréhender les problèmes de façon plus éclairée et plus efficace, et d'autre part, que les données produites devraient être mises à disposition de l'humanité entière et que leur dissémination ne devrait pas être limitée par des intérêts privés ou nationaux. Ces deux notions ont eu comme corollaire la prise de conscience que l'obtention de ces données devrait se faire au travers d'efforts collaboratifs internationaux, sans précédent dans le domaine de la biologie. Cet état d'esprit d'une communauté consciente de l'ampleur de la tâche que peut représenter l'étude du génome humain avait d'ailleurs déjà conduit à des initiatives, dont l'une des plus connues est probablement le CEPH (Centre d'Etudes du Polymorphisme Humain), fondé par le Français Jean Dausset. Le CEPH a joué un rôle fondamental dans l'organisation du travail de cartographie génétique humaine de plusieurs dizaines de laboratoires, souvent concurrents, au cours des années 80. L'idée de séquencer le génome humain a été évoquée pour la première fois dans des congrès organisés par le Département à l'énergie américain entre 1984 et 1986. Le Projet Génome Humain (en l'occurrence la phase de séquençage proprement dite) a été lancé en 1990, sur l'initiative de 4 pays (USA, Grande-Bretagne, Japon et France), qui ont chacun fondé un centre de séquençage. D'autres pays se sont ensuite joints au projet, comme l'Allemagne et la Chine : au final, plus de 20 laboratoires s'y sont impliqués. Jusqu'à 1995, le projet a progressé sur deux voies principales : la construction des cartes génétique et physique chez l'homme et la souris d'une part, et le séquençage de la levure *S. cerevisiae* et du nématode *C. elegans* ainsi que de certaines régions de génomes mammifères d'autre part. Cette première phase a prouvé que le séquençage à grande échelle était envisageable, par l'approche en deux étapes consistant tout d'abord à générer les séquences de fragments couvrant le génome de façon fortement redondante (8 à 10 fois), puis à effectuer l'assemblage et la finition c'est-à-dire à boucher les trous (« gaps ») et résoudre les ambiguïtés. Cette phase

du projet a également permis de montrer que les séquences complètes fournissent des informations sur les gènes, les régions régulatrices, et la structure des chromosomes, qui ne sont pas accessibles par l'étude exclusive des ADNs complémentaires. En 1995, l'idée de produire un brouillon du génome humain (« draft genome sequence ») a été proposée, mais elle était prématurée car l'efficacité du séquençage à grande échelle d'un génome complexe et riche en répétitions tel que le génome humain n'avait pas encore été prouvée. Des projets pilotes ont donc été lancés, afin de démontrer la faisabilité d'un tel projet, avec échéance en mars 1999 : ils ont produit avec succès des séquences de bonne qualité (99.99%) et sans « gaps », représentant 15% du génome humain tout en permettant de mettre au point des stratégies cohérentes pour le séquençage. En mars 1999, le projet de séquençage à grande échelle du génome humain a donc débuté, avec pour but initial de produire un « brouillon » de la séquence génome humain couvrant cette fois la majeure partie du génome, avant juin 2000. L'article paru en février 2001 (Lander 2001) décrit cette première version de la séquence du génome humain. Depuis, la phase de finition a été amorcée, en particulier avec le séquençage de clones permettant l'élimination des gaps. L'accomplissement de cette tâche a été annoncé en avril 2003, pour le cinquantenaire de la découverte de l'ADN (voir *Science* du 11 avril 2003 et *Nature* du 24 avril 2003). Tout au long du projet, les séquences ont été régulièrement mises à jour et ont toujours été d'accès libre. De plus, l'analyse des séquences de chromosomes complets a été publiée au fur et à mesure de leur achèvement. Ainsi, les articles concernant les chromosomes 22 (Dunham 1999) et 21 (Hattori 2000) sont parus avant la publication de la séquence brouillon. Depuis, les séquences des chromosomes 20 (Deloukas 2001), 14 (Heilig 2003) et 7 (Scherer 2003 ; Hillier 2003) ont été publiées. La France, qui est l'un des quatre pays à l'initiative du Projet Génome Humain, s'est impliquée tout particulièrement dans le séquençage du chromosome 14, par l'intermédiaire de son Centre National de Séquençage, le Génoscope (Heilig 2003).

L'analyse de la séquence « brouillon » du génome humain (22 paires d'autosomes et une paire de chromosomes sexuels), premier génome de vertébré séquencé, de taille équivalant à 8 fois l'ensemble des génomes séquencés jusque là, soit environ 3000 mégabases (Mb) permet de préciser un certain nombre d'observations antérieures (Lander 2001) :

- Le génome est constitué de zones hétérogènes à divers points de vue : densité en gènes, en éléments transposables, pourcentage en GC, îlots CpG, taux de recombinaison.
- Les séquences codantes représentent moins de 5% du génome humain, tandis que les séquences répétées en constituent plus de 50% dont : (1) des séquences dérivées de transposons qui ne sont plus actifs, appelées répétitions dispersées, dont les LINEs (« long interspersed elements » ; 20% du génome), les SINEs (« short interspersed elements » ; 13%), les transposons LTR (8%), et les transposons à ADN (3%) ; (2) des copies inactives (partiellement) de gènes provenant de rétropositions, appelées pseudogènes ; (3) des séquences répétées en tandem en majorité des répétitions de dinucléotides, (« SSR : simple sequence repeat ») qui représentent environ 3% du génome ; (4) des duplications de segments de 10 à 300 kb, copiés d'une région du génome

à une autre ; (5) des grands blocs de répétitions en tandem (ADN satellite), comme les centromères, les télomères, les bras courts des chromosomes acrocentriques et les gènes ribosomiques.

- Les éléments Alu (SINEs) sont préférentiellement localisés dans les régions riches en GC.
- Les régions péricentromériques et subtélomériques sont constituées de segments dupliqués provenant d'autres régions du génome. Ces duplications de segments sont beaucoup plus fréquentes chez l'homme que chez la levure, le nématode ou la drosophile.
- Le taux de mutation méiotique est à peu près le double chez les mâles par rapport aux femelles : la majorité des mutations surviennent chez les mâles.
- Les régions riches en GC sont bien corrélées avec les « bandes G noires » visibles sur les caryotypes.
- Les taux de recombinaison ont tendance à être plus élevés dans les régions distales (~ 20 Mb) et sur les bras courts des chromosomes, ce qui permet la survenue d'au moins un crossing-over par bras chromosomique à chaque méiose.

Cette séquence apporte également des réponses majeures :

- Il y a 30000 à 40000 gènes codant pour des protéines (environ 40% sont de fonction inconnue), seulement le double du nombre de gènes du nématode ou de la drosophile. Mais ces gènes sont plus complexes, subissant plus de phénomènes d'épissage alternatif qui génèrent un plus grand nombre de produits protéiques.
- L'ensemble des protéines codées par le génome humain (protéome) est plus complexe que celui des invertébrés : non seulement il contient des domaines protéiques et des motifs spécifiques des vertébrés, mais possède également un ensemble plus riche d'architectures de domaines, résultant de réarrangements entre des domaines pré-existants.
- Certains gènes humains semblent provenir d'éléments transposables et de transferts horizontaux depuis les bactéries. Cependant, la question du transfert horizontal des bactéries vers les eucaryotes reste très controversée, et cette hypothèse a été réfutée pour de nombreux gènes depuis la première analyse du génome humain (Genereux 2003).
- Plus d'1,4 millions de SNP (« single nucleotide polymorphism » : mutations ponctuelles) ont été identifiés, ce qui pourra permettre une étude du déséquilibre de liaison à l'échelle du génome entier.

1.1.2.1.2 Le séquençage du génome humain par la société CELERA

En 1998, la société CELERA a annoncé qu'elle construisait une plate-forme de séquençage afin de déterminer la séquence du génome humain en moins de trois ans. La méthode utilisée, appelée WGS (pour « whole-genome shotgun ») par opposition à celle du consortium public, basée sur le séquençage de BACs (« bacterial artificial chromosomes ») chevauchants de position connue sur le génome humain, a reposé sur l'assemblage de séquences aléatoires

provenant de banques plasmidiques générées à partir du génome entier de 5 individus (Venter 2001). La Figure 4 illustre ces deux stratégies de séquençage (Waterston 2002).

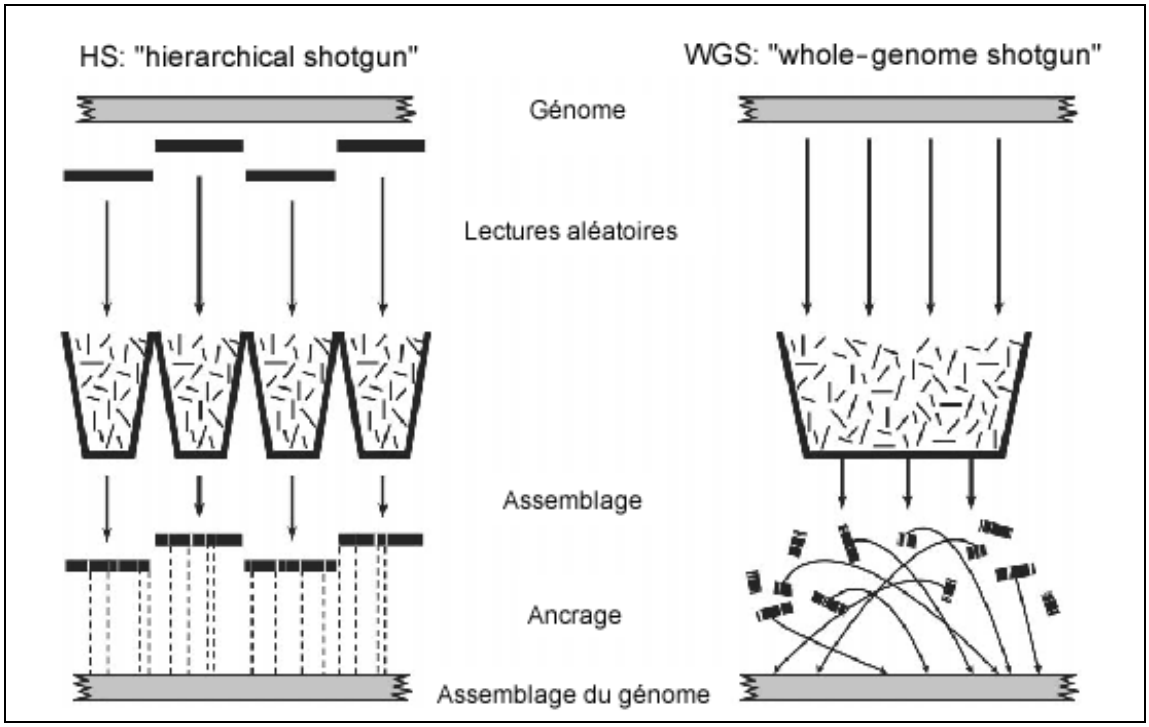


Figure 4 : Comparaison entre les stratégies de séquençage « hierarchical shotgun » et « whole genome shotgun », d'après (Waterston 2002).

L'approche WGS a été inventée et appliquée initialement pour le séquençage des génomes bactériens (voir chapitre 1.1.1). Afin de gagner du temps, la société CELERA a incorporé les données publiquement accessibles provenant du Projet Genome Humain, aux données produites par l'approche WGS afin de générer son propre assemblage. Le résultat de ce deuxième séquençage est paru le 16 février 2001 (Venter 2001), en même temps que celui du consortium public (15 février 2001 : Lander 2001), et a conduit à des conclusions similaires. Le Tableau 2 présente quelques caractéristiques du génome humain issues de la version CELERA (Venter 2001).

Le séquençage du génome humain par la société CELERA, dont le résultat n'est pas librement accessible, a déclenché de nombreuses polémiques. Par exemple, dans un article paru en mars 2002, trois membres du consortium du Projet Génome Humain affirment que la stratégie de séquençage « whole-genome shotgun » pour un génome aussi complexe et riche en répétitions que le génome humain n'a pas été prouvée par l'article de Venter *et al.*, car leur assemblage a largement profité de l'assemblage effectué par le consortium public (Waterston 2002). L'équipe CELERA a aussitôt répondu (Myers 2002) que ces assertions n'étaient pas fondées : les données provenant du Projet Genome Humain utilisées par CELERA auraient été suffisamment morcelées pour ne pas permettre de rétablir l'assemblage HGP. En tout état de cause, les séquences produites par ces deux projets ne sont pas indépendantes, ce qui n'a

jamais été nié par l'équipe CELERA. Nous en verrons une illustration dans l'article décrit au chapitre 2.3.2 (Denoeud 2003).

Tableau 2 : Quelques caractéristiques de la version du génome humain produite par CELERA, d'après (Venter 2001).

taille du génome (y compris les gaps)	2,91 Gb
taille du génome (excluant les gaps)	2,66 Gb
%AT	54
%GC	38
% bases indéterminées	9
fenêtre de 50 kb la plus riche en GC	chr 2 (65%)
fenêtre de 50 kb la moins riche en GC	chr X (25%)
% de la séquence correspondant à des répétitions	35
Nombre de gènes annotés	26383
Pourcentage de gènes annotés de fonction inconnue	42
Nombre de gènes hypothétiques et annotés	39114
Pourcentage de gènes hypothétiques et annotés de fonction inconnue	59
Gène contenant le plus d'exons	Titin (234 exons)
Taille moyenne des gènes	27 kb
Chromosome le plus riche en gènes	chr 19 (23/Mb)
Chromosome le moins riche en gènes	chr Y, chr 13 (5/Mb)
Longueur totale des déserts (>500 kb sans gènes annotés)	605 Mb
% pb dans des gènes (annotés / annotés + hypothétiques)	25,5 / 37,8
% pb dans des exons (gènes annotés / annotés + hypothétiques)	1,1 / 1,4
% pb dans des introns (gènes annotés / annotés + hypothétiques)	24,4 / 36,4
% pb intergéniques (gènes annotés / annotés + hypothétiques)	74,5 / 63,6
Chromosome avec la plus grande proportion d'ADN dans des exons annotés	chr 19 (9,33)
Chromosome avec la plus faible proportion d'ADN dans des exons annotés	chr Y (0,36)
Plus longue région intergénique	chr 13 (~3 Mb)
Taux de mutations ponctuelles (SNPs)	1/1250 pb

1.1.2.1.3 Comparaison entre les deux séquences produites (HGP et CELERA)

Les observations globales (nombre de gènes, pourcentage en GC, éléments répétés, grandes duplications…) issues des deux séquences « brouillons » du génome humain sont similaires. Cependant, même si ces deux études ont identifié environ 30000 gènes codant pour des protéines, il y a peu de recouvrement entre les gènes prédits *de novo* (Hogenesch 2001). Il apparaît que cette incohérence provient des séquences utilisées pour la prédiction de ces gènes et non pas de la méthode de prédiction : il y a des différences majeures entre ces deux assemblages du génome humain. En effet, la distribution et la longueur des contigs diffèrent (Aach 2001 ; Semple 2002), et les deux assemblages possèdent un certain nombre, comparable, de séquences uniques (Aach 2001). La Figure 5 présente une comparaison des assemblages CELERA et HGP pour les chromosomes 21 et 22, obtenue en positionnant des minisatellites identifiés au cours de cette thèse (Denoeud 2003) sur les scaffolds CELERA et les contigs HGP : on peut mettre en évidence des inversions entre les deux assemblages proposés.

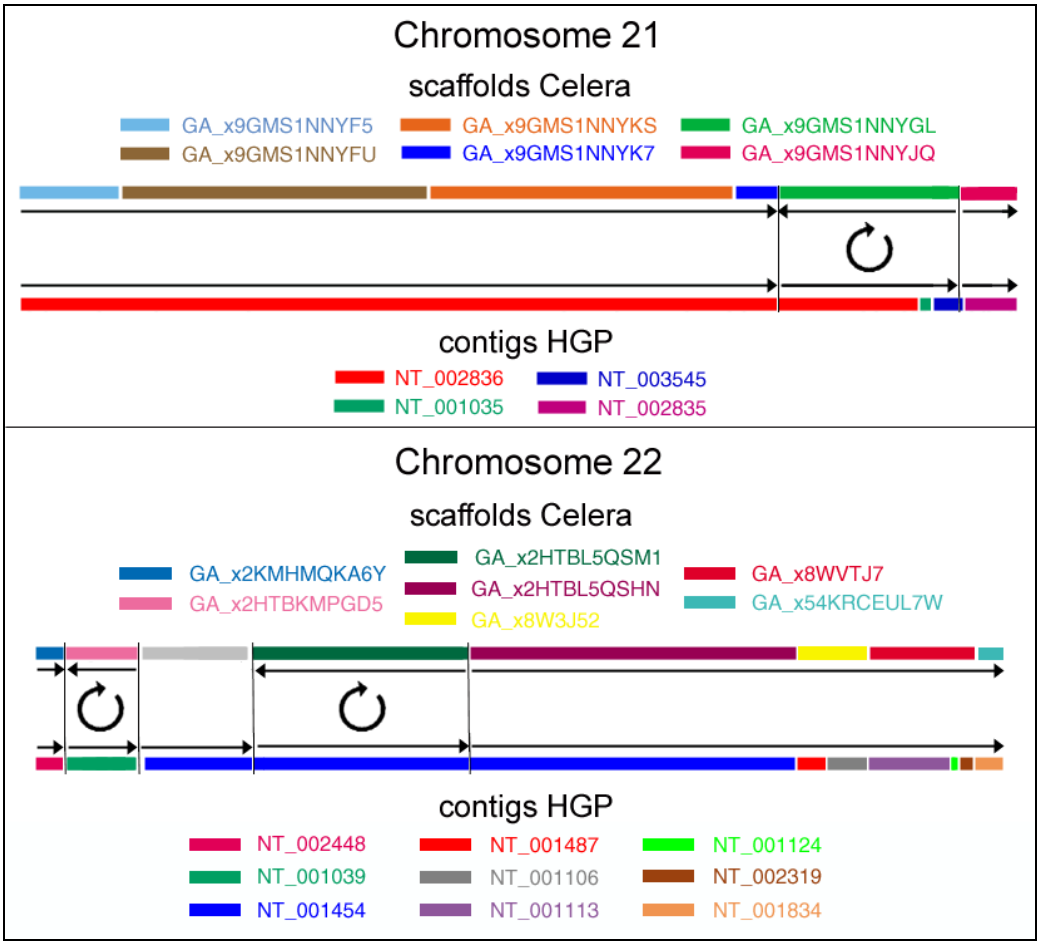


Figure 5 : Comparaison des assemblages proposés par le consortium public et la société CELERA pour les chromosomes 21 et 22.

Une comparaison des différents assemblages d'une portion du chromosome 4 suggère que le séquençage de répétitions en tandem est particulièrement délicat par l'approche « WGS » (Semple 2002). Nous présenterons au chapitre 2.3.2 une comparaison de la qualité du séquençage des répétitions en tandem dans les versions HGP et CELERA des chromosomes 21 et 22 (Denoeud 2003).

1.1.2.1.4 Applications du séquençage du génome humain

La fin du séquençage ne marque évidemment que le début de l'étude du génome humain. La première application du séquençage du génome humain porte sur les gènes responsables de pathologies : le fait de disposer de cette séquence a facilité l'identification de gènes de fonction biochimique inconnue par le clonage positionnel (Wooster 1995), ainsi que l'étude de liaisons génétiques en fournissant des candidats de marqueurs microsatellites. Cette séquence a également conduit à la mise en évidence des délétions chromosomiques à l'origine de syndromes : de telles délétions sont causées dans certains cas par des phénomènes de recombinaison entre deux régions provenant d'une duplication intrachromosomique (Shaikh 2000). De plus, la séquence du génome humain permet d'identifier des paralogues de gènes responsables de maladies, ce qui peut avoir deux intérêts : d'une part, de mettre en évidence

d'autres gènes causant des maladies génétiques apparentées (Kohl 2000), et d'autre part d'identifier de nouvelles voies thérapeutiques (Olivieri 1998).

Une seconde application porte sur l'identification de cibles pour les médicaments. L'identification de paralogues à des cibles connues de médicaments est prometteuse pour découvrir de nouvelles cibles (Davies 1999 ; Heise 2000).

Enfin, le séquençage du génome humain devrait également apporter des informations plus fondamentales sur la biologie et la physiologie humaines (Matsunami 2000).

1.1.2.2 Autres génomes eucaryotes

Chez les eucaryotes, 18 génomes (ou chromosomes) étaient séquencés en plus du génome humain au 5 août 2003 (voir GOLD [http://ergo.integratedgenomics.com/GOLD]), parmi lesquels ceux décrits dans le Tableau 3. La proportion de gènes de fonction inconnue varie entre 40 et 60% selon le génome eucaryote considéré.

Tableau 3 : Description de quelques génomes d'eucaryotes séquencés.

Organisme	Site Web dédié à la séquence	Nombre de chromosomes	Taille du génome	Nombre d'ORFs (phases ouvertes de lecture)	Densité en ORFs (nb d'ORFs/kb)
Saccharomyces cerevisiae (levure)	http://www.yeastgenome.org/	16	12069 kb	6294	0,52
Plasmodium falciparum (parasite causant la malaria)	http://www.plasmodb.org/	14	22900 kb	5268	0,23
Caenorhabditis elegans (nématode)	http://www.sanger.ac.uk/Projects/C_elegans/	6	97000 kb	19099	0,20
Arabidopsis. Thaliana (plante)	http://arabidopsis.org/home.html	5	115428 kb	25498	0,22
Drosophila melanogaster (drosophile)	http://flybase.bio.indiana.edu/	6	137000 kb	14100	0,10
Mus musculus (souris)	http://www.ncbi.nlm.nih.gov/genome/guide/mouse/	20	~2500 Mb	~30000	~ 0,01

1.2 Les répétitions en tandem

1.2.1 Définitions

Une répétition en tandem est une succession de motifs d'ADN répétés les uns derrière les autres, par opposition aux répétitions dispersées dont les unités répétées sont éparpillées dans le génome (comme décrit au paragraphe 1.1.2.1). Les différentes unités formant la répétition en tandem ne sont pas nécessairement identiques entre elles : le degré de conservation au sein d'une répétition en tandem est très variable. Les répétitions en tandem ont été en premier lieu étudiées chez les mammifères, où trois catégories ont été distinguées : les satellites, les minisatellites, et les microsatellites. Cette distinction correspond à différentes plages de taille (pour la longueur totale) et a été faite de façon plus ou moins arbitraire : l'ADN « satellite » a tout d'abord été observé et isolé par centrifugation sur gradient de densité, où il constituait une fraction particulière (dite « satellite ») (Britten 1968 ; Meneveri 1984). Ensuite, des répétitions en tandem de taille inférieure, pouvant être analysées grâce à la technique de Southern Blot, ont été caractérisées (Wyman 1980) puis appelées « minisatellites » (Jeffreys 1985a). Enfin, les répétitions en tandem de taille encore inférieure ont été nommées « microsatellites » lorsque l'avènement de la technique de PCR (« polymerase chain reaction ») en a fait des outils courants de génétique moléculaire. Avec le recul, les études menées, principalement chez les eucaryotes, sur les mécanismes de création et de mutation de ces structures souvent polymorphes laissent penser que cette distinction recouvre une réalité biologique, comme nous le verrons dans les paragraphes suivants.

Les répétitions en tandem sont présentes, dans des proportions variables, chez tous les organismes : eucaryotes, procaryotes, et même virus. Dans ces deux derniers groupes, les mécanismes sous-jacents n'ont quasiment pas été étudiés, et la distinction entre microsatellites et minisatellites est rarement faite. D'autres termes sont couramment employés, chez les procaryotes, mais également parfois chez les eucaryotes, pour désigner les répétitions en tandem : les SSR (« simple sequence repeat ») et STR (« short tandem repeat ») désignent des répétitions en tandem « simples » : elles correspondent aux microsatellites et aux minisatellites de petite taille (quelques centaines de paires de bases). Les VNTR (« variable number of tandem repeats ») désignent les répétitions en tandem polymorphes, qui peuvent appartenir à la classe des microsatellites ou des minisatellites.

1.2.2 Intérêts des répétitions en tandem chez les bactéries

Les répétitions sont une composante importante des génomes bactériens : elles peuvent constituer jusqu'à 10% de ces petits génomes, denses en gènes, et on trouve souvent dans la

littérature des allusions, peu précises, au fait que les répétitions en tandem bactériennes se trouveraient « fréquemment » dans des gènes. Même si ces observations sont fondées pour un certain nombre d'espèces bactériennes, elles restent impossibles à généraliser : la base de données des répétitions en tandem que je présente dans cette thèse (voir chapitre 2.1) devrait permettre de répondre à ces questions d'une façon plus rigoureuse.

1.2.2.1 Rôles dans l'adaptation et la virulence des bactéries : régulation de l'expression de gènes

Dans un certain nombre de pathogènes, des répétitions en tandem (SSR pour « simple sequence repeats ») présentes en amont ou dans les séquences codantes de protéines de surface sont polymorphes et contribuent à l'adaptation de la bactérie aux changements de conditions survenant au cours de l'infection de l'hôte. Ce phénomène, appelé variation de phase (revue : Henderson 1999), « allume »/« éteint » la synthèse protéique ou fait varier la structure de la protéine. Les locus à fort taux de mutation qui y sont impliqués sont appelés locus de contingence (Moxon 1994). Par exemple, chez la bactérie *Haemophilus influenzae*, qui colonise les voies respiratoires et peut causer des pneumonies et des méningites, le rôle des répétitions en tandem dans la modulation de la virulence est bien documenté. Tout d'abord, la variabilité des répétitions a été associée expérimentalement à la modulation des gènes impliqués dans la synthèse des pilus et du lipopolysaccharide (LPS) (Weiser 1989). Une répétition de dinucléotides dans un promoteur de gènes codant des sous-unités de pilus est un facteur régulateur majeur de leur expression : selon le nombre d'unités répétées, l'espacement entre les boîtes -35 et -10 est soit favorable soit défavorable à la reconnaissance de ce site par l'ARN polymérase (van Ham 1993). Les répétitions en tandem peuvent aussi moduler l'expression génique en étant à l'origine de blocages de la réplication (Krasilnikova 1998), ou en tant que terminateurs de transcription (Guerin 1998).

D'autres répétitions en tandem ont un effet au niveau de la traduction. Chez *H. influenzae*, différents gènes codant pour des enzymes de synthèse du LPS contiennent des répétitions de tétranucléotides, localisées dans les séquences codantes, ce qui est à l'origine de décalages du cadre de lecture (Weiser 1990). Un tétranucléotide, situé dans un gène homologue à une méthyltransférase de type III, est tellement instable qu'il génère même du mosaïcisme dans les cultures bactériennes (De Bolle 2000). Chez *Neisseria meningitidis*, une répétition en tandem de 7 pb dans la phase codante du gène PilQ affecte la biosynthèse des pilus de façon quantitative (Tonjum 1998).

L'implication des SSR dans la modulation de l'expression de gènes a été mise en évidence dans une grande variété d'autres bactéries (pour revue : voir van Belkum 1999b) dont *Escherichia coli* (Foster 1994), *Neisseria meningitidis* (van der Ende 1995), *Bacillus anthracis* (Jackson 1997), ou *Mycoplasma gallisepticum* (Glew 1998). Le séquençage des génomes bactériens a ouvert la voie à une analyse plus systématique des gènes potentiellement impliqués dans la variation de phase. Lorsque la séquence complète du

génome de *H. influenzae* a été connue, un catalogue des répétitions en tandem de type microsatellites a pu être établi (Fleischmann 1995) : la plupart de ces répétitions sont associées à des gènes potentiellement impliqués dans la virulence (molécules d'adhésion, enzymes de synthèse du LPS…) (Hood 1996). De la même façon, le séquençage du pathogène *Helicobacter pylori* (Tomb 1997) a permis de mettre en évidence une trentaine de gènes associés à des SSR, c'est-à-dire potentiellement impliqués dans la variation de phase (Saunders 1998). Ces gènes codent pour des enzymes de biosynthèse du LPS, des protéines de surface, et des enzymes de restriction. Le séquençage d'une seconde souche de cette bactérie (Alm 1999) a rendu possible l'identification de SSR polymorphes entre les deux souches considérées. La plupart de ces candidats subissent bel et bien une variation d'expression selon le nombre de répétitions.

Par ailleurs, des répétitions en tandem appartenant à des phases ouvertes de lecture et dont le motif est multiple de 3 génèrent des polymorphismes au niveau des protéines, ce qui peut être à l'origine d'une variation antigénique :

- Chez *Staphylococcus aureus*, des protéines de surface impliquées dans la reconnaissance des molécules d'adhésion de la matrice extracellulaire de l'hôte contiennent de nombreuses répétitions en tandem, dont le nombre de répétitions influe sur l'accessibilité du domaine actif (voir Tableau 4).

- Chez *Bacillus anthracis*, une répétition en tandem dans le gène de l'exosporium est à l'origine de variations de la longueur des filaments de la surface des spores (Sylvestre 2003).

- Chez les streptocoques du groupe A, la protéine M, protéine de surface et facteur de virulence est soumise à une grande variabilité antigénique causée vraisemblablement par des événements de recombinaison homologue (Hollingshead 1987).

- Chez les streptocoques du groupe B, la protéine alpha C, antigène de surface, contient une répétition en tandem polymorphe qui lorsqu'elle est délétée permet d'échapper à la réponse immunitaire de l'hôte (Madoff 1996 ; Gravekamp 1998).

- Chez le mycoplasme *Mycoplasma hyorhinis*, les protéines de surface du système VIp confèrent aux bactéries, par leur variation de taille, une résistance contre les anticorps produits par l'hôte (porc) (Citti 1997).

Le Tableau 4 liste des répétitions en tandem associées à des gènes de fonction connue chez différentes bactéries (van Belkum 1998) : selon la taille de leurs motifs répétés, certaines appartiennent à la classe des microsatellites et d'autres à la classe des minisatellites.

Tableau 4 : Description de gènes bactériens associés à des répétitions en tandem, d'après (van Belkum 1998).

Espèce	Motif répété	Gène	Niveau de régulation du gène		Fonction du gène
			transcription	Traduction / protéine	
H. influenzae	CAAT	*lic1-lic3*	-	+	biosynthèse du lipopolysaccharide
	GCAA	*yadA*	-	+	adhésine
	GACA	*lgtC*	-	+	glycosyltransférase
	TTGG	ND	-	+	protéines liant le fer
	AGTC	ND	-	+	méthyltransférase
	TTTA	ND	-	+	homologue d'une protéine de *Bacillus*
	TA	*HifA/B*	+	-	synthèse des fimbriae (pilus)
N. meningitidis	G	*lsi2*	+	-	biosynthèse du lipopolysaccharide
	CTCTT	*opa*	-	+	protéines d'opacité
	A	*opa*	+	-	protéines d'opacité
	G	*porA*	+	-	protéine de la membrane externe
S. aureus	93 pb	*fnb*	-	+	protéine se liant à la fibronectine
	561 pb	*cna*	-	+	adhésine du collagène
	81 pb	*coa*	-	+	coagulase
	24 pb	*spa*	-	+	protéine A
	18 pb	*clf*	-	+	récepteur du fibrinogène
Streptococcus spp.	60 pb	*pspA*	-	+	protéine de surface des pneumocoques
	69 pb	*emm*	-	+	protéine de résistance à la phagocytose
	246 pb	*αC*	-	+	protéine *αC*
E. faecalis	TAGTARR	*rep1*et *rep2*	+	-	itéron: régule la réplication et le transfert des plasmides
M. hyorhinus	36 et 39 pb	*vlp*	-	+	protéine membranaire variante
	A	*vlp*	+	-	protéine membranaire variante
M. bovis	24 pb	*vspA*	-	+	lipoprotéine de la surface membranaire
M. fermentans	A	*P78*	-	+	lipoprotéine de transporteur ABC
U. urealyticum	18 pb	*MB*	-	+	antigène spécifique de MB
B. anthracis	12 pb	*vvrA*	-	+	homologue de la protéine de la gaine microfilaire
L. monocytogenes	66 pb	*prfA*	-	+	homologue à l'internaline riche en leucine
E. coli	A et C	*lac*	+	-	β-galactosidase
A. marginale	87 pb	*Msp1α*	-	+	protéine majeure de la surface

1.2.2.2 Utilisation en épidémiologie

Les répétitions en tandem sont souvent polymorphes et leurs allèles, de tailles différentes, sont facilement identifiables par PCR à partir d'amorces spécifiques des flanquantes, suivie d'une simple migration sur gel. Cette procédure est illustrée par la Figure 6. Ces structures, dont l'analyse est simple et rapide, ont une utilité en tant que marqueurs épidémiologiques. En effet, afin de comprendre le mode de dissémination des infections dans les communautés et les hôpitaux, et d'appréhender les changements évolutifs qui ont donné lieu à des avantages sélectifs, la distinction précise entre différents isolats d'une même espèce bactérienne est indispensable. Elle peut être faite grâce aux répétitions en tandem polymorphes, en particulier

dans des espèces d'émergence récente, comme *Yersinia pestis* (Achtman 1999) et *Bacillus anthracis*, pour lesquelles elles constituent une source majeure de polymorphisme.

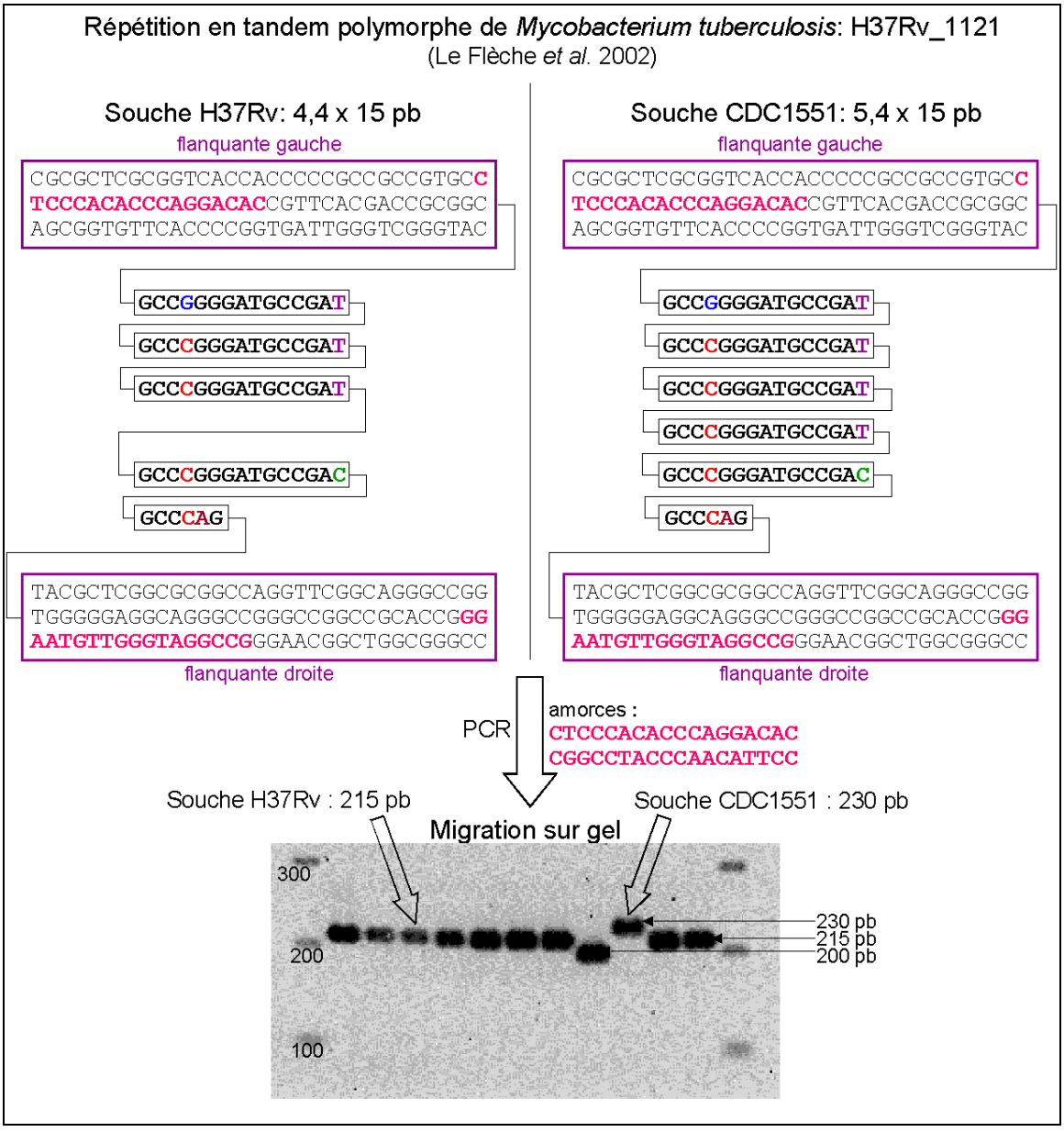


Figure 6 : Utilisation d'une répétition en tandem polymorphe afin de distinguer des souches bactériennes.

Quelques précautions doivent cependant être observées quant à l'utilisation des répétitions en tandem pour l'épidémiologie bactérienne (van Belkum 1999a). D'une part, certaines répétitions ne sont pas neutres du point de vue évolutif, c'est-à-dire que certains allèles peuvent conférer un avantage sélectif à certains isolats. Elles peuvent donc conduire à des conclusions erronées sur la proximité des souches, qui peuvent avoir le même allèle car elles ont été confrontées au même hôte et non pas parce qu'elles sont génétiquement proches. Cela dit, de nombreuses répétitions en tandem non neutres sont utilisées de façon efficace en épidémiologie : par exemple, celle de la coagulase de *Staphylococcus aureus* (Shopsin 2000).

D'autre part, les locus de contingence, subissant la variation de phase (voir paragraphe 1.2.2.1) n'ont aucune valeur épidémiologique et doivent par conséquent être évités. Ces locus sont toutefois rares : dans l'ensemble, les répétitions en tandem ne subissent pas d'altérations au cours de leur manipulation en laboratoire (van Belkum 1997 ; Stothard 1998).

Les répétitions en tandem ont été validées en tant que marqueurs épidémiologiques dans différentes espèces bactériennes, dont *Mycobacterium tuberculosis* (Frothingham 1998 ; Supply 2000), *Bacillus anthracis* (Jackson 1997 ; Keim 2000), et *Haemophilus influenzae* (van Belkum 1997). L'un des avantages majeurs de la technique de typage de répétitions en tandem est qu'elle est relativement simple et peu coûteuse par rapport aux autres techniques de génotypage communément utilisées (van Belkum 2001 ; Olive 1999). En outre, cette méthode est d'analyse facile, et reproductible, contrairement à celles qui génèrent des profils mutlibandes, comme :

- l'électrophorèse en champ pulsé (Tenover 1995)
- la RAPD (Random Amplified Polymorphic DNA) (Williams 1990)
- l'AFLP (Amplified fragment Length Polymorphism) (Vos 1995)
- la RFLP (Restriction Fragment Length Polymorphism), nommée ribotypage lorsqu'elle est appliquée à l'ADN ribosomique 16s-23s (Bingen 1994).

La technique MLST (multilocus sequence typing) (Maiden 1998), qui consiste à séquencer différents gènes de ménage, est très reproductible mais beaucoup plus coûteuse que le typage de répétitions en tandem ou MLVA (multilocus VNTR analysis). Le séquençage d'un nombre croissant de génomes bactériens va faciliter grandement l'identification de ce type de séquences. Dans la suite de cette thèse, je présenterai les différents outils informatiques que j'ai développés en tirant profit de la disponibilité de séquences, parfois de plusieurs souches d'une même espèce, afin d'identifier des marqueurs polymorphes pour le génotypage. Ces outils ont été appliqués avec succès à diverses espèces bactériennes, ce qui a fait l'objet de plusieurs publications (Le Flèche 2001; Le Flèche 2002 ; Pourcel 2003 ; Onteniente 2003).

1.2.2.3 Les familles de répétitions en tandem bactériennes

La notion de famille de répétitions en tandem a été introduite pour décrire des locus ayant des motifs répétés similaires et qui sont trouvés en plusieurs localisations. Souvent, il s'agit d'une simple coïncidence, d'une similitude de séquence due à la faible complexité du motif élémentaire et dans ce cas le terme de famille est abusif. Parfois, les « familles » existent bel et bien : deux cas se distinguent alors. Le plus simple est le cas de la famille dans laquelle la flanquante est également conservée. Par exemple, si un élément transposable contient une répétition en tandem, la répétition en tandem est alors un simple passager. Le cas de familles de répétitions en tandem sans flanquantes communes est plus curieux, et encore inexpliqué (Supply 1997). De façon encore plus intéressante, une famille de répétitions en tandem présente dans de nombreux génomes procaryotes (bactéries et archées) a été mise en

évidence (Mojica 2000). Ces éléments ont été appelés CRISPR (« clustered regularly interspaced short palindromic repeats »). Il s'agit de répétitions d'un motif bien conservé, de taille variant entre les génomes considérés (de 21 pb dans *S. typhimurium* à 37 pb dans la cyanobactérie *Anabaena sp.* (Masepohl 1996)), espacées par des séquences non conservées, mais de longueur constante, du même ordre de taille que le motif conservé (Jansen 2002). Elles peuvent être assimilées à des répétitions en tandem ayant un motif répété comprenant une région bien conservée et une région variable : ces régions variables sont appelées espaceurs.

Tableau 5 : Caractéristiques des CRISPRs trouvés dans différents génomes, d'après (Mojica 2000).

Organisme (archées, bactéries)	Taille du motif conservé répété (a)	Taille de l'espaceur (b)	Taille de l'unité répétée en tandem (a+b)	Nombre de locus dans le génome	Nombre de répétitions en tandem / locus
H. volcanii	30	nd*	nd	≥ 2	nd
H. mediterranei	30	33-39	63-69	3	21/nd/nd
M. jannaschii	28-30	31-51	59-81	$7^A + 6^B + 1^C$	4-25
M. thermoautotrophicum	30	34-38	64-68	2	124/47
A. fulgidus	$37^A/30^B$	~37	$\sim 74^A/\sim 67^B$	$1^A + 2^B$	$19^A/27^A/42^B$
S. solfataricus	25	~40	~65	≥ 2	94/102
P. abyssi	$29^A/30^B$	26-43	55-73	$1^A + 2^B$	$7^A/22^B/27^B$
P. horikoshii	29	34-58	63-87	3	18/26/66
A. pernix	$24^A/23^B$	37-52	60-76	$2^A + 1^B$	$19^A/27^A/42^B$
T . maritima	30	39-40	69-70	8	2-40
A. aeolicus	29	36-38	65-67	1	6
E. coli	29	32-33	61-62	3	2/7/13
S. typhi	29	32	61	≥ 1	6
C. jejuni	36	30	66	1	5
Y. pestis	28	32-33	60-61	2	6/9
C. difficile	29	36-38	65-67	$4^A + 2^B$	5-17
M. tuberculosis	36	38-40	74-76	1	Variable
Calothrix sp.	37	35-41	72-78	> 1	5
Anabaena sp.	37	32-43	69-80	> 1	17

* nd : non déterminé

* A, B, C : CRISPRs de différents types (plus de 3 pb de différence dans le motif conservé) dans un même organisme

Dans un génome, les CRISPRs peuvent se retrouver en plusieurs localisations, avec un motif répété très conservé d'un locus à l'autre, même si le nombre de répétitions peut varier (van Embden 2000). Dans certains cas, on trouve plusieurs types de CRISPRs, dont les motifs conservés sont différents, dans un même génome. Le nombre de répétitions peut également varier, en un même locus, d'une souche à l'autre : certains CRISPRs ont donc été exploités pour l'épidémiologie (Kamerbeek 1997). En revanche, d'une espèce à l'autre (sauf espèces proches comme par exemple *Pyrococcus abyssi*, *horikoshii* et *furiosus* (Zivanovic 2002)), les motifs conservés constituant les CRISPRs ont peu d'homologie : une exception est à noter entre *Neisseria meningitidis* et *Pasteurella multocida*. Certaines caractéristiques de séquence sont cependant communes aux CRISPRs de toutes les espèces : le motif répété conservé contient une symétrie dyade (souvent GTT à une extrémité et AAC à l'autre extrémité). En général, une séquence flanquante de quelques centaines de pb est conservée entre tous les

locus CRISPRs d'un génome donné, qui n'a pas d'homologie avec les flanquantes des CRISPRs d'autres génomes. Enfin, tous les génomes bactériens ne contiennent pas de CRISPRs et la présence de ces CRISPRs est strictement associée à des gènes appelés *cas* pour « CRISPR associated genes », qui pourraient correspondre à des protéines se liant à l'ADN (Jansen 2002). Les mécanismes de création et de dispersion dans les génomes, ainsi que la fonction biologique des CRISPRs, sont encore inconnus et méritent d'être étudiés. Le Tableau 5 présente les caractéristiques des CRISPRs trouvés dans différents génomes (Mojica 2000). Il s'agit cependant là d'une classe très particulière de répétitions en tandem, qu'il ne faut pas confondre avec les répétitions en tandem de type mini- ou microsatellites. L'existence de répétitions de type CRISPR dans les génomes eucaryotes n'a pas été documentée jusqu'à présent. Il serait toutefois intéressant de s'assurer de l'absence de tels éléments en explorant de façon systématique les génomes eucaryotes dont la séquence est achevée (une recherche effectuée sur les génomes eucaryotes accessibles dans la base de données des répétitions en tandem n'a pour l'instant pas mis en évidence de locus CRISPR).

1.2.3 Intérêts des répétitions en tandem chez l'Homme

D'après l'article analysant la séquence issue du Projet Génome Humain, les répétitions en tandem représentent environ 3% du génome humain, la majorité (en nombre de locus) étant des répétitions de dinucléotides (0.5% de l'ensemble du génome). Il y a environ 1 répétition en tandem tous les 2 kilobases, soit 437 répétitions en tandem par mégabase.

1.2.3.1 L'ADN satellite

L'ADN satellite rencontré dans les grands génomes correspond à des structures dont la longueur totale atteint plusieurs mégabases (Mb). L'ADN satellite est localisé dans les régions péricentromériques ou d'hétérochromatine télomérique des eucaryotes (Charlesworth 1994). Les satellites semblent jouer des rôles importants dans la structuration des génomes : ils sont les constituants majeurs des centromères fonctionnels, ce qui a été montré chez l'homme (Schueler 2001) et chez la drosophile (Sun 1997), où ils sont nécessaires pour le bon déroulement de la mitose et de la méiose (Csink 1998). Les séquences des satellites centromériques diffèrent même entre des organismes proches, ce qui est associé à des changements dans les histones correspondantes (Henikoff 2001). L'ADN satellite centromérique évolue rapidement (Ugarkovic 2002), ce qui pourrait, en provoquant l'évolution adaptative des histones centromériques, être à l'origine du processus de spéciation (Malik 2001).

1.2.3.2 Intérêts des microsatellites humains

Les microsatellites ont une unité répétée de 1 à 10 paires de bases (pb) environ, et une longueur totale de quelques dizaines à quelques centaines de paires de bases. Ils sont répartis

de façon relativement homogène le long des chromosomes eucaryotes, même si leur abondance dans différentes régions du génome varie selon le type de répétition (mononucléotides, dinucléotides…) (Toth 2000) et que dans certaines espèces, ils sont moins abondants aux environs des centromères, par exemple chez *A. thaliana* (Lin 1999) ou ont tendance à s'organiser en clusters, par exemple chez la drosophile (Bachtrog 1999). En particulier, les régions codantes sont moins riches en microsatellites que les régions non-codantes sauf en ce qui concerne les motifs de 3 ou 6 pb, ce qui peut correspondre à une contre-sélection des mutations de décalage du cadre de lecture (Metzgar 2000 ; Morgante 2002 ; Li 2002). Certaines de ces observations restent approximatives et mériteront d'être ré-examinées, puisque, comme nous le montrons dans cette thèse, l'étude des répétitions en tandem à l'échelle de génomes entiers est dorénavant possible.

1.2.3.2.1 Utilisation comme marqueurs génétiques

Les microsatellites, qui sont distribués de façon homogène sur tout le génome humain, et sont fréquemment polymorphes, représentent la source la plus abondante de marqueurs génétiques chez l'Homme. Ils sont à la base de l'élaboration de la carte du génome humain par le Généthon, en particulier les microsatellites de type (CA/GT) (Gyapay 1994 ; Dib 1996). Les répétitions de mononucléotides poly (A/T), de dinucléotides (CA/GT) et de trinucléotides (AAT) sont les plus communes (Beckmann 1992 ; Hancock 1998 ; Lander 2001). Le Tableau 6 présente le nombre de répétitions en tandem de longueur comprise entre 1 et 10 sur le génome humain (Lander 2001).

Tableau 6 : Description des répétitions en tandem trouvées dans la séquence du génome humain produite par le consortium public, d'après (Lander 2001).

Taille du motif répété (pb)	Nombre moyen de bases appartenant à une répétition en tandem, par Mb	Nombre moyen de répétitions en tandem par Mb
1	1660	36,7
2	5046 (dont AC: 50%, AT: 35%, AG: 15%, GC: 0,1%)	43,1
3	1013 (dont AAT: 33%, AAC: 21%, ACC: 4%, AGC: 2,2%, ACT: 1,4%, ACG: 0,1%)	11,8
4	3383	32,5
5	2686	17,6
6	1376	15,2
7	906	8,4
8	1139	11,1
9	900	8,6
10	1576	8,6

1.2.3.2.2 Différents rôles biologiques attribués aux microsatellites

Les microsatellites sont associés à un certain nombre de processus cellulaires, chez l'homme comme chez d'autres organismes :

- Structure de l'ADN :
  Dans différentes espèces, l'abondance relative de répétitions dinucléotidiques semble être associée à différentes caractéristiques structurales à grande échelle de l'ADN, comme la courbure ou le surenroulement (Karlin 1998 ; Baldi 2000).

- Structure des télomères :
  Les télomères humains sont constitués d'une répétition en tandem d'un motif de 6 paires de bases, de type microsatellite : TTAGGG. L'ADN des mammifères étant linéaire, deux problèmes majeurs se posent : comment répliquer les molécules d'ADN sans en raccourcir les extrémités, et comment les extrémités de l'ADN sont-elles protégées de l'activité des exonucléases ? Chez la plupart des eucaryotes, les extrémités des chromosomes (télomères) sont répliquées par une polymérase particulière, appelée télomérase. Cette ribonucléoprotéine qui a une activité de transcriptase inverse utilise son propre ARN comme matrice pour la synthèse d'ADN. L'activité de la télomérase est élevée dans les cellules embryonnaires et devient inexistante dans les cellules différenciées. De plus, son activité varie lorsque les cellules entrent ou sortent du cycle cellulaire. Ces observations ont conduit à formuler l'hypothèse suivante : le raccourcissement progressif des télomères dans les cellules somatiques conduirait à l'arrêt de la division cellulaire et à la sénescence ; les télomères seraient ainsi impliqués dans le processus de vieillissement cellulaire et de cancérogénèse (Autexier 1996). A l'inverse, en conservant une longueur constante de leurs télomères, les cellules germinales (télomères courts) et cancéreuses (télomères longs) continueraient de se diviser indéfiniment. La séquence répétée constituant les télomères humains (TTAGGG)n est riche en G, ce qui confère à ces séquences simple-brin la capacité de se replier sous forme de structures 4-brins ou tétraplexes (Henderson 1987), stables dans les conditions physiologiques et insensibles aux endo- et exonucléases.

- Effet sur la recombinaison :
  Les microsatellites pourraient correspondre à des points chauds de recombinaison. En effet, une répétition $(GT)_{30}$ insérée dans un chromosome de levure, augmente la fréquence de recombinaison méiotique et de conversion génique dans les régions adjacentes (Treco 1986). De même, dans des cellules humaines en culture, cette répétition $(GT)_{30}$ augmente la recombinaison entre vecteurs plasmidiques (Wahls 1990). De plus, les répétitions dinucléotidiques ont une forte affinité pour les enzymes de recombinaison (Biet 1999), qui augmente avec le nombre de copies (Dutreix 1997). Il a également été proposé que les microsatellites influencent directement la recombinaison par l'intermédiaire de leurs effets sur la structure de l'ADN (conformation Z) (Biet 1999). Enfin, sur le chromosome 22 humain, la distribution des répétitions de type (GT/CA) est corrélée aux points chauds de recombinaison (Majewski 2000).

- Effet sur la réplication de l'ADN et le cycle cellulaire :
  Les microsatellites pourraient affecter la réplication. Par exemple, dans des cellules de mammifères, des phénomènes d'amplification sont co-sélectionnables avec un

microsatellite instable (Caligo 1999). De plus, certaines enzymes du contrôle du cycle cellulaire sont codées par des gènes contenant des répétitions en tandem, qui si elles sont mutées, par exemple dans un contexte d'instabilité causé par des mutations du système de réparation des mésappariements, peuvent contribuer à la transformation cancéreuse (Johannsdottir 2000).

- Association avec les enzymes de réparation des mésappariements (« MMR ») :
  Les enzymes de réparation des mésappariements corrigent les erreurs de réplication et inhibent la recombinaison entre des séquences divergées, ce qui contrôle le taux de mutation et l'adaptation évolutive. Les gènes MMR mineurs contiennent des répétitions de A chez une grande variété d'eucaryotes : *Homo sapiens*, *Mus musculus*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Drosophila melanogaster*, *Arabidopsis thaliana* ainsi que chez le procaryote *Escherichia coli*. Ces microsatellites, vulnérables aux insertions et délétions spontanées, peuvent introduire des décalages du cadre de lecture dans les protéines correspondantes. La perte d'activité d'un de ces gènes mineurs de réparation des mésappariements provoquerait alors des phénotypes mutateurs modérés (par rapport à des mutations dans les gènes MMR majeurs), ce qui permettrait de moduler le taux de mutations adaptatives nécessaires à l'évolution (Chang 2001).

- Régulation de gènes :
  Les répétitions en tandem présentes dans les promoteurs semblent pouvoir affecter l'activité des gènes en agissant au niveau transcriptionnel : l'augmentation du nombre de répétitions peut faire diminuer le taux de transcription (Lanz 1995 ; Gebhardt 2000), ou le faire augmenter (Okladnova 1998 ; Xu 1998). Des microsatellites présents dans des introns semblent aussi affecter la transcription, soit en agissant comme des régulateurs transcriptionnels (Meloni 1998), soit en modifiant la courbure de l'ADN pour rapprocher le promoteur de régions régulatrices (Gebhardt 1999). Par ailleurs, chez *S. Cerevisiae*, les répétitions trinucléotidiques sont plus fréquentes dans les gènes régulateurs de la transcription et de la transduction du signal, et sont sous-représentées dans les gènes de protéines de structure, ce qui suggère également une influence de ces éléments sur la transcription (Young 2000).

  Certaines séquences microsatellites ont également la capacité de lier des protéines (éléments régulateurs). Par exemple, des protéines se liant aux poly(GA) et poly(GT) ont été identifiées dans des fibroblastes humains (Aharoni 1993), et des séquences (GT)n ou (GT)n(GA)n présentes dans les introns de gènes du système immunitaire se lient à des protéines nucléaires (Epplen 1993). Le nombre de répétitions peut influer sur ces liaisons, la protéine ne se liant qu'à des répétitions en tandem d'une certaine taille (Winter 1989).

  Enfin, différentes études ont montré que les microsatellites avaient un effet sur la traduction, par exemple chez la bactérie *Escherichia coli* (Martin-Farmer 1999), et chez l'homme (Sandberg 1997).

De plus, le polymorphisme des microsatellites codants peut avoir des conséquences sur des pathologies humaines, comme nous le décrirons dans le paragraphe suivant.

1.2.3.2.3 Pathologies associées aux microsatellites

1.2.3.2.3.1 Les maladies neurodégénératives liées à l'expansion de triplets

Tableau 7 : Description des pathologies humaines causées par l'expansion de triplets, d'après (Timchenko 1999).

Maladie	Motif répété (nb copies individus atteints)	Localisation de la répétition	Transmission	Conséquences de la mutation	Références
SBMA : atrophie musculaire spinale et bulbaire	(CAG)40-62	ORF	dominante liée à l'X	Traduction de protéines mutantes contenant de longues régions poly-glutamine, qui conduisent à la mort cellulaire (voir Robitaille 1997 ; La Spada 2003)	(Brooks 1995)
HD : maladie de Huntigton	(CAG)36-121	ORF	autosomale dominante		(Hofferbert 1997)
DRPLA : atrophie dentatorubrale et pallidoluysienne	(CAG)49-88	ORF	autosomale dominante		(Nagafuchi 1994)
MJD/SCA3 : maladie de Macado-Joseph/ ataxie spinocérébelleuse de type 3	(CAG)68-79	ORF	autosomale dominante		(Cancel 1995)
SCA1 : ataxie spinocérébelleuse de type 1	(CAG)41-81	ORF	autosomale dominante		(Orr 1993)
SCA2 : ataxie spinocérébelleuse de type 2	(CAG)35-59	ORF	autosomale dominante		(Imbert 1996)
SCA6 : ataxie spinocérébelleuse de type 6	(CAG)21-27	ORF	autosomale dominante		(Zhuchenko 1997)
FraX : syndrome de l'X fragile	(CGG)60-2000	5' UTR	dominante liée à l'X	plus de transcription de FMRP1, protéine se liant à l'ARN	(Verkerk 1991; Fu 1991)
DM : dystrophie myotonique	(CTG)50-2000	3' UTR	autosomale dominante	1- plus de maturation de l'ARN de DMPK 2- plus de transcription de DMPK	(Brook 1992)
FRDA : ataxie de Friedreich	(GAA)200-900	intron 1	autosomale récessive	plus d'ARNs messagers de la frataxine dans les mitochondries conduisant à l'altération du métabolisme du fer	(Campuzano 1996)

Des expansions de répétitions trinucléotidiques sont associées à des maladies génétiques neurodégénératives (pour revue, voir : Monckton 1995 ; Wells 1996 ; Timchenko 1999) : le Tableau 7 décrit les différentes maladies de ce type. Elles sont caractérisées par l'expansion de triplets qui passent d'un nombre de copies inférieur à 15 chez les individus normaux à un nombre beaucoup plus élevé, pouvant atteindre plusieurs centaines de copies selon la maladie, chez les individus atteints. Un phénomène d'anticipation y est associé : la pathologie est de plus en plus sévère et survient de plus en plus tôt de génération en génération, ce qui correspond à une augmentation du nombre de répétitions (Wells 1996).

D'autres maladies sont associées à des microsatellites dont l'unité n'est pas un triplet. C'est le cas pour la rétinopathie diabétique, associée à une répétition (CA)n en 5' de l'aldose réductase (Fujisawa 1999), ainsi qu'à une répétition (CCTTT)n dans le promoteur de NOS2A (Warpeha 1999) et pour la fibrose du poumon après greffe, associée à une répétition (CA)n dans le premier intron de l'interféron γ (Awad 1999).

1.2.3.2.3.2 Microsatellites et cancer

Environ 15% des cancers sporadiques colorectaux (Liu 1995), gastriques (Leung 2000) et endométriaux (Basil 2000) sont associés à une instabilité des microsatellites ou MSI (« microsatellite instability ») (revue : Atkin 2001). On distingue deux types de variations des microsatellites associées aux cancers humains : le type A correspond à des petites variations de taille, de moins de 6 paires de bases, et le type B à des variations plus importantes, supérieures à 8 pb. L'instabilité de type A a été associée à des défauts du système de réparation des mésappariements. Une telle corrélation n'est toujours pas claire pour l'instabilité de type B, qui semble par ailleurs être associée à des prédispositions familiales (revue : Li 2002). Lorsque le système de réparation des mésappariements n'est plus efficace, le taux de mutation des microsatellites augmente 100 fois et les protéines contenant ce type de séquences risquent donc d'être altérées. Des mutations de décalage du cadre de lecture ont effectivement été trouvées, avec des fréquences variables, dans des gènes codant pour des protéines impliquées dans la transduction du signal, l'apoptose et les mécanismes inflammatoires, la régulation de la transcription, et la réparation de l'ADN (Jacob 2002). Cependant, il est difficile de savoir si ces gènes sont les véritables cibles du processus d'inactivation par décalage du cadre de lecture et sont donc impliqués dans l'initiation des tumeurs ou leur progression, ou si leur mutation n'est qu'une conséquence sans pertinence de l'altération du système de réparation des mésappariements (Woerner 2003). Différentes cibles, suppresseurs de tumeurs, contenant des répétitions mutant à forte fréquence dans le contexte « MSI » ont déjà été identifiées, par exemple, TGFβ-RII (récepteur d'hormone de croissance) (Markowitz 1995), et APC (Laken 1997).

1.2.3.3 Intérêts des minisatellites humains

Les minisatellites sont constitués d'unités d'une quinzaine à plusieurs centaines de paires de bases répétées sur une centaine à quelques milliers de paires de bases. Le premier minisatellite humain a été mis en évidence en 1980 par Wyman *et al.* (Wyman 1980), et le premier minisatellite de levure en 1984 par Horowitz et Haber (Horowitz 1984). Le nom de « minisatellites » leur a été donné par Jeffreys en 1985, qui a découvert la grande utilité de ces structures extrêmement polymorphes pour réaliser les premières « empreintes génétiques » (Jeffreys 1985a ; Jeffreys 1985b). Des minisatellites ont été identifiés et caractérisés dans divers génomes (pour revue, voir Bois 2003). Chez l'homme, 90% des minisatellites sont localisés dans les régions subtélomériques (Royle 1988), ce qui n'est pas le cas pour d'autres eucaryotes comme la souris, le rat ou le porc (Amarger 1998).

1.2.3.3.1 Utilisation pour les empreintes génétiques et en tant que marqueurs génétiques

Les minisatellites ont été largement utilisés pour établir des empreintes génétiques (Jeffreys 1985b). Il s'agit d'hybrider des sondes contenant des séquences riches en GC sur des Southern Blots. La technique de Southern Blot consiste à digérer l'ADN génomique d'un individu par des enzymes de restriction. Les fragments d'ADN sont ensuite séparés par électrophorèse puis transférés sur une membrane. La sonde s'hybridant à de nombreuses répétitions en tandem de taille variable, des profils complexes, uniques pour chaque individu, sont générés. Cette technique a été très utilisée dans les débuts de la médecine légale et pour les recherches de paternité mais a été depuis remplacée par le typage de microsatellites : elle avait en effet l'inconvénient d'être difficile à interpréter à cause de la multiplicité des bandes (revue : Bennett 2000). Par ailleurs, même si les microsatellites sont, grâce à leur distribution homogène, un outil de choix pour la cartographie génétique, les minisatellites restent une source de marqueurs génétiques pour des investigations concernant les extrémités télomériques où ils sont plus abondants. Par exemple, le retard mental est souvent associé à des délétions d'extrémités chromosomiques, et certaines ont pu être mises en évidences grâce à des marqueurs minisatellites (Giraudeau 1997 ; Giraudeau 2001).

1.2.3.3.2 Rôles biologiques attribués aux minisatellites

La plupart des minisatellites sont localisés dans des régions intergéniques. Ils peuvent cependant avoir des fonctions. Elles sont évoquées dans la revue (Vergnaud & Denoeud 2000) qui sera présentée au chapitre 2.3.1:

- Certains sont soupçonnés de jouer un rôle dans les phénomènes d'empreinte parentale (Chaillet 1995 ; Neumann 1995).
- D'autres constituent des sites fragiles (revue : Handt 2000), cassures chromosomiques survenant lors de la métaphase à des localisations caractéristiques et induites par différents types de traitements des cellules en culture (Sutherland 1998). Par exemple, le site fragile FRA10B est un minisatellite d'unité répétée de 42 pb (Hewett 1998) et FRA16B est un minisatellite de motif de 33 pb riche en AT (Yamauchi 2000). Le minisatellite situé dans le locus IGH (Tableau 8) joue peut-être également un rôle dans les événements de recombinaison qui ont lieu à ce locus (Brusco 1999).
- Enfin, de façon plus hypothétique, les minisatellites pourraient jouer un rôle dans l'initiation de l'appariement des chromosomes (Ashley 1994 ; Sybenga 1999), ce qui peut être lié à leurs propriétés recombinogènes supposées (Boan 1998 ; Wahls 1998).
- Des minisatellites associés à des gènes, dans les séquences régulatrices, les introns, ou les séquences codantes, peuvent avoir un effet sur certaines pathologies humaines : nous en dressons une liste dans le paragraphe suivant.

1.2.3.3.3 Minisatellites associés à des gènes

La revue que nous présenterons au chapitre 2.3.1 (Vergnaud & Denoeud 2000) ne détaille pas les minisatellites associés à des gènes. Je présente donc ici une synthèse bibliographique des minisatellites présents dans des introns, exons, ou au voisinage de gènes, et dont le polymorphisme a été étudié : certains sont impliqués dans des pathologies.

1.2.3.3.3.1 Minisatellites à proximité de gènes ou dans des introns (non codants)

Le Tableau 8 liste les minisatellites situés dans des introns, des régions flanquantes en 3' ou en 5' de gènes, ou au niveau de jonctions introns/exons, et dont le polymorphisme a été étudié. D'autres répétitions en tandem polymorphes, ou « VNTRs » (pour « Variable Number of Tandem Repeats »), associées à des gènes ont été étudiées, mais elles n'appartiennent pas à la classe des minisatellites, et ne sont donc pas listées ici. Cette recherche bibliographique a permis de dénombrer (par des requêtes dans PubMed [http://www.ncbi.nlm.nih.gov/PubMed/] du type « minisatellite+gene+polymorphism» qui génère plus de 600 résultats) : 23 minisatellites polymorphes localisés dans des introns, dont 3 au niveau de sites d'épissage, et 20 localisés en amont ou en aval de gènes. Pour la plupart de ces minisatellites, une association entre certains allèles et des pathologies humaines a été décelée, soit indirectement, l'association reflétant alors un déséquilibre de liaison dû à la proximité entre le minisatellite et la mutation responsable du phénotype, soit directement, le minisatellite lui-même étant impliqué dans la pathologie par l'une de ces deux voies majeures:

- soit en modulant l'expression du gène situé à proximité (pour revue, voir : Nakamura 1998) : c'est par exemple le cas pour les minisatellites en amont des gènes l'insuline (Bennett 1995) et de la monoamine oxydase A (Sabol 1998).
- soit, pour les minisatellites localisés dans des introns, en jouant un rôle sur l'épissage (Turri 1995) : par exemple, le minisatellite de l'intron 7 du facteur de coagulation VII est le site d'épissages anormaux conduisant à une pathologie (Borensztajn 2003).

Parmi les 43 minisatellites décrits dans le Tableau 8, 6 sont fortement polymorphes (ils ont 20 allèles ou plus). Par exemple, le minisatellite en amont du gène HRAS compte 4 allèles dit « communs » (30, 46, 68 et 84 copies) et plus d'une trentaine d'allèles rares, qui sont associés à la cancérogenèse (Krontiris 1993). Certains de ces minisatellites très polymorphes sont instables. C'est le cas des allèles « prémutateurs » du minisatellite du gène CSTB impliqué dans l'épilepsie, comptant 12 à 17 copies d'un dodécamère, qui mutent à une fréquence élevée pour générer des allèles « mutants » comptant plus de 45 copies, responsables de la pathologie via une diminution du nombre de transcrits (Lalioti 1997).

Tableau 8 : Minisatellites polymorphes au voisinage de gènes ou dans des introns.

Nom du gène (nom du minisatellite)	fonction(s) de la protéine	pathologie(s) associée(s) au gène	association du polymorphisme du ms avec la(les) pathologie(s)	localisation du minisatellite dans le gène	taille du motif répété	plage de nombre de copies	nombre d'allèles identifiés	numéro d'accession (position du minisatellite dans la séquence)	Locus ID* (Locus Link)	Chromosome	références
ABO	glycosyltransférase qui génère les antigènes spécifiques des groupes sanguins A ou B		? (fixation de facteurs de transcription)	5' UTR	43 pb	1-4	2	AF014105 (108-280)	28	9q34.1-q34.2	(Kominato 1997 ; Irshaid 1999)
APOB	apolipoprotéine B: rôle dans l'homéostasie du cholestérol	diverses pathologies dues à l'hypercholestérolémie	-	région hypervariable en 3'	15 pb	15-55	> 25	J05157	338	2p24-p23	(Huang 1987 ; Destro-Bisol 2000)
CBS	cystathionine beta-synthase: métabolisme de l'homocystéine	hyperhomocystéinémie légère: facteur de risque pour les maladies cardiovasculaires	+	frontière exon 13-intron 13	31 pb	16-21	5	AF042836 (22990-23516)	875	21q22.3	(Yang 2000 ; Lievers 2001)
COL2A	collagène de type II, chaîne alpha 1	collagénoses : achondroplasie (défaut de formation du cartilage), ostéoarthrose…	+	région flanquante en 3'	34 et 31 pb	8-17	>20	L10160	1280	12q13.11-q13.2	(Berg 1993 ; Katsuyama 1998; Meulenbelt 1999)
COL3A1	collagène de type III, chaîne alpha 1	syndrome coronarien aigu	+	intron 25	15 pb	3-13	11	AY016295 (57618-57692)	1281	2q31	(Mays 1992 ; Muckian 2002)
CSTB (EPM1 minisatellite)	cystatine B: inhibiteur de protéase à cystéine	épilepsie myoclonique progressive du type de Unverricht-Lundborg	+ : pathologie due à l'expansion du dodécamère (fait baisser le taux d'ARNm)	région flanquante en 5'	12 pb	2-70	> 30	AF208234 (3255-3294)	1476	21q22.3	(Lalioti 1997 ; Larson 1999)
DAT	transporteur de la dopamine	troubles psychiatriques, maladie de Parkinson, hyperactivité…	+ (agit sur le niveau de transcription)	3' UTR	40 pb	3-11	≥ 8	NM_001044 (2741-3144)	6531	5p15.3	(Vandenbergh 1992 ; Doucette-Stamm 1995 ; Mill 2002 ; Lin 2003)
F7/ IVS7 repeated element	facteur de coagulation VII: dépendant de la vitamine K	hémorragie sévère	+ (effet sur l'épissage)	Frontière exon7-intron 7	37 pb	4-8	5?	J02933 (10234-10461)	2155	13q34	(O'Hara 1988 ; Borensztajn 2003)
FSTL3	« follistatin-like 3 » (glycoprotéine sécrétée)	rôle dans l'apparition de leucémies ?	?	3' UTR	24 pb	?-13-?	≥2	U76702 (1213-1524)	10272	19p13	(Hayette 1998)
G1P3	protéine inductible par l'interféron alpha : fonction inconnue		site d'épissage alternatif	frontière exon 2 - intron 2	12 pb	1-3	3	NM_022873 (170-206)	2537	1p35	(Turri 1995)
HRAS1	proto-oncogène : impliqué dans la transduction du signal mitogène et la différenciation	cancer	+ (effet sur la transcription ; allèles rares = risques de cancer)	1 kb en aval (3')	28 pb	30-100	> 30 (4communs : 30, 46, 68, 84 copies)	a1:AF105318 a2:AF105319 a3:AF105320 a4:AF105321	3265	11p15.5	(Krontiris 1993 ; Weitzel 2000 ; Langdon 2003)
hTERT (hTERT-VNTR 2-1)	sous-unité catalytique de la télomérase	cancer : télomérase=oncogène	? (fixation de facteurs de transcription)	intron 2	42 pb	40-111	6	AY007685 (27650-29650)	7015	5p15.33	(Leem 2002)
hTERT (hTERT-VNTR 2-2)	sous-unité catalytique de la télomérase	cancer : télomérase=oncogène	? (fixation de facteurs de transcription)	intron 2	61 pb	40-44	4	AY007685 (34111-35453)	7015	5p15.33	(Leem 2002)

Nom du gène (nom du minisatellite)	fonction(s) de la protéine	pathologie(s) associée(s) au gène	association du polymorphisme du ms avec la(les) pathologie(s)	localisation du minisatellite dans le gène	taille du motif répété	plage de nombre de copies	nombre d'allèles identifiés	numéro d'accession (position du minisatellite dans la séquence)	Locus ID* (Locus Link)	Chromosome	références
hTERT (hTERT-VNTR 6-1)	sous-unité catalytique de la télomérase	cancer : télomérase=oncogène	-	intron 6	38 pb	27-47	8	AY007685 (41485-42324)	7015	5p15.33	(Leem 2002)
hTERT (hTERT-VNTR 6-2)	sous-unité catalytique de la télomérase	cancer : télomérase=oncogène	-	intron 6	36 pb	23-88	30	AY007685 (44225-45115)	7015	5p15.33	(Leem 2002)
5-HTT	transporteur de la sérotonine	désordres affectifs, dépression…	+/-	intron 2	17 pb	9-12	3	X76754 (843-1018)	6532	17q11.1-q12	(Ogilvie 1996 ; Ito 2002)
IDUA	iduronidase, alpha-L	mucopolysaccharidose de type I	?	intron 2	86 pb	5-7	3	M88001 (1146-1749)	3425	4p16.3	(Scott 1992 ; Gallegos-Arreola 1999)
IGHA1	chaîne alpha1 de l'immunoglobine H	néphropathie à IgA : maladie glomérulaire	+ (agit sur le niveau de transcription)	3' hs1,2 enhancer	52 pb	1-3	3	Y14407 (849-963)	3493	14q32.33	(Pinaud 1997 ; Aupetit 2000 ; Denizot 2001)
IL-1A	Interleukine 1 alpha	maladies infectieuses, autoimmunes, inflammatoires…	? (agit sur le niveau de transcription)	intron 6	46 pb	5-18	6	X03833 (8912-9137)	3552	2q14	(Bailly 1993 ; Bailly 1996)
IL-1RN	antagoniste du récepteur de l'interleukine 1	maladies infectieuses, autoimmunes, inflammatoires…	+	intron 2	86 pb	2-6	5	AY196903 (14659-15002)	3557	2q14.2	(Vamvakopoulos 2002 ; Ma 2002 ; Sehouli 2002)
IL-4	Interleukine 4: cytokine anti-inflammatoire	maladies autoimmunes: arthrite…	+	intron 3	70 pb	2-4	3	AF395008 (19049-19269)	3565	5q31.1	(Mout 1991 ; Buchs 2000)
IL6	Interleukine 6, interféron beta 2: produite dans le système périphérique et le système nerveux central	inflammations du système nerveux central: maladie d'Alzheimer, sclérose en plaques…	+	région flanquante en 3'	9 pb	52-72	6	NT_007819 (avril 2003)	3569	7p21	(Bowcock 1989 ; Murray 1997 ; Bagli 2000 ; Vandenbroeck 2000)
INS	insuline	diabète	+ (agit sur le niveau de transcription)	région flanquante en 5'	14-15 pb	28-213	$\geq$ 38	L15440 (3396-3887)	3630	11p15.5	(Bennett 1995 ; Stead 2000)
KLK14 (C9)	kallikréine : protéase à sérine	troubles des fonctions supérieures, cancer ?	+ (étude préliminaire)	3'UTR	36 pb	8-11	3	NM_022046 (1067-1215 ?)	43847	19q13.3-q13.4	(Yousef 2001)
KLK4 (C3)	kallikréine : protéase à sérine	troubles des fonctions supérieures, cancer ?	+ (étude préliminaire)	3'UTR	36 pb	7-10	3	NM_004917 (970-1175?)	9622	19q13.41	(Yousef 2001)
locus IGH	locus de la partie constante de la chaîne lourde des immunoglobulines		? (proche d'un point de cassure : recombinaison ?)	2 régions homologues: en 5' de IGHA1 et IGHA2	33 pb	25-45	$\geq$ 5	AJ238959	3493 et 3494	14q32.3	(Brusco 1999)
MAOA	monoamine oxydase A	troubles de l'humeur et du tempérament: agressivité, impulsivité…	+ (agit sur le niveau de transcription)	promoteur: 1,2 kb en 5'	30 pb	3-4,5	4	AJ004833 (57-189)	4128	Xp11.4-p11.3	(Sabol 1998 ; Denney 1999 ; Manuck 2000)
MCC	« mutated in colorectal cancers »: gène suppresseur de tumeurs colorectales	cancer du colon	?	région flanquante en 5'	33 pb	5-11	$\geq$ 3	AJ223013	4163	5q21-q22	(Bugert 1998)
MUC5B	mucine 5B: bronchique, salivaire, vésicale, cervicale		? (fixation de facteurs de transcription)	intron 36	59 pb	3-8	5	Y09788 (4000-4451)	4587	11p15.5	(Desseyn 1999 ; Vinall 1998)
NOS3 (eNOS2)	synthèse endothéliale d'acide nitrique (NO)	hypertension, maladies cardiovasculaires	?	intron 2	32 pb	?-38-?	$\geq$2	D26607 (3184-4334)	4846	7q36	(Miyahara 1994)

Nom du gène (nom du minisatellite)	fonction(s) de la protéine	pathologie(s) associée(s) au gène	association du polymorphisme du ms avec la(les) pathologie(s)	localisation du minisatellite dans le gène	taille du motif répété	plage de nombre de copies	nombre d'allèles identifiés	numéro d'accession (position du minisatellite dans la séquence)	Locus ID* (Locus Link)	Chromosome	références
NOS3 (eNOS4)	synthèse endothéliale d'acide nitrique (NO)	hypertension, maladies cardiovasculaires	+	intron 4	27 pb	4-6	3	D26607 (5132-5284)	4846	7q36	(Miyahara 1994 ; Song 2003)
PAH	phénylalanine hydroxylase	phénylcétonurie	+	3 kb en aval (3')	30 pb	6-9	4	S41936	5053	12q22-q24.2	(Goltsov 1992)
PDGFA (IVS3 minisatellite)	facteur de croissance « platelet-derived », chaine A : mitogène		?	intron 3	44 pb	4-12	≥6	AJ238420 (7500-7602)	5154	7p22	(Bonthron 1999)
PKD1	protéine membranaire impliquée dans l'interaction avec la matrice extracellulaire et dans la transduction du signal	polykystose rénale autosomique dominante	?	intron 1	26-27 pb	2-15	≥2	L39891 (10976-11382)	5310	16p13.3	(Burn 1995 ; De Fonzo 1998)
PLA2G4C	phospholipase cytosolique A2-gamma	dans la région du suppresseur de gliome	?	5' UTR	27 pb	1-3	3	NM_003706 (1 copie?)	8605	19q13.3	(Hartmann 2002)
POMT1	protéine-O-mannosyltransferase	syndrome de Walker-Warburg (trouble sévère de la migration neuronale)	?	intron 13	17-19 pb	36-56	8	AF095146 (150-1127)	10585	9q34.1	(Jurado 1999)
PTHRP	« parathyroid hormone-related peptide »	développement : squelettogénèse	?	intron 6 ?	10 à 24 pb	3-17	8	L07553	5744	12p12.1-p11.2	(Pausova 1993)
RB1	rétinoblastome-1 : suppresseur de tumeur	rétinoblastome, ostéosarcome, cancer de la vessie	?	intron 16	53 pb	~10-~35	11	L11910 (123912-125501)	5925	13q14.2	(Scharf 1992)
RECQL4 (IVS12 minisatellite)	« RecQ protein-like 4 »: hélicase	syndrome de Rothmund-Thomson	?	intron 12	30 pb	17-20	>2	NT_037704 (màj avril 2003)	9401	8q24.3	(Roversi 2003)
SEC14L1	« SEC14 S. cerevisiae-like 1 » : rôle possible dans un système de transport intracellulaire		?	3' UTR	13 pb	20-60	>3	NM_003003 (4139-4789)	6397	17q25.1-17q25.2	(Chinen 1996 ; Kalikin 2001)
SNAPC1	« small nuclear RNA activating complex »: active la transcription des gènes de petits ARNs nucléaires		?	5' UTR	17 pb	2-5	4	NT_026437 (màj avril 2003)	6617	14q22	(Maeng 1998)
TMPT	thiopurine méthyltransférase	sensibilité à la 6-mercaptopurine	+ : effet sur le taux de transcription et sur le niveau d'activité de l'enzyme	région flanquante en 5' : promoteur	17 ou 18 pb	3-9	7	AF060074	7172	6p22.3	(Spire-Vayron de la Moureyre 1999 ; Yan 2000)
TPO	thyroide péroxydase: glycoprotéine liée à la membrane, jouant un rôle central dans la fonction de la glande thyroïde	maladies en rapport à la synthèse d'hormones thyroïdiennes: hypothyroïdie congénitale, goitre congénital, déficience de synthèse de l'hormone thyroïdienne, IIA	? (pas d'effet sur l'épissage alternatif)	intron 10	50 pb	4-34	>20	M68651 (1918-2370)	7173	2p25	(Bikker 1992)

*Locuslink : http://www.ncbi.nlm.nih.gov/LocusLink/

1.2.3.3.3.2 Minisatellites appartenant à des séquences codantes

Le Tableau 9 liste les minisatellites polymorphes déjà étudiés appartenant à des séquences codantes, c'est-à-dire générant des répétitions en tandem dans la séquence en acides aminés de la protéine correspondante. La grande majorité des gènes décrits codent pour des glycoprotéines, généralement des protéines sécrétées, constituants de la matrice extracellulaire : aggrécane (cartilage), involucrine (kératocytes), mucines (muqueuses). Le collagène est également une protéine de la matrice extracellulaire contenant des répétitions en tandem : cette famille de protéines ne figure pas dans le Tableau 9 car le motif répété qui la caractérise, glycine-*X*-*Y*, est de la classe des microsatellites. Les répétitions en tandem d'acides aminés, qui occupent en général la majeure partie de la longueur de ces macro-protéines, jouent un rôle important dans leur structure fonctionnelle, en étant le site de glycosylations (attachement des glucosaminoglycanes pour l'aggrécane, O-glycosylations pour les mucines). Les mucines, protéines exprimées par les cellules épithéliales, sécrétées ou membranaires, entrent dans la composition des mucus. Elles constituent une famille de protéines contenant des répétitions en tandem polymorphes d'acides aminés riches en sérine et en thréonine, qui sont le site d'O-glycosylations (Perez-Vilar 1999 ; Kinarsky 2003). Les répétitions en tandem couvrent 50% ou plus de la protéine et sont plus ou moins polymorphes, pouvant faire varier du simple au double la taille de la protéine (Vinall 1998 ; Debailleul 1998). Ces protéines ont un intérêt médical car elles sont souvent surexprimées par les cellules cancéreuses ou lors de réactions inflammatoires. De nombreuses autres protéines contiennent des domaines d'O-glycosylation, appelés « mucin-like domains » : par exemple, GP1BA et PSGL-1 pour lesquelles un polymorphisme au niveau des répétitions en tandem a également été caractérisé (Tableau 9).

Un autre intérêt des minisatellites codants réside dans le fait que certaines répétitions en tandem d'acides aminés ont la particularité d'être immunogènes (Mollick 2003) : des épitopes répétitifs stimulent la production d'immunoglobulines.

Tableau 9 : Minisatellites polymorphes appartenant à des séquences codantes.

Nom du gène	fonction(s) de la protéine	pathologie(s) associée(s) au gène	association du polymorphisme du ms avec la(les) pathologie(s)	localisation du minisatellite dans le gène	taille du motif répété	Plage de nombre de copies	nombre d'allèles identifiés	numéro d'accession (position du minisatellite dans la séquence)	Locus ID* (LocusLink)	Chromosome	références	répétition d'acides aminés
AGC1	aggrécane : protéoglycane, constituant majeur du cartilage	scoliose	-	exon G3	57 pb	13-33	13	M55172 (2866-4495)	176	15q26.1	(Doege 1997 ; Zorkol'tseva 2002)	19aa:PGVEDISGLPSGEVLETAA
CEL	carboxyl ester lipase: rôle dans l'absorption intestinale du cholestérol et des vitamines liposolubles	athérosclérose	?	exon 11	33pb	13-18	6	NM_001807 (1693-2218)	1056	9q34.3	(Higuchi 2002)	11aa: PVPPTGDSGAP
DRD4	récepteur de la dopamine	troubles de la personnalité, hyperactivité; recherche de nouveauté (dépendance aux drogues)…	+/-	exon 3	48 pb	1-10	9	L12398 (744-1080)	1815	11p15.5	(Lichter 1993 ; Swanson 2000 ; Li 1997)	16aa: LPQDPCGPDCAPPAPG
GP1BA	glycoprotéine Ib alpha des plaquettes (« mucin like »)	thrombose artérielle	+	exon 2	39 pb	1-3	3	S34439	2811	17pter-p12	(Lopez 1992 ; Gonzalez-Conejero 1998 ; Jilma-Stohlawetz 2003)	13aa: EPTSEPAPSPTTP
IVL	involucrine: précurseur de l'enveloppe cornée des kératocytes différenciés		?	exon 3	30 pb	?-39-?	>8	NM_005547 (398-1567)	3713	1q21	(Simon 1991 ; Urquhart 1993)	10aa: ELPEQQEGQL
MUC1	mucine 1: protéine membranaire (face apicale des cellules épithéliales sécrétoires), antigène associé au cancer du sein	carcinogenèse	+	exon 2	60 pb	20-120	>30	M61170 (3821-3881)	4582	1q21	(Gendler 1988 ; Engelmann 2001 ; Silva 2001 ; Kinarsky 2003)	20aa: PDTRPAPGSTAPPAHGVTSA
MUC2	mucine 2: mucine intestinale, trachéale et bronchique	asthme atopique	+	exon	69 pb	51-115	>20	NM_002457(5713-12612)	4583	11p15.5	(Gum 1989 ; Toribara 1991 ; Vinall 2000)	23aa: PTTTPITTTTTVTPTPTPTGTQT
MUC3	mucine 3: mucine intestinale	colite ulcérante	+	exon	51 pb	? (9,38-18kb)	>30	AF113616 (2-1025)	4584	7q22	(Gum 1990 ; Van Klinken 1997)	17aa: HSTPSFTSSITTTETTS
MUC4	mucine 4: trachéale et bronchique, associée à la membrane	surexprimée dans les tumeurs pancréatiques	?	exon 2	48 pb	~145-395	> 30	NM_018406 (3393-3485: 2 copies)	4585	3q29	(Nollet 1998 ; Debailleul 1998 ; Vinall 2000)	16aa: ATPLPVTSTSSASTGH

Nom du gène	fonction(s) de la protéine	pathologie(s) associée(s) au gène	association du polymorphisme du ms avec la(les) pathologie(s)	localisation du minisatellite dans le gène	taille du motif répété	Plage de nombre de copies	nombre d'allèles identifiés	numéro d'accession (position du minisatellite dans la séquence)	Locus ID* (LocusLink)	Chromosome	références	répétition d'acides aminés
MUC5AC	mucine 5AC: bronchique, gastrique; exprimée anormalement dans les adénomes colorectaux, cancers pancréatiques	désordres respiratoires liés à l'hypersécrétion	?	exon	24 pb	? (? +/- 0,5kb)	4	AJ298318 (1-4047)	4586	11p15.5	(Escande 2001 ; Vinall 1998 ; Vinall 2000)	8aa: TTSTTSAP
MUC5B	mucine 5B: bronchique, salivaire, vésicale, cervicale		?	exon	87 pb	?(15,5kb-19,5kb)	3	Z72496 (2188-10584)	4587	11p15.5	(Vinall 1998 ; Desseyn 1997)	29aa: SSTPGTTWILTELTTAATTTAGTGP TATP
MUC6	mucine 6, mucine gastrique : protection du tractus gastro-intestinal contre acides, protéases, microorganismes pathogènes, traumatismes mécaniques…	prédisposition au développement de carcinomes gastriques	+	exon	507 pb	8-13,5kb	>12	U97698 (1-507: 1 copie)	4588	11p15.5	(Toribara 1993 ; Garcia 1997 ; Vinall 1998)	169aa: SPFSSTGPMTATSFQTTTTYPTPSH PQTTLPTHVPPFSTSLVTPSTGTVIT PTHAQMATSASIHSTPTGTIPPPTTL KATGSTHTAPPMTPTTSGTSQAHS SFSTAKTSTSLHSHTSSTHHPEVTP TSTTTITPNPTSTGTSTPVAHTTSAT SSRLPTPFTTHSPPTGS
MUC7	mucine 7: mucine salivaire	se lie avec les bactéries orales, rôle dans l'asthme	+	exon 3	69 pb	5-8	3	NM_152291 (534-1009)	4589	4q13-q21	(Bobek 1993 ; Kirkbride 2001)	23aa: TTAAPPTPPATTPAPPSSSAPPE
OGFR	récepteur de facteur de croissance opioïde: module la prolifération cellulaire et l'organisation des tissus lors du développement. (« mucin like »)	rôles dans le cancer, le renouvellement cellulaire, la cicatrisation, et le rejet de tumeurs dirigé par l'antigène de l'angiogénèse	?	exon	60 pb	0-8	5	NM_007346 (1576-1996)	11054	20q13.3	(Zagon 2000)	20aa: SPSETPGPRPAGPAGDEPAE
PRNP	prion (glycoprotéine membranaire ayant tendance à former des agrégats)	maladie de Creutzfeldt-Jakob héréditaire	+	exon	24 pb	4-13	≥ 11	NM_000311(260--372)	5621	20p12	(Campbell 1996 ; Laplanche 1995)	8 aa: GGGWGQPH
PSGL-1	« P-selectin glycoprotein ligand 1 (mucin like) »: rôle dans la liaison plaquettes-neutrophilles (liaison à la sélectine sur l'endothélium des vaisseaux sanguins)	accidents vasculaires cérébraux ischémiques	+	exon 2	30 pb	14-16	3	U25956 (590-1063)	6404	12q24	Afshar-Kharghan, 2001 ; Lozano, 2001	10aa: ATEAQTTPPA

*Locuslink : http://www.ncbi.nlm.nih.gov/LocusLink/

1.2.3.3.3.3 Localisation chromosomique des gènes associés à des minisatellites

La Figure 7 présente la localisation des 59 gènes associés à des minisatellites polymorphes : dans leur voisinage, leurs introns (Tableau 8) ou leurs exons (Tableau 9). Cette répartition est biaisée vers les extrémités des chromosomes. En particulier, l'extrémité 11p15.5 est riche en gènes associés à des minisatellites : DRD4, HRAS, INS, mucines MUC2, MUC5AC, MUC5B, MUC6. Il a été montré que ce cluster de mucines dérive probablement d'une série de duplications (Desseyn 1998), et que cette région du chromosome 11 est riche en recombinaisons méiotiques (Pigny 1996), ce qui peut expliquer la présence de nombreux minisatellites fortement polymorphes. En effet, comme nous le verrons au paragraphe 1.2.4.2., l'instabilité des minisatellites (à l'origine de leur polymorphisme) peut provenir d'événements de recombinaison méiotique, impliquant des cassures double-brin.

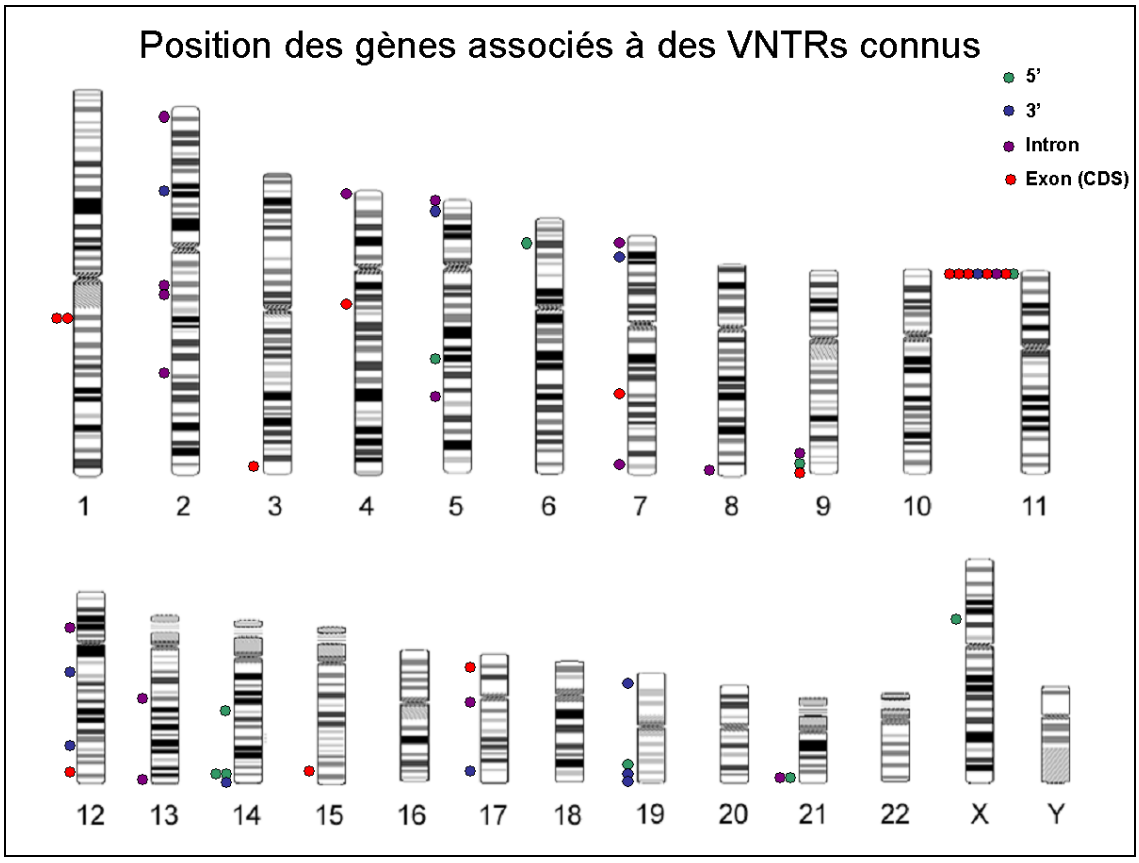


Figure 7 : Localisation chromosomique des gènes associés à des minisatellites polymorphes étudiés.

1.2.3.4 Intérêts des minisatellites hypermutables

Les minisatellites hypermutables humains, décrits en détail dans la revue présentée au chapitre 2.3.1 (Vergnaud & Denoeud 2000), sont des structures très instables dont le taux de mutation dans la lignée germinale excède 0,5% : cette définition correspond au seuil de détection imposé par les études de pedigrees. Ces répétitions en tandem ont été beaucoup étudiées, afin de comprendre les mécanismes qui provoquent leur instabilité. Comme nous le

verrons au paragraphe 1.2.4.2.1, ces minisatellites mutent par des mécanismes complexes probablement initiés par des points chauds de cassures double-brin. Ils ont donc un intérêt fondamental pour l'étude de ces points chauds, mal connus chez l'homme, ainsi que des mécanismes de réparation des cassures double-brin.

Par ailleurs, les minisatellites hypermutables s'avèrent être des outils de choix pour l'étude de l'effet d'agents génotoxiques, en particulier les faibles doses de radiations ionisantes, car leur fort taux de mutation spontanée permet de détecter des changements de taux de mutation dans des échantillons de population relativement petits. Chez l'homme, ces minisatellites ont permis de détecter une augmentation du taux de mutation dans des populations vivant dans des zones contaminées comme la région de Tchernobyl, et donc soumises à une exposition chronique à des faibles doses de radiations (Dubrova 1996 ; Dubrova 1997 ; Dubrova 2002 ; Dubrova 2002). Ces observations sur l'effet des radiations ionisantes ont été étendues aux souris (Dubrova 1993 ; Dubrova 1998), oiseaux (Ellegren 1997), et plantes (Kovalchuk 2000). D'autres agents cytotoxiques induisent également une augmentation de l'instabilité des minisatellites, comme montré par des études menées chez le goéland (Yauk 1996) et la souris (Hedenskog 1997). Cependant, de façon contradictoire, aucune augmentation substantielle du taux de mutation des minisatellites n'a pu être corrélée chez l'homme avec des hautes doses de radiations ionisantes, reçues de façon ponctuelle (exposition aiguë). Cette observation provient d'études menées sur les descendants de personnes ayant subi une radiothérapie testiculaire (Armour 1999 ; May 2000) ou ayant survécu à une exposition aiguë, lors de l'explosion de bombes atomiques à Hiroshima et Nagasaki (Kodaira 1995), comme du nettoyage de la centrale nucléaire de Tchernobyl après l'accident de 1986 (Livshits 2001). Ce phénomène pourrait s'expliquer par un rôle du temps d'exposition aux radiations. Les dommages causés par une irradiation ponctuelle pourraient être réparés (les minisatellites ne seraient alors pas instables), tandis que les dommages provenant d'une exposition chronique à des faibles doses de rayonnements ne seraient pas réparés (ce qui générerait une instabilité des minisatellites) (May 2000). Par ailleurs, les radiations ionisantes pourraient avoir un effet sur les cellules germinales en cours de maturation, par exemple lors de la méiose, mais pas sur les cellules souches de la lignée germinale. Il n'y aurait donc pas d'effet de l'exposition aiguë sur des descendants conçus plusieurs années plus tard : la mise en évidence de mutations nécessiterait alors l'étude de grands échantillons d'enfants conçus juste après l'exposition (Livshits 2001).

1.2.4 Mécanismes de mutation des répétitions en tandem

1.2.4.1 Les microsatellites

Le premier mécanisme de mutation des microsatellites suggéré suit le modèle de glissement et mésappariement ou SSM (« slipped-strand mispairing ») (Levinson 1987). Ce phénomène, se produisant lors de la réplication, conduit à l'augmentation ou à la réduction du nombre de copies selon le brin subissant le mésappariement, comme illustré par la Figure 8.

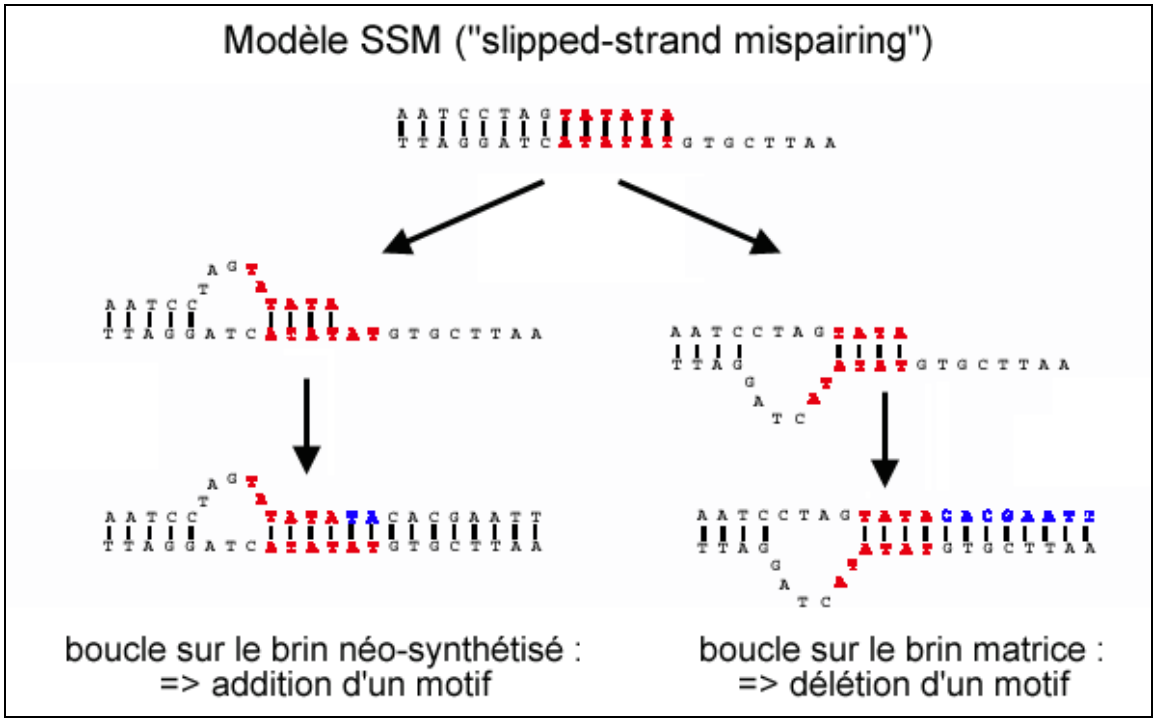


Figure 8 : Modèle de mutation des répétitions en tandem par glissement lors de la réplication.

Certaines de ces erreurs de réplication sont corrigées par l'activité de relecture de l'exonucléase et par le système de réparation des mésappariements, mais d'autres échappent à ces réparations et deviennent des mutations. Par exemple, chez la levure, les structures CTG/CAG ou CGG/CCG forment des structures secondaires qui échappent au système de réparation de l'ADN (Moore 1999). L'instabilité des microsatellites correspondrait donc à un équilibre entre la génération d'erreurs de réplication par glissement et la correction de certaines de ces erreurs par les systèmes de réparation, comme le système de réparation des mésappariements et, à un degré moindre, l'activité de relecture de l'exonucléase (Kruglyak 1998). Des mutations dans des gènes du système de réparation des mésappariements (MMR) sont à l'origine d'une instabilité des microsatellites chez *Escherichia coli* (Levinson 1987), *Saccharomyces cerevisiae* (Strand 1993) et l'homme (Boyer 1995). Différentes études génétiques menées chez la levure ont montré que le système de réparation des

mésappariements ne corrige efficacement que les boucles de 1 à 14 nucléotides mais pas celles de plus de 16 nucléotides (Tran 1996 ; Sia 1997), ce qui démontre que les mécanismes de mutation des répétitions en tandem ne peuvent se réduire au « SSM », du moins pour les unités répétées de taille supérieure à 16 pb (minisatellites).

Des études menées sur différentes espèces ont montré que le taux de mutation des microsatellites augmente avec leur nombre de répétitions, entre locus de longueur moyenne différente (Weber 1990 ; Brinkmann 1998), mais également entre différents allèles du même locus (Wierdl 1997 ; Schlotterer 1998 ; Ellegren 2000b). Pour expliquer cette observation, on peut émettre l'hypothèse que le glissement lors de la réplication soit susceptible de survenir plus fréquemment si le microsatellite compte davantage d'unités répétées. D'autre part, lorsque des mutations ponctuelles surviennent dans les microsatellites, conduisant à une répétition imparfaite, les locus sont moins variables (Petes 1997), ce qui correspondrait à un taux de mutation par glissement de la réplication inférieur dans les répétitions en tandem interrompues (Jin 1996). En effet, il semble que la présence de variations internes stabilise ces structures car elle favorise la reconnaissance par le système de réparation des mésappariements des régions où un glissement a eu lieu lors de la réplication (Strand 1993). Le taux de mutation des microsatellites dépendrait donc de la taille de la plus grande répétition non interrompue, et non pas du nombre de répétitions total, dégénérées ou non.

Des études structurales ont montré que les microsatellites (en particulier les triplets CNG) simple-brin peuvent former des structures secondaires en épingle à cheveu (Mitas 1997 ; McMurray 1999). En outre, lors de la réplication de l'ADN, des régions d'ADN simple-brin ont plus de chances de se former lors de la synthèse du brin retardé que lors de la synthèse du brin direct, ce qui peut générer des structures secondaires à l'origine d'erreurs de réplication. En effet, l'ADN double-brin étant antiparallèle, l'un des deux brins (brin direct) est synthétisé de façon continue au fur et à mesure que la fourche de réplication avance, mais l'autre s'allonge dans le sens opposé à la progression de la fourche (brin retardé). Sur ce brin, des petits fragments d'ADN, appelés fragments d'Okazaki, sont synthétisés à partir d'amorces d'ARN, qui sont ensuite éliminées par la RNAse H au fur et à mesure de la réplication, puis les fragments d'ADN sont reliés pour achever la synthèse. Pendant la réplication du brin retardé, l'extension de l'amorce peut continuer alors que l'ARN initiateur est en cours de destruction. Le déplacement d'une amorce en aval peut alors résulter en la formation d'une structure « flap » (extrémité 5' battante). Gordenin et collègues ont suggéré que la présence d'une répétition en tandem formant une structure secondaire dans le flap pourrait conduire à une expansion (Gordenin 1997). L'enzyme FEN1 chez l'homme (Li 1995 ; Bambara 1997), homologue de Rad27 chez la levure, est impliquée dans la résolution de ces structures. Il s'agit d'une exonucléase 5'-3', qui est également nécessaire à l'élimination du dernier ribonucléotide en 5' des fragments d'Okazaki. Chez la levure, la fréquence d'insertion/délétion dans les microsatellites augmente dans un contexte mutant pour Rad27 (Schweitzer 1998 ; Freudenreich 1998 ; Kokoska 1998).

Kokoska *et al.* ont proposé un modèle expliquant l'effet de Rad27 sur l'instabilité des répétitions, qui fait intervenir des glissements (Kokoska 1998) : Figure 9. Le maintien du dernier ribonucléotide en 5' du fragment d'Okazaki (représenté par l'astérisque) retarderait la ligation, ce qui favoriserait la formation d'une structure « flap ». Cette structure serait reconnue par l'exonucléase 5'-3' de la polymérase qui générerait un gap adjacent au « flap ». La réassociation du « flap » avec cette région simple-brin pourrait alors conduire à une addition de répétitions en tandem (par glissement) : Figure 9, gauche. Le maintien du dernier ribonucléotide pourrait aussi conduire à un blocage de la synthèse d'ADN et une activation de l'activité de relecture de l'exonucléase sans passer par la structure flap. Pendant la synthèse d'ADN, un glissement sur le brin matrice conduirait alors à une délétion de motifs : Figure 9, droite.

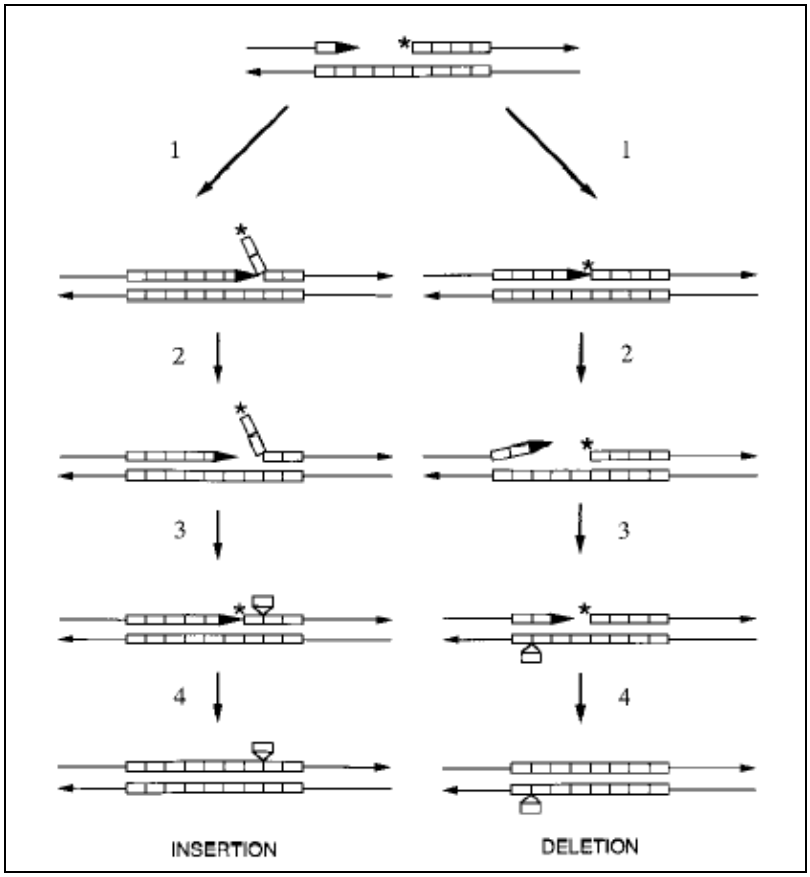


Figure 9 : Modèle de mutation des répétitions en tandem dans un contexte mutant pour Rad27 (où le dernier ribonucléotide en 5' des fragments d'Okazaki est maintenu), d'après (Kokoska 1998).

Chez l'homme, il a été montré que la résolution par FEN-1 des structures « flap » contenant des répétitions de triplets dépend de leur taille - seules les répétitions de moins de 11 CAG/CTG sont clivées - et de leur orientation, ce qui pourrait refléter la différence de stabilité entre les structures en épingles à cheveu constituées de triplets CAG ou CTG - (Lee 2002).

Des mécanismes impliquant la recombinaison ont aussi été proposés pour la mutation des microsatellites : des phénomènes de crossing-over inégal ou de conversion génique pourraient faire varier la taille de ces structures (Brohede 1999 ; Jakupciak 2000 ; Parniewski 2002).

Cependant (d'après Ellegren 2000a), de plus en plus de preuves penchent en faveur du mécanisme de glissement de la réplication par rapport aux phénomènes de recombinaison :

- Les allèles mutants sont en général non-recombinants au niveau des marqueurs flanquants, ce qui correspondrait à des événements de mutation intra-alléliques (Morral 1991 ; Mahtani 1993).
- Le taux de mutation des microsatellites ne diffère pas entre les chromosomes sexuels et les autosomes, ce qui suggère que les événements de mutation ne nécessitent pas le contact entre chromosomes homologues (Kayser 2000 ; Heyer 1997).
- Le type de mutations rencontrées est cohérent avec le mécanisme de glissement de la réplication : augmentation du taux de mutation avec la taille et l'homogénéité. De plus, des études *in vitro* montrent que les microsatellites ont la capacité intrinsèque à subir ce glissement (Schlotterer 1992).
- Chez la levure, la stabilité des microsatellites n'est pas affectée par des mutations dans les gènes de recombinaison (Henderson 1992) mais elle est fortement réduite par des mutations dans les gènes affectant le système de réparation des mésappariements (Strand 1993), ce qui est également le cas chez *E. coli* (Levinson 1987).

Des mécanismes de mutation des microsatellites impliquant simultanément le glissement et la recombinaison sont également envisageables. En effet, dans certains cas, une instabilité plus importante des microsatellites a été observée dans la lignée germinale (Cohen 1999 ; Seznec 2000). Lors de la recombinaison entre deux chromosomes homologues, la jonction de Holiday, structure 4-brins, contient des régions d'ADN hétéroduplex (avec des mésappariements). Ces régions subissent une correction dépendante de la réplication, qui fait intervenir des cassures double-brin, et au cours de laquelle des phénomènes de glissement peuvent survenir et générer de nouveaux allèles (Gendrel 2000 ; Richard 2000 ; Jankowski 2000).

Aux locus microsatellites, les insertions/délétions d'une seule unité sont les plus fréquentes (Henderson 1992), mais pas les seules. La fréquence d'insertion/délétion de plus d'une unité, en général de 2 à 5, varie selon les espèces ; elle est de 4 à 14% pour l'homme (Kayser 2000 ; Brinkmann 1998 ; Xu 2000). De plus, les mutations survenant dans les microsatellites sont biaisées vers une augmentation de taille (Amos 1996 ; Yamada 2002) : ce phénomène est appelé « évolution directionnelle » (Amos 1996). Cependant, les microsatellites excèdent rarement quelques dizaines de répétitions, ce qui laisse supposer l'existence de facteurs empêchant leur croissance indéfinie:

- Plusieurs études ont montré que les délétions d'unités répétées étaient plus fréquentes ou de plus grande taille plus le nombre de copies était grand (Wierdl 1997 ; Parniewski 2002 ; Ellegren 2000b ; Harr 2000). La taille des microsatellites pourrait donc correspondre à un équilibre où les taux d'expansion et de contraction seraient égaux (Xu 2000). Ces observations nécessitent encore d'être généralisées aux différents types de répétitions et espèces (Ellegren 2000a).
- Une autre hypothèse expliquant la limitation de l'accroissement de taille des microsatellites serait un équilibre entre le processus de mutation biaisée en faveur de l'accroissement de taille et les mutations ponctuelles qui tendent à faire diminuer cette taille en ralentissant les expansions ultérieures (Kruglyak 1998). Ce modèle pourrait expliquer les différences de taille des microsatellites entre organismes, qui dépendraient du taux de mutations ponctuelles et du taux de glissement réplicatif. On prédirait alors des répétitions plus courtes dans les génomes avec de faibles taux de glissement, ce qui est le cas pour *Drosophila melanogaster* (Schlotterer 1998). De façon intéressante, le taux de glissement ne serait pas faible parce que les répétitions sont courtes mais les répétitions seraient courtes à cause du faible taux de glissement (Ellegren 2000a).
- Enfin, la dernière hypothèse pourrait être une action de la sélection naturelle sur la taille des microsatellites. Les grands allèles pourraient être contre-sélectionnés, ce qui introduirait un plafond pour la longueur des répétitions (Garza 1995). La sélection naturelle pourrait aussi introduire des limites de taille supérieure et inférieure (Li 2000). Des pressions de sélection très différentes semblent par ailleurs agir sur les régions non-codantes en 5' et en 3' de gènes et les régions codantes, comme suggéré par l'étude de Morgante *et al.* (Morgante 2002) sur le génome *d'A. thaliana*.

1.2.4.2 Les minisatellites

Plusieurs catégories de minisatellites méritent d'être distinguées, en tout cas chez l'homme, où ont été découverts des minisatellites dits hypermutables. Ces derniers ont la particularité d'être extrêmement instables. Leur taux de mutation en lignée germinale dépasse 0,5%, seuil de détection imposé par l'étude de pedigrees (pour revue, voir : Jeffreys 1999 ; Vergnaud 2000), et peut même atteindre jusqu'à 20%, pour le minisatellite CEB1 (Vergnaud 1991) en méiose mâle (Buard 1998). Les minisatellites hypermutables n'ont pas été découverts dans d'autres espèces à ce jour (Bois 2003), mais les minisatellites hypermutables humains ont été largement étudiés, y compris dans des organismes modèles (c'est-à-dire la souris, et surtout la levure *Saccharomyces cerevisiae* (Appelgren 1997 ; Debrauwère 1999; Paques 2001; Lopes 2002)), afin d'appréhender leurs mécanismes de mutation. Ils seront détaillés dans le paragraphe suivant.

1.2.4.2.1 Minisatellites hypermutables

La première caractéristique des mutations survenant au niveau des minisatellites hypermutables est qu'elles n'ont pas la même nature ni la même fréquence entre les lignées

somatique et germinale. Ces mutations sont beaucoup plus fréquentes dans les cellules germinales que dans les cellules somatiques, comme montré par des études menées sur les minisatellites MS31A, MS32 (Jeffreys 1994 ; Jeffreys 1997), MS205 (May 1996), CEB1 (Buard 2000) et B6-7 (Tamaki 1999). De plus, les mutations survenant dans les cellules somatiques correspondent à des événements intra-alléliques relativement simples par comparaison aux réarrangements complexes survenant dans les cellules germinales (Jeffreys 1997 ; Buard 2000). Ces observations suggèrent que la déstabilisation de ces minisatellites s'opère par des voies différentes dans les lignées somatique et germinale. Il faut cependant noter qu'un même mécanisme, comme la réparation d'une cassure double brin se produisant entre chromatides sœurs, dans les cellules somatiques, ou entre chromosomes homologues, dans les cellules germinales, pourrait parfaitement expliquer la création de ces deux types d'allèles mutants.

1.2.4.2.1.1 Historique : les deux premiers modèles de mutation proposés

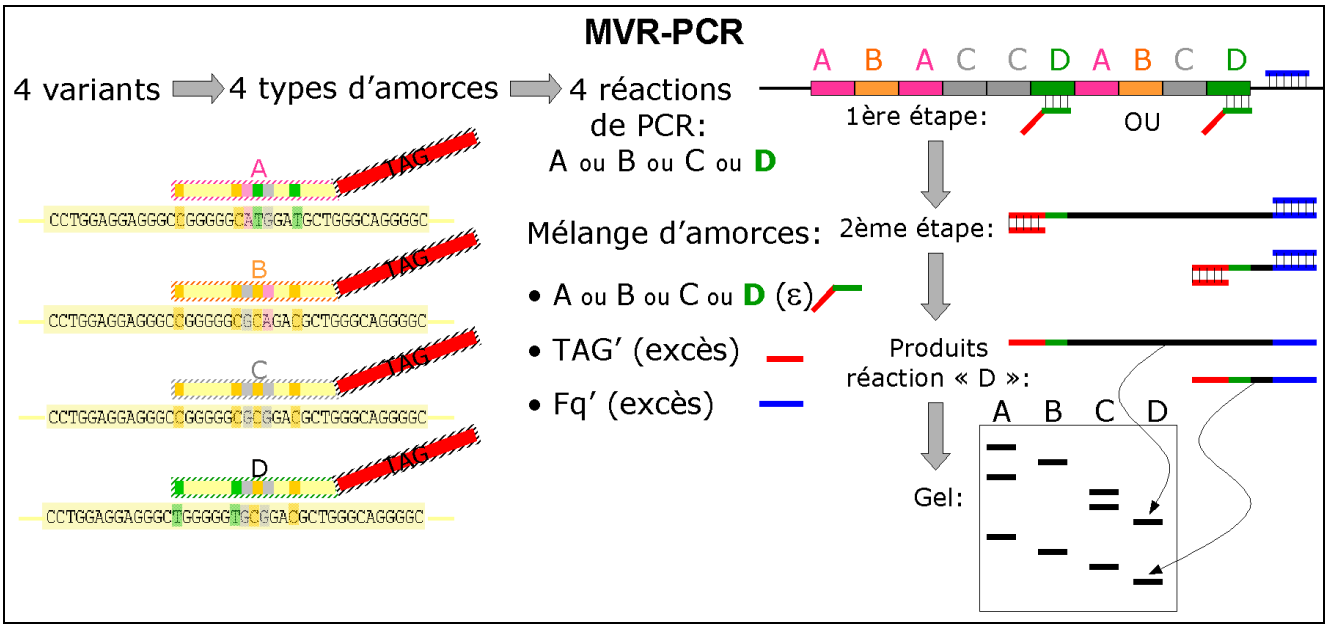


Figure 10 : Principe de la technique MVR-PCR.

En 1994, deux études majeures ont été menées sur les minisatellites humains CEB1 (Buard 1994) et MS32 (Jeffreys 1994), en mettant à profit la technique appelée « MVR-PCR » (« minisatellite variant repeat -PCR» : voir Figure 10). Cette technique, développée par Jeffreys et collègues (Jeffreys 1991), permet de typer les positions variantes dans les allèles de minisatellites en utilisant des amorces spécifiques. En effet, tous les minisatellites hypermutables caractérisés jusqu'à présent (leur liste figure dans la revue présentée au chapitre 2.3.1 (Vergnaud 2000)) possèdent des variants internes, ce qui permet de caractériser les mutations survenues.

Ces études d'allèles mutants dans les cellules germinales ont mis en évidence des événements de recombinaison complexes, inter-alléliques (de type conversion génique) et intra-alléliques,

et ont conduit à l'hypothèse que ces mutations pourraient être initiées par des cassures double-brin et impliquer un intermédiaire d'ADN hétéroduplex, suite à l'invasion d'un brin du chromosome homologue : Figure 11.

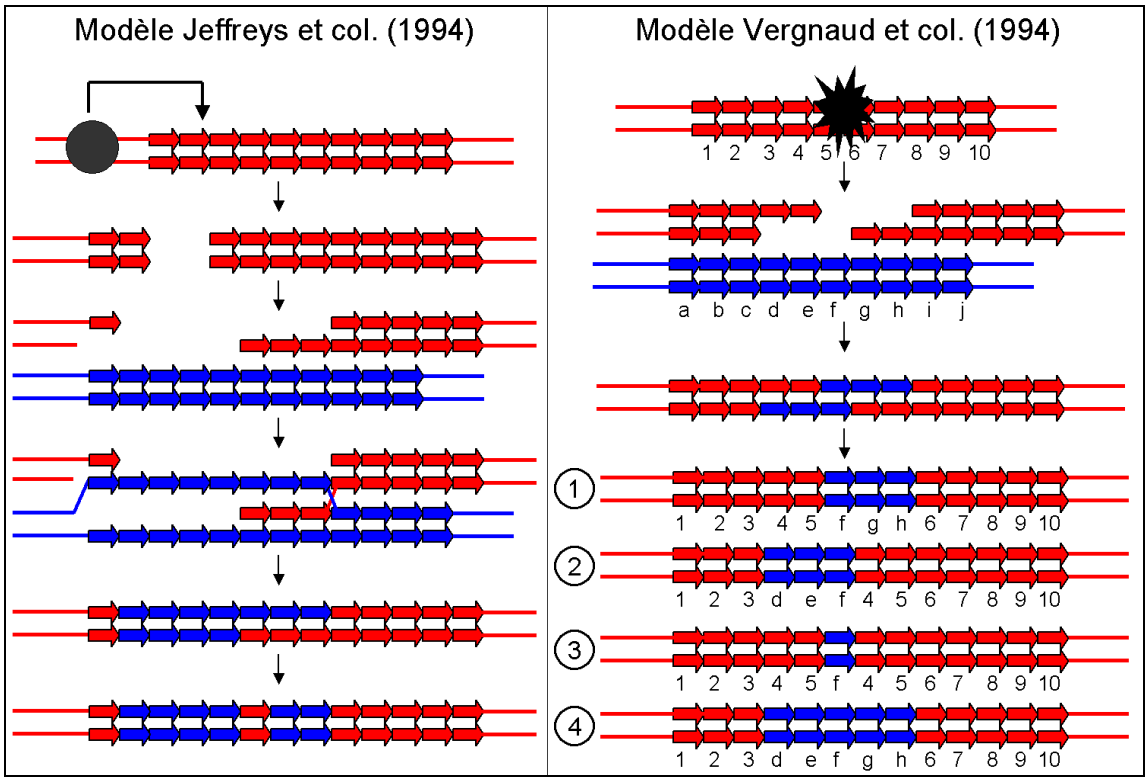


Figure 11 : Deux mécanismes de mutation méiotique des minisatellites hypermutables proposés en 1994 par (Buard 1994) et (Jeffreys 1994).

La différence majeure entre ces deux modèles réside dans la localisation du point chaud de cassure double-brin. Selon Jeffreys et collègues (Jeffreys 1994), il se situerait dans la flanquante. Ils ont en effet observé, pour les minisatellites MS32, MS31A et MS205, que les mutations surviennent plus fréquemment vers une extrémité de la répétition que vers l'autre, ce qui laisse présager d'un effet de la flanquante sur l'instabilité des minisatellites. De plus, une mutation ponctuelle dans la flanquante du minisatellite MS32 est associée à une forte réduction de l'instabilité du minisatellite (Monckton 1994). Selon le modèle proposé par Vergnaud et collègues (Buard 1994), la cassure double-brin surviendrait dans la répétition en tandem elle-même, avec la possibilité d'un décalage entre les cassures des deux brins. Ce détail était nécessaire pour rendre compte de la mise en évidence au locus CEB1 d'événements de conversion génique associés à la duplication de motifs de part et d'autre de l'insertion (produit 3 dans la Figure 11). Ces deux modèles, même s'ils sont similaires du point de vue des mécanismes (cassure double-brin, intermédiaire hétéroduplex), correspondent à deux conceptions des minisatellites hypermutables : dans l'un des cas, la séquence du minisatellite n'aurait pas de rôle sur l'instabilité, qui serait générée par la séquence flanquante ou les caractéristiques de la chromatine environnante, et dans l'autre cas, l'hypermutabilité serait liée à la structure du minisatellite.

1.2.4.2.1.2 Présentation du mécanisme de mutation proposé dans la revue (Vergnaud & Denoeud, 2000)

Pendant mon DEA, j'ai participé à la rédaction de la revue intitulée « Minisatellites : mutability and genome architecture » (Minisatellites : mutabilité et architecture des génomes) (Vergnaud & Denoeud 2000), qui sera présentée plus en détail au chapitre 2.3.1, car elle propose, outre une vue d'ensemble sur les minisatellites hypermutables humains, de nouveaux résultats concernant la distribution des répétitions en tandem dans les génomes eucaryotes.

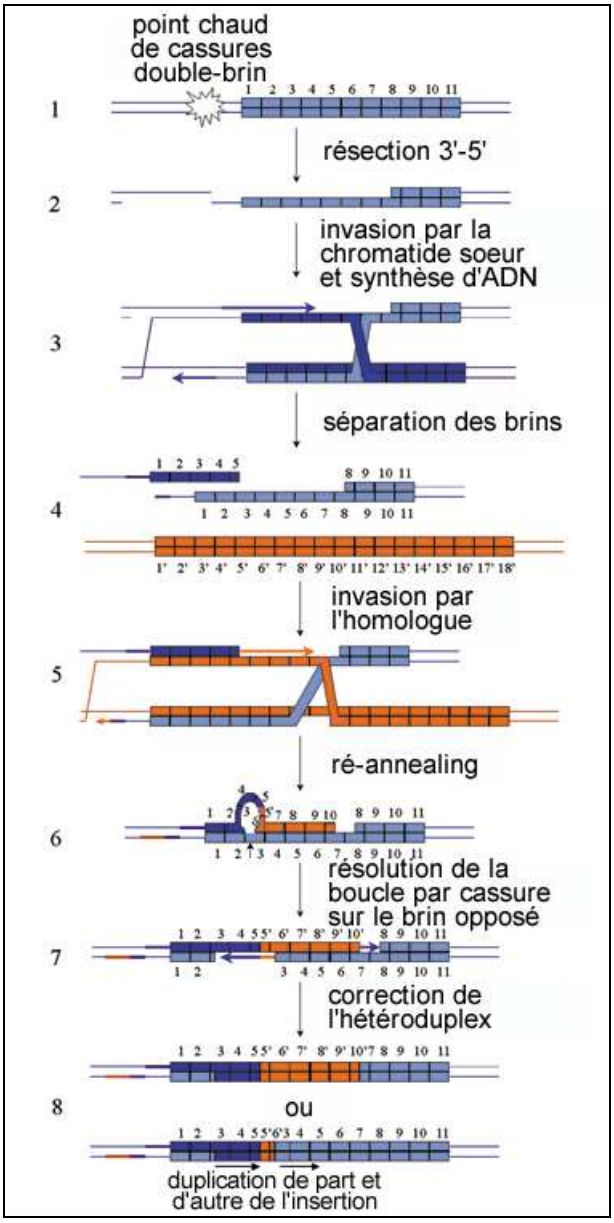


Figure 12 : Mécanisme de mutation méiotique des minisatellites hypermutables proposé en 2000, dans la revue (Vergnaud & Denoeud 2000).

Cette revue propose un modèle de mutation des minisatellites hypermutables qui prend en compte toutes les observations faites jusque là. L'avancée essentielle par rapport à 1994 provient d'une étude menée sur le minisatellite humain CEB1 inséré dans l'organisme modèle

Saccharomyces cerevisiae. Cette étude a montré que l'instabilité méiotique dépendait du site d'insertion (Debrauwère 1999) : le minisatellite est beaucoup plus instable lorsqu'il est inséré à proximité d'un point chaud de cassures double-brin, et cette instabilité dépend des cassures double-brin induites par l'enzyme Spo11. Le déplacement d'un minisatellite de levure (non instable) à proximité d'un point chaud de cassures double-brin permet également de générer une instabilité méiotique (Bishop 2000). Ces observations ont conduit au modèle dans lequel une cassure double-brin générée dans la flanquante du minisatellite, comme initialement proposé par Jeffreys et collègues, s'étend dans la répétition en tandem. Ensuite, l'invasion par le chromosome homologue peut conduire à des phénomènes de conversion génique, impliquant un intermédiaire hétéroduplex, comme cela a été mis en évidence dans certains allèles mutants lors de l'étude de Debrauwère *et al.* (Debrauwère 1999). Ce modèle, présenté sur la Figure 12, propose deux épisodes successifs d'invasion suite à la cassure double-brin : par la chromatide sœur puis par le chromosome homologue. Plusieurs étapes d'invasion pourraient expliquer les réorganisations extrêmement complexes observées fréquemment pour le minisatellite B6.7, et qui ne pouvaient pas être expliquées par les modèles proposés en 1994 (Tamaki 1999). Le mécanisme faisant intervenir l'invasion par la chromatide sœur peut en outre être invoqué en situation haploïde, par exemple pour le minisatellite MSY1, situé sur le chromosome Y humain. L'instabilité de MSY1 est estimée à 5% dans les cellules germinales (mâles, exclusivement, étant donné sa localisation sur le chromosome Y) (Vogt 1997). La comparaison des allèles de ce minisatellite suggère un mécanisme de mutation provoquant la diffusion des variants de façon linéaire, probablement par glissement de la réplication dans les groupes de répétitions homogènes (Jobling 1998). On observe en outre un processus d'homogénéisation des répétitions (une substitution dans un motif va se répercuter tout le long de la répétition). Il pourrait être causé par la réparation biaisée des mésappariements dans de l'ADN hétéroduplex généré par des mécanismes de glissement ou d'échange inégal entre chromatides sœurs (Bouzekri 1998). La séquence de ce minisatellite, riche en AT, a tendance à former des structures en épingle à cheveux et pourrait donc être impliquée directement dans les phénomènes de mutation en cause.

Le modèle présenté sur la Figure 12 réconcilie les deux modèles initiaux de 1994 : il permet de créer les événements de conversion génique avec duplications de motifs de part et d'autre, tout en initiant la cassure dans la flanquante du minisatellite.

1.2.4.2.1.3 Evolutions récentes de la compréhension des mécanismes de mutation

Le modèle proposé dans la revue (Vergnaud & Denoeud 2000), présenté au paragraphe précédent, reste toujours de mise pour expliquer les mécanismes de mutation méiotique des minisatellites hypermutables humains. Cependant, d'autres avancées ont été faites depuis, en particulier en ce qui concerne les mutations somatiques.

Il s'avère que des mutations dans les protéines Rad27 (enzyme impliquée dans la résolution des structures « flap » générées sur le brin retardé lors de la synthèse des fragments

d'Okazaki : voir paragraphe 1.2.4.1), Pol3-t (DNA polymérase δ (Kokoska 1998)), et PCNA (impliquée dans la réplication et la réparation de l'ADN (Kokoska 1999)), déstabilisent par des mécanismes apparemment différents, un minisatellite de 3 fois 20 pb (à un degré moindre que des microsatellites). Par ailleurs, une étude récente (Lopes 2002) a montré que deux allèles du minisatellite hypermutable CEB1 insérés dans la levure sont déstabilisés en mitose par une mutation de Rad27, de façon plus importante pour le grand allèle que pour le petit allèle. Cette déstabilisation, quoique plus marquée lorsque le minisatellite est inséré à proximité d'un point chaud de cassure double-brin, reste effective quand l'insertion a lieu en un autre locus. Les réarrangements complexes observés ont conduit au modèle présenté sur la Figure 13 (Lopes 2002).

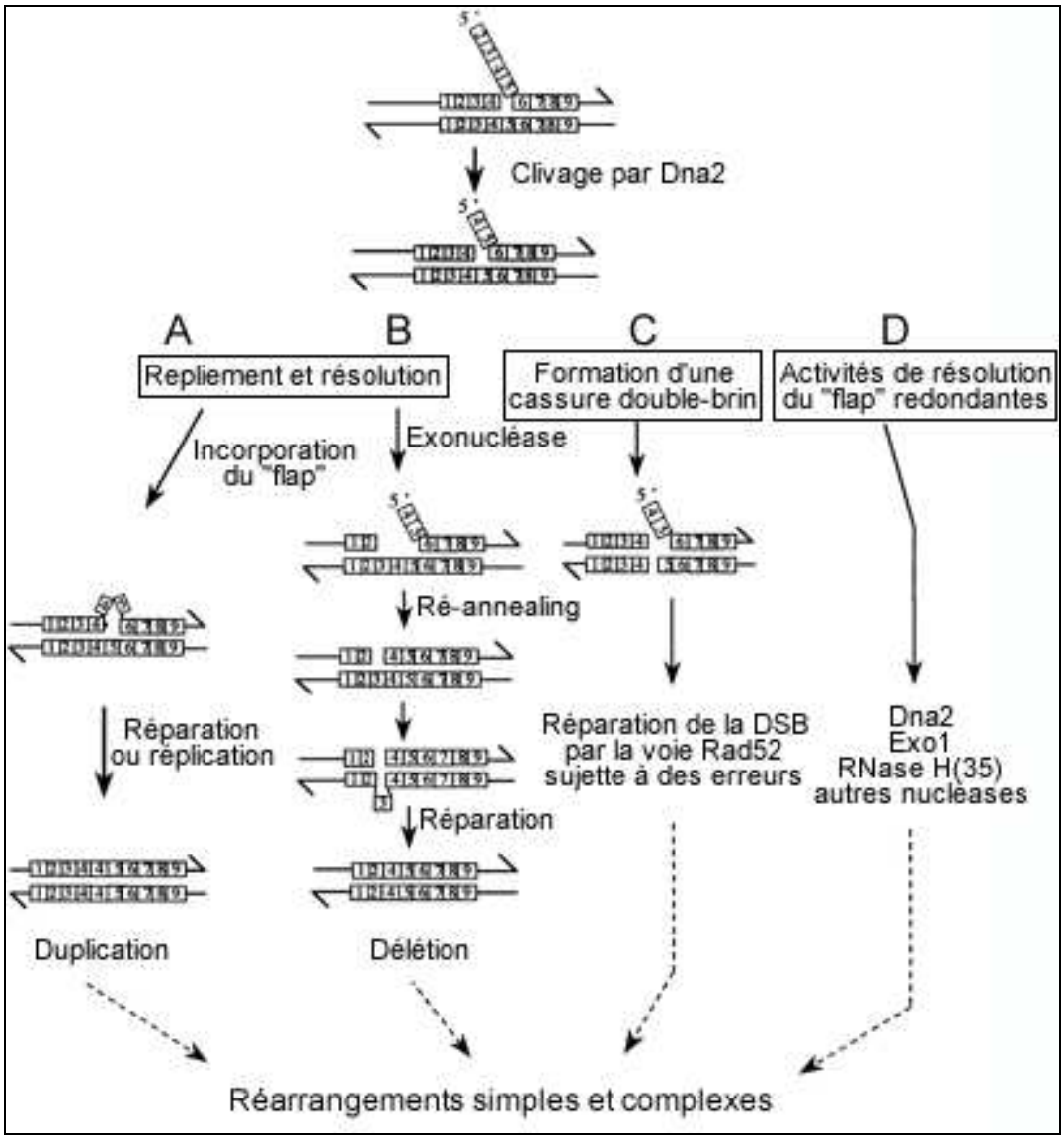


Figure 13 : Mécanisme de mutation des minisatellites hypermutables en mitose, dans un contexte mutant pour Rad27, proposé par (Lopes 2002).

Les voies A et B illustrent les mécanismes de ré-annealing du « flap » avec la matrice, ce qui peut conduire à des duplications ou des délétions de motifs (modèle adapté d'après celui

proposé par Kokoska *et al.* (Kokoska 1998) : voir Figure 9). La voie C montre une résolution des structures flap conduisant à la formation d'une cassure double-brin mal réparée par les mécanismes de réparation par recombinaison, comme proposé par Tishkoff *et al.* (Tishkoff 1997). La voie D correspond à une résolution de la structure « flap » par une des exonucléases Dna2, Exo1 ou RNase H(35).

Ce modèle de mutation somatique est proche de celui de 1994 proposant une cassure dans le minisatellite (Buard 1994). La séquence du minisatellite pourrait alors jouer un rôle direct sur ce mécanisme de mutation : par exemple, la tendance à former des structures secondaires pourrait influer sur la probabilité de suivre telle ou telle voie pour la réparation de la structure « flap ». Un rôle de la séquence des différents allèles pourrait alors expliquer les résultats disparates qui ont été observés aux locus de minisatellites humains, concernant la relation entre la taille des allèles et leur instabilité : pour CEB1 (Buard 1998) et B6-7 (Tamaki 1999), l'instabilité augmente avec la taille de la répétition en tandem, mais elle varie toutefois entre différents allèles de même taille, tandis que pour MS32 et MS205, le taux de mutation est indépendant de la taille des allèles (Jeffreys 1994) ; (May 1996). L'instabilité peut également être influencée par la divergence entre les deux allèles du minisatellite : le taux de mutations germinales du minisatellite MS205 intégré dans la levure est deux fois plus élevé chez un hétérozygote (31-51 répétitions) que chez un homozygote pour le plus grand allèle (51-51). Une hypothèse pour expliquer ce phénomène serait que le système de réparation des mésappariements, détectant l'ADN hétéroduplex, induirait des cassures double-brin secondaires, qui seraient alors réparées par une étape de recombinaison supplémentaire (He 2002).

Il apparaît donc que les minisatellites hypermutables insérés dans le génome de levure mutent par deux mécanismes, l'un en méiose, faisant intervenir l'initiation de la recombinaison homologue à proximité d'un minisatellite (Debrauwère 1999), l'autre en mitose, faisant intervenir une mauvaise maturation des fragments d'Okazaki au cours de la réplication (Lopes 2002) : étant donné que les mutations survenant en mitose chez la levure délétée pour Rad27 ont beaucoup de similitudes avec celles observées en méiose chez les minisatellites CEB1 et B6-7 chez l'homme (Buard 1998 ; Tamaki 1999), on peut imaginer qu'en plus des événements de recombinaison se produisant en prophase méiotique, des défauts spontanés de réplication pourraient aussi se produire lors de la gamétogène chez l'homme (Lopes 2002).

Les mécanismes de mutation des minisatellites chez les bactéries n'ont jusqu'à présent jamais été étudiés. En revanche, le mécanisme de mutation par glissement lors de la réplication s'applique aux répétitions en tandem bactériennes de la classe de taille des microsatellites : voir chapitre 1.2.4.1.

1.2.4.2.2 Minisatellites non hypermutables

Les minisatellites non-hypermutables, parfois polymorphes mais ayant un degré d'instabilité faible (inférieur au seuil de détection imposé par les études de familles, de 0.5% de mutation),

ont été beaucoup moins étudiés du point de vue de leurs mécanismes de mutation. Pour la plupart des minisatellites de ce type, on observe une distribution des longueurs totales bimodale (minisatellites de l'insuline (Bell 1984) et D2S44 (Holmlund 1998)) ou tri-modale (minisatellites D19S20 (Nakamura 1988) et MS51(D11S97) (Armour 1989)). Ce type de distributions résulte probablement d'une fixation de certains allèles dans la population par la dérive génétique, les mutations survenant dans ces allèles ne faisant ensuite varier leur taille que de quelques unités. On peut également imaginer des processus de mutation biaisés vers certaines tailles d'allèles ou même des phénomènes de sélection (Stead 2000).

Une étude menée par Stead et Jeffreys (Stead 2000) sur le minisatellite de l'insuline a permis de distinguer deux modes de mutation dans les lignées germinales : le premier, se produisant à un taux de 10^{-3}, implique des insertions et délétions de 1-2 unités, survenant en majorité dans des régions homogènes, ce qui suggère un mécanisme de glissement de la polymérase, comme pour les microsatellites. Les auteurs émettent l'hypothèse que de tels événements surviennent en mitose, même si on ne peut pas distinguer ces mutations d'événements de recombinaison méiotique intra-alléliques. Le deuxième mode de mutation, qui survient à un taux de 2.10^{-5} environ, correspond à des réarrangements complexes, de type conversion génique, semblables à ceux observés –de façon beaucoup plus fréquente- chez les minisatellites hypermutables. La diversité des allèles du minisatellite de HRAS1 suggère que des événements de conversion surviennent à ce locus relativement stable, mais à une fréquence inconnue (Ding 1999). Dans ce minisatellite, les mutations d'allèles associés au cancer surviennent préférentiellement dans trois régions : événements complexes de recombinaison inter-allélique dans une région, simples événements de duplication/délétion dans une autre, et événements de différents types dans la troisième région.

Il semble donc que chez ces minisatellites, comme chez les minisatellites hypermutables, deux mécanismes de mutation soient impliqués. Le premier, survenant en mitose (dans la lignée germinale ou dans la lignée somatique) impliquerait des glissements lors de la réplication, et des phénomènes de maturation erronée des fragments d'Okazaki, à des taux semblables entre les deux types de locus (Jeffreys 1997 ; Stead 2000). Le second, survenant en méiose, impliquerait des réarrangements complexes de type conversion génique, initiés par des cassures double-brin . Ce deuxième type d'événement serait rare chez les minisatellites « ordinaires » mais très fréquent chez les minisatellites hypermutables. Dans la revue que nous présenterons au chapitre 2.3.1, nous émettons donc l'hypothèse que les minisatellites seraient hypermutables lorsqu'ils se trouvent à proximité d'un point chaud de cassures double-brin (Vergnaud & Denoeud 2000). Ensuite, si une mutation survient dans la flanquante et affecte les propriétés de recombinaison à ce locus (une transversion G->C en amont de MS32 a en effet été associée à une diminution de l'instabilité de ce minisatellite (Monckton 1994)), la dérive génétique (Jeffreys 2002) peut conduire à l'extinction du minisatellite hypermutable, qui redeviendra alors un minisatellite ordinaire. Le site de cassure double-brin pourrait également provenir de l'arrangement de la chromatine et non de la

séquence elle-même : l'extinction de l'hypermutabilité proviendrait alors de mécanismes de réarrangement du génome à plus grande échelle. Pour les minisatellites non-hypermutables, les rares mutations survenant dans la lignée germinale pourraient résulter de cassures double-brin aléatoires survenant en méiose ou d'événements de type somatique, survenant lors des divisions mitotiques préalables.

1.2.4.3 Distinction entre microsatellites et minisatellites

Les études menées sur les mécanismes de mutation des répétitions en tandem montrent que les microsatellites comme les minisatellites sont déstabilisés en mitose, et en méiose (à une fréquence très importante pour les minisatellites hypermutables). Les mécanismes en cause font intervenir le glissement de la polymérase et la résolution des structures « flaps » générées par les fragments d'Okazaki sur le brin retardé lors de la réplication et/ou la réparation de cassures double-brin survenues dans ou à proximité de la répétition en tandem. La distinction entre leurs modes de mutation est plus quantitative (fréquences relatives des différents événements) que qualitative. Elle sera rendue plus aisée lorsque certains de ces mécanismes, encore très énigmatiques, pourront être élucidés.

Des distinctions majeures peuvent toutefois être déduites de l'analyse des mutations survenant dans les microsatellites et les minisatellites (les répétitions en tandem présentant l'intérêt de conserver parfois la trace d'anomalies du métabolisme de l'ADN, et de reconstituer a posteriori la séquence des événements, dans des systèmes biologiques peu propices à des investigations directes) :

- L'homogénéité interne des microsatellites est supérieure à celle des minisatellites, ce qui est cohérent avec un mécanisme de mutation impliquant le glissement de la polymérase pour les premiers, et un mécanisme compatible avec l'hétérogénéité des motifs, comme la réparation de cassures double-brins, pour les deuxièmes.
- Des mutations dans le système de réparation des mésappariements déstabilisent les microsatellites mais pas les minisatellites (sauf MS1, minisatellite d'unité répétée de « type » microsatellite : 9 pb, décrit plus bas (Berg 2003)). Cette observation peut s'expliquer par le fait que les boucles générées lors de glissements ne sont corrigées par ce système que lorsque leur taille est inférieure à 16 paires de bases (Sia 1997).
- Des mutations dans les enzymes permettant la résolution des structures « flap » générées par les fragments d'Okazaki déstabilisent les microsatellites et les minisatellites, mais à des degrés très différents (Kokoska 1998) : les minisatellites y sont beaucoup moins sensibles, ce qui pourrait être lié à la taille des fragments d'Okazaki (100 à 150 paires de bases (MacNeill 2001)). Le fait que plusieurs unités répétées soient contenues sur un fragment d'Okazaki pourrait en effet augmenter la probabilité d'événements de mutation (Monckton 1995).

Les caractéristiques structurales des répétitions en tandem semblent donc moduler l'importance relative des différents mécanismes de mutation à l'œuvre, et distinguent donc la classe des microsatellites de celle des minisatellites. Cependant, certaines répétitions en tandem se situent à la « frontière » entre ces deux classes, ce qui rend difficile l'établissement d'une définition stricte. Deux exemples en sont présentés ci-après :

- Le minisatellite hypermutable EPM1, situé en 5' du gène de la cystatine B, et dont l'amplification est responsable de l'épilepsie myoclonique progressive de type 1 (voir Tableau 8), a un motif répété de 12 paires de bases. Ce locus cause la maladie par un phénomène d'amplification semblable à ceux qui surviennent pour les maladies neurodégénératives causées par des expansions de triplets (voir Tableau 7). Les phénomènes de mutations correspondent en majorité à des expansions/contractions d'une seule unité (Lalioti 1997). Larson *et al.* ont émis l'hypothèse que l'instabilité, apparemment germinale, extrême des allèles amplifiés (taux de mutation de 0.47) pourrait être causée par des erreurs de réplication dans une région de la répétition contenant 100% de GC sur une longueur considérable (600-800 pb). Le taux de mutation légèrement plus faible des allèles pré-mutateurs résulterait alors de leur longueur plus faible (Larson 1999).
- Le minisatellite hypermutable MS1 (D1S7) a une unité répétée de 9 paires de bases seulement : son nombre de copies varie de 60 à plus de 1000, avec un taux d'hétérozygotie (relatif au polymorphisme de taille des allèles) supérieur à 99% (Wong 1987 ; Royle 1988). Son taux de mutation en lignée germinale a été estimé à 5.2% (Jeffreys 1988). Le grand nombre d'unités répétées et l'instabilité extrême de cette répétition en tandem la classent dans la catégorie des minisatellites plutôt que des microsatellites, même si son motif répété est court. Cependant, il a certaines particularités qui le rapprochent plus des microsatellites que des minisatellites. En particulier, c'est le seul minisatellite connu qui soit déstabilisé dans des cellules de cancer du colon montrant une instabilité des microsatellites (Berg 2003), ce qui suggère que la réplication ou les erreurs de réparation peuvent contribuer à son instabilité. De plus, inséré dans la levure, ce minisatellite mute à haute fréquence en méiose comme en mitose, ce qui n'est pas le cas d'autres minisatellites étudiés dans le même contexte, pour lesquels un fort taux de mutation n'est observé qu'en méiose (Appelgren 1997 ; He 1999). En mitose, l'instabilité de MS1 est dépendante de la taille des allèles (seuil d'instabilité à 750 pb) (Maleki 1997), et de la structure interne des allèles. En méiose, l'instabilité correspond principalement à des événements de type conversion génique, intra- ou inter-allélique, comme pour les autres minisatellites hypermutables (Berg 2000). Récemment, des études ont été menées non plus dans le modèle levure mais chez l'homme (Berg 2003). Les événements méiotiques intra-alléliques y sont plus fréquents que les événements inter-alléliques, et un autre phénomène de mutation a lieu : il s'agit de grandes délétions qui surviennent dans de longues régions homogènes, constituées d'au moins 12 unités du type « C » (108 pb), ce qui est similaire au seuil d'instabilité des triplets CGG du microsatellite associé au

syndrome de l'X fragile (Eichler 1994) (34-38 répétitions soit 102-114 pb). De plus, on observe une stabilisation de la répétition lorsque l'homogénéité est interrompue, ce qui rappelle également les mécanismes de mutation de microsatellites. Ce phénomène n'ayant pas été mis en évidence dans les lignées somatiques, il semble peu probable qu'il soit dû à un glissement survenant lors de la réplication mais plutôt lors de la recombinaison méiotique (Berg 2003). Certains mécanismes interprétés comme des glissements réplicatifs pourraient en fait correspondre à des recombinaisons entre chromatides sœurs : il est possible que jusqu'à présent, le phénomène de glissement lors de la réplication ait été surestimé au détriment de mécanismes plus complexes. Cependant, ces événements complexes se produisent bel et bien à certains locus, et ils mériteraient donc d'être invoqués également dans les cas « simples ».

Enfin, en plus des différences entre microsatellites et minisatellites inhérentes à leurs mécanismes de mutation, ces deux types de répétitions en tandem peuvent être distingués par leur distribution chromosomique, du moins dans le génome humain, comme nous le montrons dans la revue présentée au chapitre 2.3.1 (Vergnaud & Denoeud 2000). En effet, parmi l'ensemble des répétitions en tandem du chromosome 22 humain, il apparaît que la transition entre une répartition homogène et une répartition biaisée vers les télomères correspond à une unité répétée de 17 pb, taille des boucles non réparables par le système de réparation des mésappariements (Sia 1997), ou à une longueur totale de 120-140 pb, seuil similaire à celui décrit plus haut pour l'instabilité des triplets. Ainsi, il semble que les mécanismes de création et/ou de maintien (pression de sélection…) des microsatellites et des minisatellites dans le génome humain soient distincts, comme le reflète leur différence de distribution chromosomique, mais ces phénomènes restent pour l'instant énigmatiques.

Les minisatellites étaient définis usuellement comme des structures de longueur totale de l'ordre du kilobase, ce seuil étant imposé par les contraintes expérimentales qui ont accompagné leur première caractérisation (Southern Blot). Nous proposons d'élargir cette définition (nous le verrons dans la revue présentée au paragraphe 2.3.1 (Vergnaud & Denoeud 2000)) à des répétitions en tandem de longueur totale supérieure à 140 pb environ et/ou d'unité répétée supérieure à 15 pb environ. Des structures d'unité répété plus petite (entre 6 et 15 pb) pourront être considérées comme des minisatellites si elles sont très étendues : c'est le cas des minisatellites MS1 et EPM1, mais nous avons vu que certaines de leurs caractéristiques restent similaires à celles de microsatellites.

1.2.5 Origine des répétitions en tandem

Etant donné que les répétitions en tandem sont retrouvées chez toutes les espèces, eucaryotes comme procaryotes, analysées jusqu'à présent, et qu'il n'existe pas d'homologie entre les répétitions en tandem de différentes espèces, sauf très proches (Taylor 1999), la formation de

nouvelles répétitions en tandem doit être un événement survenant fréquemment dans les génomes.

Le premier modèle de génération des répétitions en tandem, proposé par Levinson et Gutman (Levinson 1987), implique la formation de petites répétitions en tandem survenant par hasard suite à des mutations ponctuelles, qui seraient ensuite amplifiées par glissement lors de la réplication. Cependant, même si ce modèle est plausible pour de petites répétitions comme les microsatellites (par exemple une étude chez les primates a montré que deux répétitions en tandem de motif de 2pb et 4pb avaient pu être générées par des mutations aléatoires (Messier 1996)), il n'est pas envisageable pour des répétitions en tandem de motifs plus grands (minisatellites). De plus, une distinction entre les mécanismes générant les microsatellites et les minisatellites pourrait expliquer les différentes distributions chromosomiques observées pour ces deux types de répétitions en tandem.

Pour certains minisatellites, on observe que la région répétée est flanquée de part et d'autre par quelques (5 à 10) nucléotides identiques, c'est-à-dire que la répétition se termine par un motif incomplet de quelques paires de bases : un exemple en est présenté sur la Figure 14A. Ce type de structure a été décrit chez la levure *S. cerevisiae* (Haber 1998) ainsi que d'autres champignons (Giraud 1998), l'homme (Haber 1998 ; Murray 1999), les oiseaux (Gyllensten 1989), le saumon (Goodier 1998) et les mitochondries de divers organismes (Taylor 2000).

Ces observations ont conduit à l'hypothèse selon laquelle des répétitions non-contiguës seraient à l'origine d'une duplication de la région comprise entre ces deux répétitions, par glissement lors de la réplication (un modèle est proposé par Taylor *et al.* : Taylor 2000), ou par crossing-over inégal (Haber 1998). Cette répétition pourrait ensuite être amplifiée par les mécanismes susceptibles de faire évoluer les répétitions en tandem que nous avons vus précédemment, tels que le glissement réplicatif ou la conversion génique (Haber 1998 ; Taylor 2000) : voir Figure 14B. Selon le temps écoulé depuis la création du minisatellite, l'homologie entre les petites séquences flanquant la répétition en tandem pourra avoir été perdue ou conservée (Haber 1998). Ce modèle nécessite la présence fortuite de répétitions non adjacentes, qui ont plus de chances de survenir dans des génomes de faible complexité, à fort biais de composition (Achaz 2002).

Cependant, de manière plus remarquable, de nombreux minisatellites comptent un nombre entier de motifs. Le modèle proposé dans la Figure 14 n'explique pas la formation de telles répétitions. L'hypothèse proposant que de courtes répétitions non adjacentes, survenant par hasard, seraient à l'origine de la formation des minisatellites reste plausible. Mais pour tenir compte des minisatellites « parfaits » il faudrait un modèle selon lequel ces petites répétitions seraient ensuite « avalées » dans la répétition en tandem. Jusqu'à présent, peu d'attention a été portée au problème de l'origine des minisatellites, et aucun modèle satisfaisant n'a pu être proposé : les mécanismes en cause restent encore très énigmatiques.

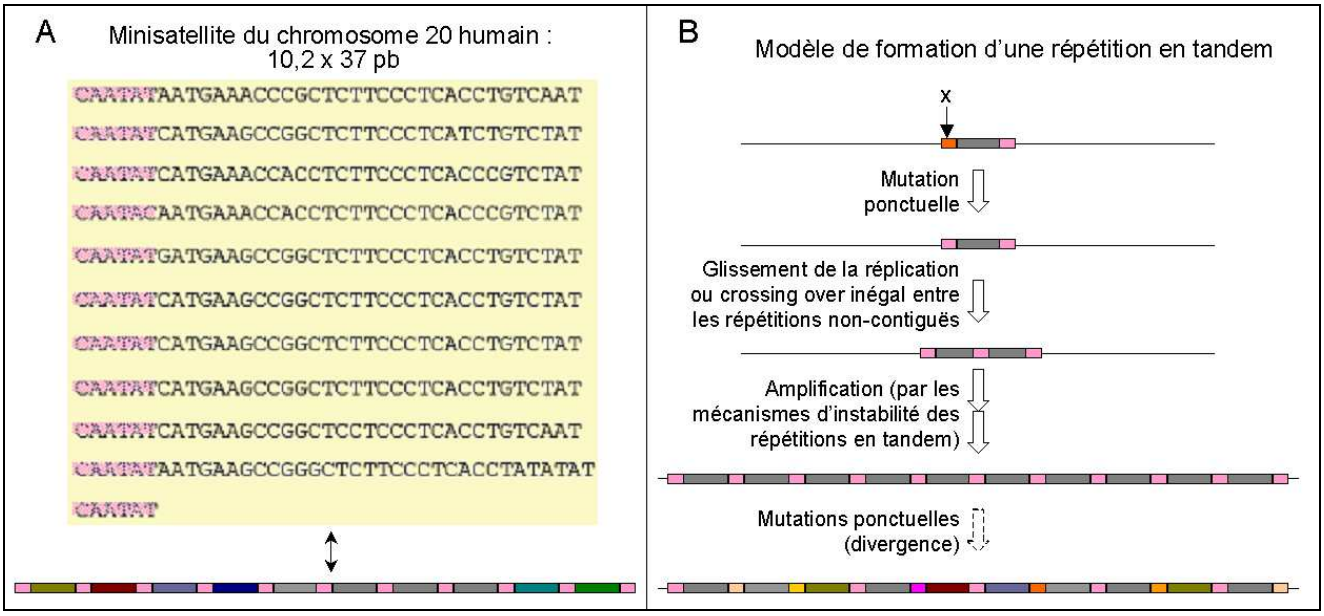


Figure 14 : Une hypothèse sur l'origine des minisatellites.

1.3 Identification des répétitions en tandem dans les génomes

1.3.1 La bioinformatique

La bioinformatique est née lorsque, suite à l'essor des techniques de séquençage, deux nécessités sont apparues pour la recherche en biologie : le stockage de séquences dans des bases de données informatiques, et l'analyse de ces séquences (nucléiques ou protéiques). La bioinformatique correspond à un domaine interdisciplinaire et peut être définie comme « l'application de l'informatique pour l'acquisition, la gestion, et l'analyse de données biologiques » (Brown 2003). La plupart des algorithmes qui sont à la base des outils bioinformatiques actuels sont nés dans les années 70-80, et proviennent en fait de recherches théoriques :

- 1970 : algorithme de Needleman et Wunsch pour la comparaison de séquences (alignement global) (Needleman 1970).
- 1978 : algorithme de Chou et Fasman pour la prédiction de structures secondaires de protéines à partir de leur séquence d'acides aminés (Chou 1978).
- 1979 : package Staden pour l'assemblage de fragments d'ADN chevauchants provenant de projets de séquençage (Staden 1979 ; Staden 1980).
- 1981 : algorithme de Smith et Waterman pour l'alignement de séquences (alignement local) (Smith 1981).

- 1982 : profils d'hydrophobicité des protéines par Kyte et Doolittle (Kyte 1982).
- 1987 : alignements multiples de séquences par Feng et Doolittle (Feng 1987).
- 1987 : méthode du « neighbor joining » pour la reconstruction phylogénétique par Saitou et Nei (Saitou 1987).
- 1988 : Pearson et Lipman créent FASTA, logiciel de recherche de similitudes entre une séquence et une base de séquences (Pearson 1988).
- 1990 : Altschul et collègues implémentent BLAST au NCBI, autre méthode de comparaison entre une séquence et une base de séquences (Altschul 1990 ; Altschul 1994). Cette méthode sera détaillée au paragraphe 2.1.1.3.

Le terme « bioinformatics » a été indexé dans les mots-clés ou titres de PubMed/Medline [http://www.ncbi.nlm.nih.gov/PubMed/] à partir de 1993. Depuis, le nombre d'occurrences de ce terme comme du terme « computational biology » n'a cessé de croître (Brown 2003).

Ces dernières années, le nombre de génomes séquencés a augmenté de façon spectaculaire, en particulier chez les bactéries (voir chapitre 1.1.1.1), ce qui a été accompagné de progrès dans les domaines de la génomique, de la protéomique mais aussi de la bioinformatique : raffinements aux algorithmes de recherche d'homologie dans les bases de données, prédiction de structures protéiques secondaires et tertiaires, logiciels de recherche de motifs protéiques ou nucléiques, analyse des séquences répétées (Heringa 1998). En particulier, comme nous le détaillons au paragraphe suivant, différents outils permettant l'identification de répétitions en tandem dans les séquences ont été élaborés.

1.3.2 Logiciels d'identification de répétitions en tandem

Jusqu'à ces dernières années, les répétitions en tandem étaient identifiées de façon expérimentale (hybridation de sondes sur des Southern Blots…) mais dorénavant la disponibilité de données de séquence permet leur identification systématique, *in silico*, dans les génomes. Lorsque mon DEA a commencé, en 1999, les outils disponibles étaient peu nombreux : la plupart permettaient la recherche de répétitions dans les séquence protéiques, de façon plus ou moins efficace (Heringa 1998 ; Pellegrini 1999). Les programmes détectant les répétitions en tandem nucléiques avaient certaines lacunes : certains, comme le programme « REPuter » (Kurtz 1999), ne détectaient que les répétitions en tandem strictement conservées. Ce type d'outils a une utilité pour identifier les microsatellites, mais devient inapproprié pour les minisatellites, très rarement strictement conservés. D'autres permettaient la détection de répétitions en tandem dégénérées mais seulement dans la plage de taille des microsatellites (par exemple sputnik : http://espressosoftware.com:8080/esd/pages/sputnik.jsp, créé par Abajian en 1994), ou nécessitaient de préciser certains paramètres tels que la longueur du motif répété recherché et son degré de conservation (Korotkov 1997 ; Sagot 1998), et/ou requéraient des temps de calcul très longs, ce qui rendait

la tâche consistant à analyser des chromosomes entiers pour toutes leurs répétitions en tandem beaucoup trop fastidieuse. Une telle tâche est devenue envisageable avec la création du logiciel « Tandem Repeats Finder » (TRF) par Gary Benson en 1999 (Benson 1999). Cet algorithme est efficace, il peut donc traiter de longues séquences nucléiques, et ne nécessite pas d'entrer des paramètres précis sur les répétitions en tandem recherchées, ce qui permet leur identification « à l'aveugle ». Nous le décrivons plus en détail au paragraphe suivant.

1.3.2.1 Le « Tandem Repeats Finder » (TRF)

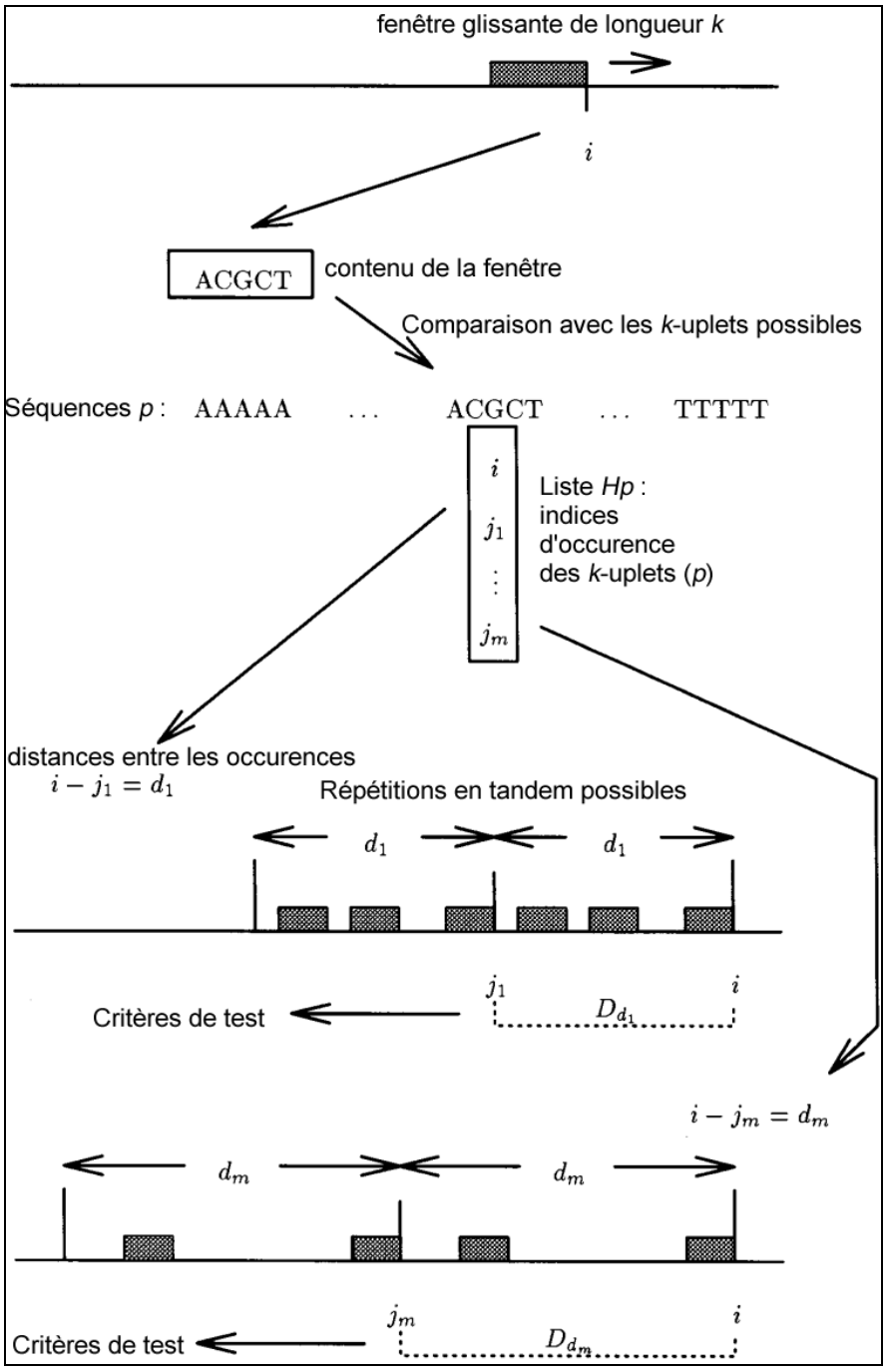


Figure 15 : Algorithme de détection des répétitions en tandem implémenté par le « Tandem Repeats Finder » (Benson 1999).

L'algorithme utilisé est de type probabiliste, heuristique, c'est-à-dire qu'il n'est pas garanti de trouver toutes les répétitions en tandem existantes, mais il a l'avantage, par rapport aux algorithmes exacts, de ne pas nécessiter un temps de calcul trop long ou de rentrer des paramètres trop précis sur les répétitions en tandem recherchées. La recherche repose sur le fait que dans une répétition en tandem, une même base, ou un groupe de *k* bases successives (*k*-uplet), sera retrouvé à intervalles réguliers (distance *d*), sauf mésappariements. La Figure 15 présente les différentes étapes pour la détection des répétitions en tandem. Tout d'abord, la séquence est parcourue pour détecter les positions *i* de tous les *k*-uplets (où k =5 par exemple) possibles (4^k) : cette liste sera notée *Hp* (ensemble des positions où se trouve le k-uplet *p*). Lorsqu'une position *i* est ajoutée à la liste *Hp* des occurrences d'un *k*-uplet donné (*p*), toutes les occurrences de *p* déjà identifiées sont recherchées : si une occurrence a eu lieu à la position *j*, la distance $d = i-j$ est la taille possible d'un motif répété en tandem. Dans ce cas, d'autres *k*-uplets seront trouvés au voisinage, éloignés de la même distance *d*. La liste *Dd* contient l'ensemble des distances *d* survenant entre des *k*-uplets dont l'une des occurrences se situe entre les positions *j* et *i*. Ces listes de distances sont alors testées selon un critère statistique, basé sur la loi de Bernoulli, qui dépend de la taille du motif, de la probabilité de match, de la probabilité d'insertion/délétion, et de la taille du n-uplet. Si le test est favorable, un motif candidat allant des positions j+1 à *i* est alors aligné avec la séquence environnante : si au moins deux copies du motif s'alignent avec la séquence, la répétition en tandem est mentionnée. Un motif consensus est ensuite généré, correspondant au motif s'alignant le mieux avec toutes les unités répétées, et l'alignement des unités avec ce motif consensus est proposé en sortie. La taille de la période (« period size ») correspond à la distance entre les caractères matchant et peut être différente de la taille du consensus.

Ce logiciel présente l'inconvénient de générer des sorties redondantes (pourtant limitées à trois par répétition en tandem détectée) : par exemple, une même plage peut être proposée avec un motif répété de 24 et un motif de 48 pb. D'autre part, certaines plages de répétitions en tandem se chevauchent, ou certaines répétitions en tandem peuvent être segmentées si elles contiennent une portion mal conservée. Cet inconvénient ne constitue cependant pas une vraie limite, étant donné que le choix de la « meilleure » répétition en tandem n'est pas évident, et nécessite donc dans la plupart des cas l'intervention de l'utilisateur (en particulier en ce qui concerne la détermination exacte des bornes des répétitions en tandem). De plus, comme nous le verrons dans le paragraphe suivant, les différents autres logiciels proposés depuis le TRF ne s'avèrent pas meilleurs. Le TRF, logiciel d'utilisation conviviale, accessible à l'adresse [http://tandem.biomath.mssm.edu/trf/trf.html] reste donc un outil de référence pour la détection de répétitions en tandem dans les génomes (Yeramian 1999) : par exemple, il a été utilisé lors de l'analyse de la séquence brouillon du génome humain par le consortium public (Lander 2001).

1.3.2.2 Autres outils de détection des répétitions en tandem

Depuis 1999, d'autres outils de recherche et de manipulation des répétitions en tandem ont été élaborés et des outils existants ont été améliorés:

- Le programme REPuter, basé sur des arbres de suffixes, a été amélioré pour permettre la détection de répétitions approximatives. Cependant, il ne fournit qu'une vue globale des répétitions mais pas leurs séquences détaillées (Kurtz 2000) [http://www.genomes.de].

- Le package EMBOSS (Rice 2000) propose des logiciels de détection des répétitions en tandem : etandem et equicktandem, qui ont l'intérêt de pouvoir traiter des séquences nucléiques comme protéiques. L'interface est toutefois peu conviviale, et nécessite d'entrer un seuil de score peu intuitif. De plus, la sortie par défaut ne propose pas la séquence des répétitions [http://www.hgmp.mrc.ac.uk/Software/EMBOSS/Apps/equicktandem.html].

- Une méthode basée sur des codes couleurs attribués aux différentes bases permet une identification visuelle des répétitions en tandem (Yoshida 2000).

- Un algorithme, accessible via une interface web [http://csweb.cs.haifa.ac.il/library/appro_try1.html], permet de rechercher des répétitions en tandem approximatives (Landau 2001). Ce programme présente l'inconvénient de nécessiter que l'utilisateur entre le nombre de mésappariements maximal dans les répétitions : pour des taux de mésappariement égaux, ce nombre variera en même temps que la taille du motif répété. Pour faire des recherches exhaustives, il sera donc nécessaire de lancer plusieurs exécutions avec différentes valeurs du paramètre (l'augmentation de cette valeur ralentissant notablement l'exécution).

- Un programme de détection et de classification des répétitions (dont les répétitions en tandem) dans les génomes, basé sur les arbres de suffixes a été mis au point par Volfovsky *et al.* (Volfovsky 2001), mais il ne dispose pas d'une interface web.

- Le logiciel TROLL identifie de façon très efficace les petites répétitions en tandem correspondant à un « dictionnaire » de répétitions connues (Castelo 2002).

- En 2002, Hauth *et al.* ont proposé un algorithme basé sur l'« extension de graines » pour détecter les répétitions en tandem « classiques » mais aussi de deux autres types, les « VLTRs » (variable length tandem repeats) et les « MPTRs » (multi-period tandem repeats) : voir Figure 16. Ces répétitions en tandem étaient également identifiées par le TRF par exemple, lors de sorties redondantes, mais elles n'étaient pas spécifiées comme telles (Hauth 2002).

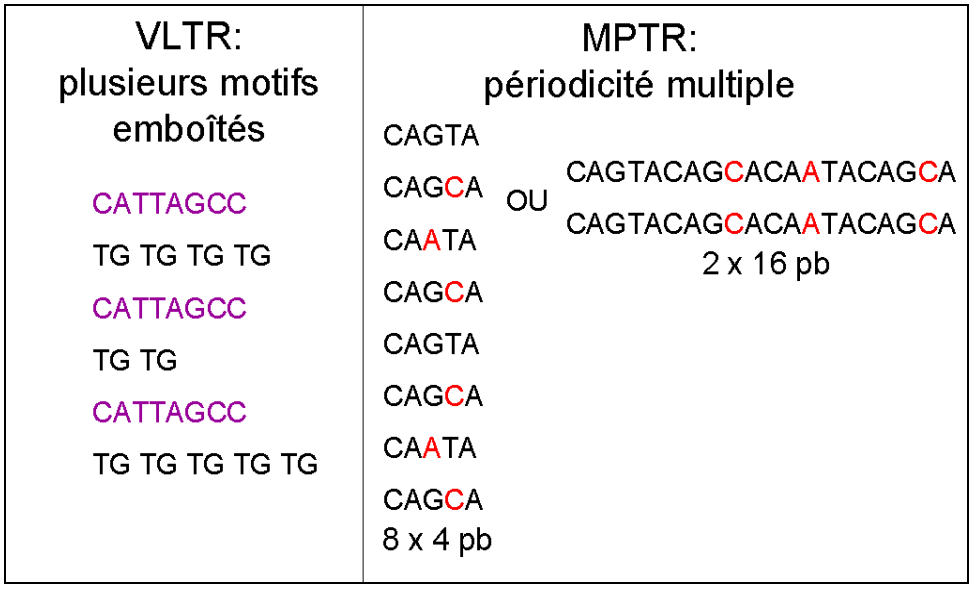


Figure 16 : Description de deux « nouveaux » types de répétitions en tandem, les VLTR et les MPTR (Hauth 2002).

- Le programme « MREPS », accessible à l'adresse [http://bioweb.pasteur.fr/seqanal/interfaces/mreps.html], possède une interface conviviale et permet une identification efficace des répétitions en tandem (Kolpakov 2003). Les résultats obtenus ne sont toutefois pas meilleurs que ceux provenant du TRF, comme nous le verrons dans le paragraphe suivant. De plus, ce logiciel requiert que l'utilisateur saisisse un paramètre de « résolution » peu intuitif.

- Le logiciel TRAP améliore l'assemblage des répétitions en tandem (Tammi 2003), une des principales sources d'erreurs lors du séquençage des génomes.

1.3.2.3 Comparaison entre les logiciels de détection des répétitions en tandem disponibles sur le web

Pour différentes répétitions en tandem, les sorties des logiciels accessibles sur Internet : TRF (Benson 1999) [http://tandem.biomath.mssm.edu/trf/trf.html], MREPS (Kolpakov 2003) [http://bioweb.pasteur.fr/seqanal/interfaces/mreps.html], et « A Program for Approximate Tandem Repeat » (Landau 2001) [http://csweb.cs.haifa.ac.il/library/appro_try1.html], ont été comparées. Les résultats sont présentés sur la Figure 17.

Le microsatellite, D21S11 a une structure répétée irrégulière : les logiciels TRF et MREPS ont pu la mettre en évidence (31 x 4 pb dégénérées). Toutefois, le logiciel MREPS propose de nombreuses autres répétitions en tandem dans cette plage, alors que seul ce microsatellite est proposé par le TRF. On note en outre que pour le logiciel TRF, cette répétition en tandem a été détectée avec tous les paramètres proposés, tandis que pour MREPS, le paramètre par défaut, ne permettant pas de détecter des répétitions en tandem dégénérées, ne l'a évidemment pas détectée : il faut utiliser un paramètre de « résolution » de 5 ou plus. Le logiciel « A

Program for Approximate Tandem Repeats » ne détecte pas la séquence entière du microsatellite : elle est fragmentée en zones bien conservées. Même une augmentation du nombre de mismatchs autorisés n'améliore pas sa détection. En outre, ce microsatellite n'est même pas partiellement détecté par les paramètres par défaut, qui n'autorisent la détection que d'unités répétées de 6 paires de bases ou plus.

Ces différents logiciels ont également été testés sur le minisatellite CEB268 du chromosome 21 humain, qui est relativement mal conservé (74% de conservation moyenne entre les unités répétées et le motif consensus) (Denoeud 2003). Le TRF détecte des motifs redondants, différents selon les paramètres choisis : la répétition en tandem considérée comme la plus représentative de la région (10,2 x 54 pb) n'est détectée dans sa totalité qu'avec les paramètres les moins restrictifs (2,3,5 : match,mismatch,indel). Le logiciel MREPS ne détecte cette répétition, de façon un peu décalée par rapport au TRF (les bornes des répétitions en tandem sont en effet souvent difficiles à définir), qu'avec un paramètre de résolution maximal : 140 (la valeur de 145 est qualifiée de trop élevée pour la séquence à traiter). Le logiciel « A Program for Approximate Tandem Repeats » produit une erreur : le site mentionne qu'il accepte les séquences jusqu'à 1024 pb, mais il semble que ce minisatellite soit trop complexe, ce qui génère un temps de calcul trop long. Le logiciel « A Program for Approximate Tandem Repeats » n'est clairement pas adapté pour traiter des longues séquences : les seuls outils vraiment efficaces sont TRF et MREPS.

Un avantage du TRF est que les paramètres sont pré-définis, et seuls 4 choix sont possibles. En revanche, pour MREPS (comme pour le troisième logiciel), le paramètre permettant de jouer sur l'homogénéité des répétitions, ou paramètre de « résolution », est libre, et l'utilisateur ne peut pas savoir quelles valeurs pertinentes lui donner (en particulier en ce qui concerne le seuil maximal). Pour le TRF, la restriction du nombre de paramètres accessibles pourrait être déplorée car elle reflète des choix faits par l'algorithme lui-même, que l'utilisateur ne peut pas contrôler et difficilement comprendre. Cependant, pour l'utilisation que nous aurons du TRF (identification rapide de répétitions en tandem dans des génomes complets), il est essentiel de pouvoir effectuer une seule exécution du programme par séquence à traiter, même si l'identification systématique des répétitions en tandem n'est pas garantie. Du point de vue de la facilité d'utilisation, qui ne requiert pas trop de mises au point pour définir les meilleurs paramètres, comme de la présentation des résultats, qui propose l'alignement des répétitions avec le consensus, le TRF reste le meilleur outil disponible à l'heure actuelle.

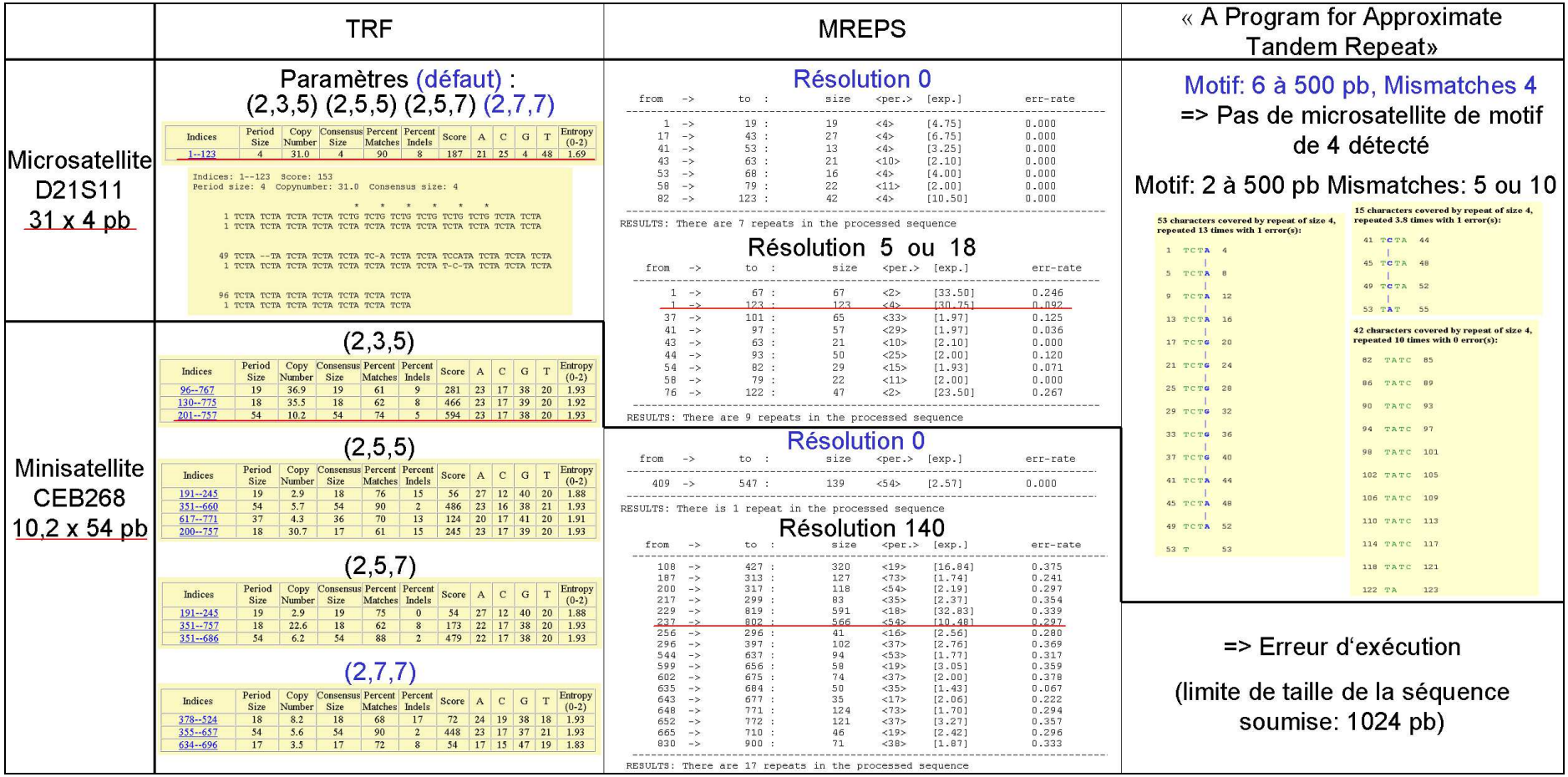


Figure 17 : Comparaison de programmes de détection de répétitions en tandem accessibles sur Internet.

1.3.3 Identification des répétitions en tandem polymorphes

Nous avons vu que des logiciels permettaient la détection de répétitions en tandem dans les séquences. Cependant, les répétitions en tandem d'intérêt sont les répétitions polymorphes, soit pour l'épidémiologie bactérienne, soit pour l'étude de pathologies humaines associées à ces polymorphismes, ou encore si on recherche des minisatellites hypermutables, en premier lieu extrêmement polymorphes. Pour l'instant, il n'existe pas d'outil « miracle » permettant de détecter les répétitions en tandem polymorphes. Cependant, quelques pistes peuvent être suivies, en particulier pour les microsatellites.

1.3.3.1 Microsatellites

Comme nous l'avons vu au paragraphe 1.2.4.1, le mode de mutation des microsatellites par glissement de la réplication dépend de la longueur et de la conservation de ces répétitions en tandem (Weber 1990). Il est donc probable que les microsatellites les plus polymorphes seront les mieux conservés, et les plus longs. Une étude menée par Wren *et al.* a en effet montré qu'une requête sélectionnant des microsatellites très conservés fournissait une grande proportion de locus polymorphes (Wren 2000).

Cependant, d'une part, le glissement de la réplication n'est pas le seul mécanisme en cause dans la mutation des microsatellites, et d'autre part, si des microsatellites autrefois instables car bien conservés sont devenus plus stables, par exemple suite à des mutations ponctuelles interrompant leur homogénéité, plusieurs allèles pourront subsister dans la population : ces locus « inactifs » mais pourtant polymorphes, risquent de ne pas être détectés par ces critères de longueur et de conservation, ce qui, selon le but recherché, peut être problématique. En effet, si on cherche des locus instables pour étudier les modes de mutation, ce type de locus ne sera pas intéressant. En revanche, si on recherche des locus polymorphes pour l'identification de souches bactériennes et la phylogénie, les bons marqueurs épidémiologiques devront au contraire être stables de génération en génération, et refléter le plus possible de « vrais événements » évolutifs, auquel cas les locus inactifs seront privilégiés.

1.3.3.2 Minisatellites

Pour les minisatellites, la tâche visant à déterminer le polymorphisme à partir de la séquence est encore plus ardue. En effet, les mécanismes de mutation semblent peu influencés par les caractéristiques de séquence (Buard 1998), et aucune caractéristique commune aux séquences de minisatellites hypermutables n'a pu être mise en évidence (Murray 1999). Cependant, jusqu'à présent, l'analyse des séquences de répétitions en tandem ne s'est intéressée qu'à des critères simples (longueur, composition…) et pas à l'organisation des mutations dans les motifs répétés. Ce type d'information mériterait toutefois d'être utilisé : par exemple, une répétition « active » aura probablement une plus grande proportion de mutations communes à

plusieurs motifs qu'une répétition inactive, dans laquelle des mutations ponctuelles auront pu survenir sans être distribuées dans l'ensemble de la répétition (mutations uniques, spécifiques d'une unité particulière). Par ailleurs, la séquence des allèles de minisatellites reflète les mécanismes de mutation ayant encore cours, mais également ceux qui sont révolus : pour les minisatellites anciennement hypermutables qui sont devenus stables, des allèles extrêmement variables, en ce qui concerne la taille comme la complexité de l'arrangement des motifs, seront observés dans la population. Si l'extinction du point chaud de cassures double-brin à l'origine de leur instabilité a eu lieu depuis longtemps, l'analyse des séquences alléliques sera encore compliquée par des mutations ponctuelles. Dans cette thèse, nous présenterons un article traitant de la prédiction du polymorphisme des minisatellites humains, à partir de caractéristiques de séquence simples et complexes (arrangement des motifs) (Denoeud 2003) : voir chapitre 2.3.2.

1.3.3.3 But de la thèse

Le but de cette thèse est de mettre à profit les données de séquençage des génomes afin d'identifier des répétitions en tandem polymorphes. L'étape visant à identifier les répétitions en tandem dans les génomes a été menée à bien par l'élaboration d'une base de données des répétitions en tandem [http://minisatellites.u-psud.fr], grâce au logiciel TRF : cette base de données sera présentée dans le chapitre 2.1. La seconde étape, visant à identifier parmi le grand nombre de répétitions en tandem, celles qui sont polymorphes, a été traitée chez les bactéries pour lesquelles les séquences de plusieurs souches étaient disponibles, par la comparaison entre ces séquences. Nous présenterons dans le chapitre 2.2 l'utilisation de la base de données pour l'identification de marqueurs polymorphes utilisables en épidémiologie bactérienne. Enfin, une étude menée sur les minisatellites humains a permis de définir des critères prédictifs de la séquence d'un seul allèle de minisatellite sur le polymorphisme (les séquences CELERA et HGP n'étant pas indépendantes, leur comparaison s'est avérée décevante), et a en outre identifié un nouveau minisatellite hypermutable situé dans une séquence codante putative. Ces résultats seront exposés dans le chapitre 2.3, ainsi qu'une application à la détection de minisatellites associés à des séquences codantes, et donc potentiellement impliqués dans des processus biologiques d'intérêt.

2 Résultats

2.1 Présentation de la base de données des répétitions en tandem

Cette base de donnée peut être définie comme un outil pour l'identification, la caractérisation, et la capitalisation des connaissances concernant les répétitions en tandem.

De multiples ressources « bioinformatiques » accessibles via Internet fleurissent presque quotidiennement. Certaines sont des logiciels d'analyse couplés à une interface permettant à un utilisateur une analyse commode de ses données. En France, Infobiogen [http://www.infobiogen.fr/] est probablement le fournisseur le plus connu de ce type de service. D'autres rendent accessibles des données qu'elles hébergent. Le site le plus consulté est le site NCBI aux Etats-Unis [http://www.ncbi.nlm.nih.gov/]. En plus de ces sites majeurs, il existe une multitude d'autres sites, souvent d'utilité restreinte et de ce fait éphémères. Il est donc raisonnable, avant de développer un nouvel outil de ce type, de se poser la question du bien-fondé de la démarche. Le site que j'ai développé a avant tout cherché à répondre à des besoins immédiats du laboratoire, et son premier utilisateur et évaluateur en est donc le laboratoire. Le site a été rendu accessible sur Internet, dans l'espoir que cette ressource sera également utile à d'autres utilisateurs, par exemple pour la recherche en épidémiologie bactérienne.

L'outil comporte plusieurs volets qui se sont ajoutés progressivement :

- Une base de données proprement dite, qui contient, pour des génomes ou des chromosomes entièrement séquencés (virus, archées, bactéries, eucaryotes, plasmides et organelles), l'ensemble des répétitions en tandem détectées par le logiciel Tandem Repeats Finder (TRF) (Benson 1999). La base permet d'effectuer des requêtes sur cette collection de répétitions en tandem, visant par exemple à sélectionner des candidats dont on testera le polymorphisme. Elle contient également des informations sur les répétitions en tandem déjà étudiées. Ce dernier aspect nécessite un travail éditorial manuel (étude de la littérature) et les données incluses ne sont donc pas exhaustives.

- Une base de données similaire, mais permettant la comparaison de génomes très proches (par exemple, lorsque deux souches bactériennes de la même espèce ont été séquencées), ce qui permet d'identifier les répétitions en tandem polymorphes.

- Des fonctionnalités permettant la recherche d'un locus particulier, soit par mots-clés, soit par similitude de séquence, en utilisant le logiciel « BLAST ».

- Une page permettant l'identification de souches après génotypage, par comparaison avec des données que nous avons produites, et rendues accessibles dans une base de génotypes.

J'ai effectué les premiers développements de cette ressource à l'occasion de mon travail de DEA. Ces développements apparaissent dans la revue (Vergnaud & Denoeud 2000) présentée au chapitre 2.3.1. Les développements ultérieurs m'ont permis d'être associée à plusieurs

articles d'épidémiologie moléculaire du laboratoire (chapitre 2.2). La version actuelle du site fait l'objet de deux publications (présentation au congrès ECCB 2003 et article paru dans BMC Bioinformatics (Denoeud 2004)), présentées au chapitre 2.1.2. Dans la section suivante, je vais décrire plus en détail la mise en place des différentes fonctionnalités de la base de données, accessibles sur le web à l'adresse [http://minisatellites.u-psud.fr].

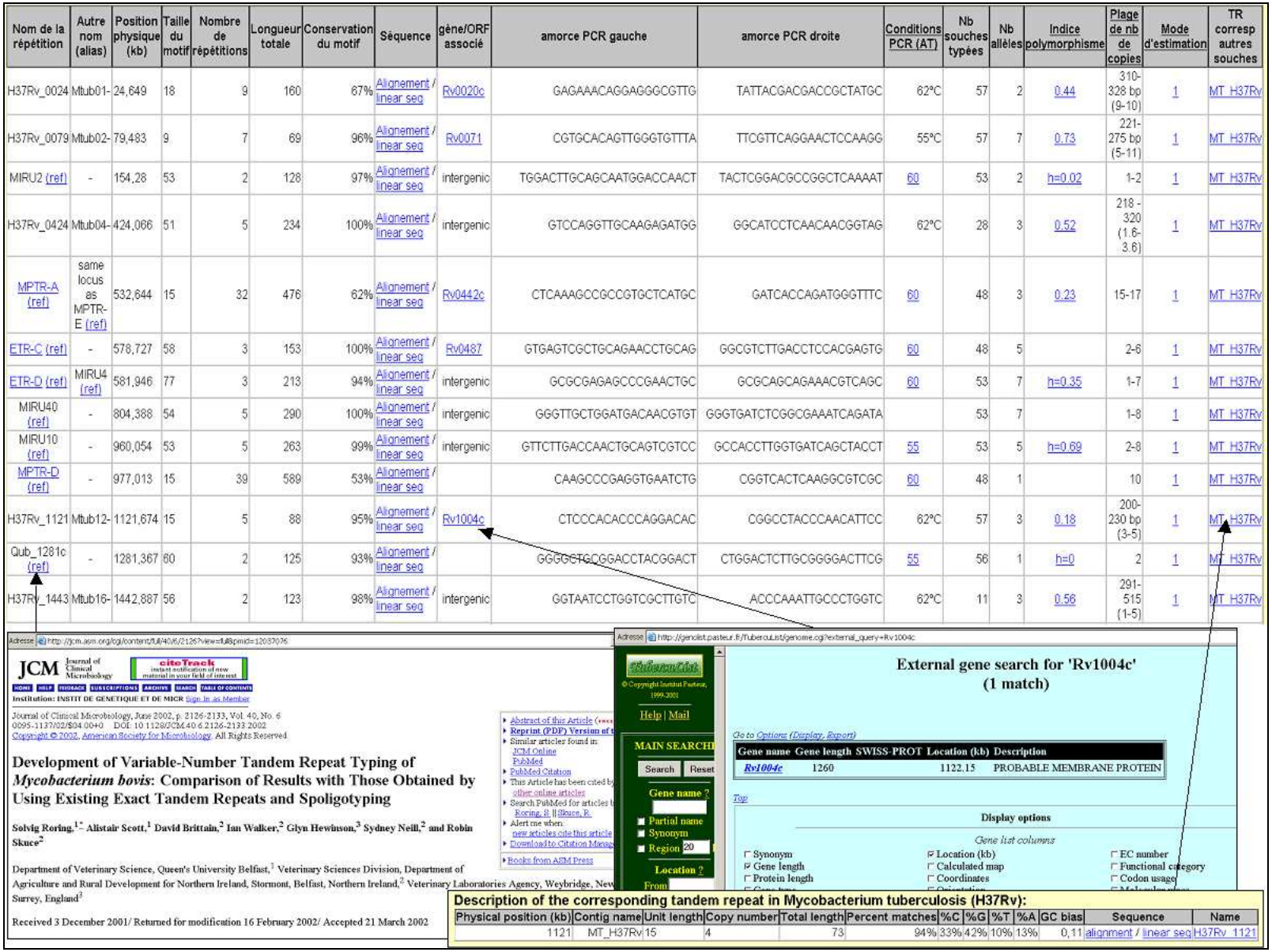
2.1.1 Élaboration de la base de données

2.1.1.1 Traitement des séquences et import dans la base de données

La base de données tire parti des données de séquences rendues publiques, qui sont pré-traitées afin d'identifier les répétitions en tandem. Lorsqu'une requête est lancée dans la base, il n'y a donc pas de nouvelle analyse. Les séquences de chromosomes complets ont été rapatriées à partir du serveur ftp du NCBI [ftp://ftp.ncbi.nih.gov/genomes/]. Elles ont ensuite été soumises au logiciel Tandem Repeats Finder (TRF) (Benson 1999) [http://tandem.bu.edu/trf/trf.html] (voir paragraphe 1.3.2.1) avec les paramètres : « match, mismatch, indels » : (2,3,5) ; « Minimum Alignment Score To Report Repeat » : 50 ; « Maximum Period Size » : 500. Afin de générer les fichiers à importer dans la base de données, j'ai développé une procédure, implémentée en Perl. Son rôle principal est d'éliminer la redondance présente dans les sorties du TRF. En effet, une répétition en tandem peut assez souvent être vue de différentes façons, en terme de taille de motif, et de point de départ et de terminaison, ce qui génère des ambiguïtés : par exemple 10 fois 20 pb et 5 fois 40 pb. Cependant, pour des raisons de simplicité évidentes, il a paru préférable que chaque répétition en tandem n'apparaisse (au moins en première analyse) que sous une seule forme. Cette élimination de la redondance a donc fait l'objet de choix. Nous avons décidé de conserver la répétition en tandem ayant le plus court motif, parmi les répétitions en tandem les plus étendues et ayant la meilleure conservation interne, comme expliqué dans (Denoeud 2003). L'Annexe 1 présente les fonctions relatives au traitement de la redondance extraites du programme Perl générant les fichiers à importer dans la base de données. Nous utilisons une base de données Access, sous Windows 2000 Server.

2.1.1.2 Interrogation de la base de données

La base de données est interrogeable sur le web, à l'adresse [http://minisatellites.u-psud.fr]. La procédure d'interrogation utilise la technologie ASP (« Active Server Pages »), avec VBscript et Perlscript.



Nom de la répétition	Autre nom (alias)	Position physique (kb)	Taille du motif	Nombre de répétitions	Longueur totale	Conservation du motif	Séquence	gène/ORF associé	amorce PCR gauche	amorce PCR droite	Conditions PCR (AT)	Nb souches typées	Nb allèles	Indice polymorphisme	Plage de nb de copies	Mode d'estimation	TR corresp autres souches
H37Rv_0024	Mtub01-	24,649	18	9	160	67%	Alignement / linear seq	Rv0020c	GAGAAACAGGAGGGCGTTG	TATTACGACGACCGCTATGC	62°C	57	2	0.44	310-328 bp (9-10)	1	MT_H37Rv
H37Rv_0079	Mtub02-	79,483	9	7	69	96%	Alignement / linear seq	Rv0071	CGTGCACAGTTGGGTGTTTA	TTCGTTCAGGAACTCCAAGG	55°C	57	7	0.73	221-275 bp (5-11)	1	MT_H37Rv
MIRU2 (ref)	-	154,28	53	2	128	97%	Alignement / linear seq	intergenic	TGGACTTGCAGCAATGGACCAACT	TACTCGGACGCCGGCTCAAAAT	60	53	2	h=0.02	1-2	1	MT_H37Rv
H37Rv_0424	Mtub04-	424,066	51	5	234	100%	Alignement / linear seq	intergenic	GTCCAGGTTGCAAGAGATGG	GGCATCCTCAACAACGGTAG	62°C	28	3	0.52	218 - 320 (1.6-3.6)	1	MT_H37Rv
MPTR-A (ref)	same locus as MPTR-E (ref)	532,644	15	32	476	62%	Alignement / linear seq	Rv0442c	CTCAAAGCCGCCGTGCTCATGC	GATCACCAGATGGGTTTC	60	48	3	0.23	15-17	1	MT_H37Rv
ETR-C (ref)	-	578,727	58	3	153	100%	Alignement / linear seq	Rv0487	GTGAGTCGCTGCAGAACCTGCAG	GGCGTCTTGACCTCCACGAGTG	60	48	5		2-6	1	MT_H37Rv
ETR-D (ref)	MIRU4 (ref)	581,946	77	3	213	94%	Alignement / linear seq	intergenic	GCGCGAGAGCCCGAACTGC	GCGCAGCAGAAACGTCAGC	60	53	7	h=0.35	1-7	1	MT_H37Rv
MIRU40 (ref)	-	804,388	54	5	290	100%	Alignement / linear seq	intergenic	GGGTTGCTGGATGACAACGTGT	GGGTGATCTCGGCGAAATCAGATA		53	7		1-8	1	MT_H37Rv
MIRU10 (ref)	-	960,054	53	5	263	99%	Alignement / linear seq	intergenic	GTTCTTGACCAACTGCAGTCGTCC	GCCACCTTGGTGATCAGCTACCT	55	53	5	h=0.69	2-8	1	MT_H37Rv
MPTR-D (ref)	-	977,013	15	39	589	53%	Alignement / linear seq		CAAGCCCGAGGTGAATCTG	CGGTCACTCAAGGCGTCGC	60	48	1		10	1	MT_H37Rv
H37Rv_1121	Mtub12-	1121,674	15	5	88	95%	Alignement / linear seq	Rv1004c	CTCCCACACCCAGGACAC	CGGCCTACCCAACATTCC	62°C	57	3	0.18	200-230 bp (3-5)	1	MT_H37Rv
Qub_1281c (ref)	-	1281,367	60	2	125	93%	Alignement / linear seq		GGGGGTCGGGACCTACGGACT	CTGGACTCTTGCGGGGACTTCG	55	56	1	h=0	2	1	MT_H37Rv
H37Rv_1443	Mtub16-	1442,887	56	2	123	98%	Alignement / linear seq	intergenic	GGTAATCCTGGTCGCTTGTC	ACCCAAATTGCCCTGGTC	62°C	11	3	0.56	291-515 (1-5)	1	MT_H37Rv

Figure 18 : Une requête dans la base de données des répétitions en tandem.

La Figure 18 illustre le résultat d'une requête via la page de recherche de répétitions en tandem déjà étudiées, appliquée à *Mycobacterium tuberculosis*. Cette requête fournit les caractéristiques des répétitions en tandem, les conditions de typage (amorces PCRs, température d'hybridation) et le polymorphisme. Elle comporte également des liens vers les articles correspondants, les informations sur les gènes associés, et, quand plusieurs souches sont présentes dans la base, les informations sur la(les) répétition(s) en tandem correspondante(s) dans la(les) autre(s) souche(s). Les autres requêtes pouvant être effectuées dans la base de données seront détaillées dans les articles présentés dans les sections suivantes (Vergnaud 2000 ; Le Flèche 2001 ; Le Flèche 2002 ; Denoeud 2003 ; Denoeud 2004).

2.1.1.3 BLAST dans la base de données

BLAST (« basic local alignment search tool ») (Altschul 1990 ; Altschul 1994 ; Altschul 1997) est le logiciel de recherche d'homologie et d'alignement de séquences le plus connu et le plus utilisé. Il fonctionne en trois grandes étapes :

- le pré-traitement de la séquence requête (où tous les mots (w) de longueur donnée de la séquence requête sont comparés, et les mots similaires -selon un seuil T- sont identifiés)
- la génération des « hits » (couples de mots similaires entre les deux séquences)
- l'extension des « hits » (afin d'obtenir des segments de similitude de score supérieur à un seuil S, ou HSP : « high scoring segment pairs »).

Lors de la comparaison de deux séquences, le HSP obtenant le meilleur score est appelé MSP (« Maximal Segment Pair ») : on peut lui associer une P-value, soit la probabilité d'obtenir par hasard un score au moins supérieur ou égal à celui du MSP.

La page de BLAST dans la base de données fonctionne via un script Perl, qui fait appel au programme de BLAST téléchargé sur le site ftp du NCBI [ftp://ftp.ncbi.nih.gov/blast/executables/]. Les requêtes BLAST peuvent être effectuées dans les répétitions en tandem et leurs flanquantes, ce qui permet d'identifier des « familles » de répétitions en tandem. Ces familles peuvent avoir un motif similaire mais des flanquantes différentes : c'est par exemple le cas des « MIRUs » (mycobacterial interspersed repetitive units) observés chez les mycobactéries (Supply 1997). Elles peuvent aussi être constituées de régions plus vastes telles que des éléments IS contenant des répétitions en tandem, pour lesquels le motif répété et les flanquantes seront conservés.

Afin de faciliter la recherche d'amorces PCRs, une page dédiée au BLAST de couples d'amorces a également été créée [http://minisatellites.u-psud.fr/Blast]. Elle fournit la taille des produits PCRs attendus dans les différents souches où les amorces matchent. Un extrait du script Perl correspondant est présenté en Annexe 2.

2.1.1.4 Comparaison de souches

Lorsque les séquences de plusieurs souches d'une même espèce bactérienne sont disponibles, on peut effectuer la comparaison de ces souches. Il s'agit de trouver les correspondances entre les répétitions en tandem des différents génomes, ce qui repose sur la comparaison des souches deux à deux. Si plus de deux souches sont disponibles, on effectuera alors une synthèse de plusieurs comparaisons deux à deux pour générer les correspondances entre les répétitions en tandem.

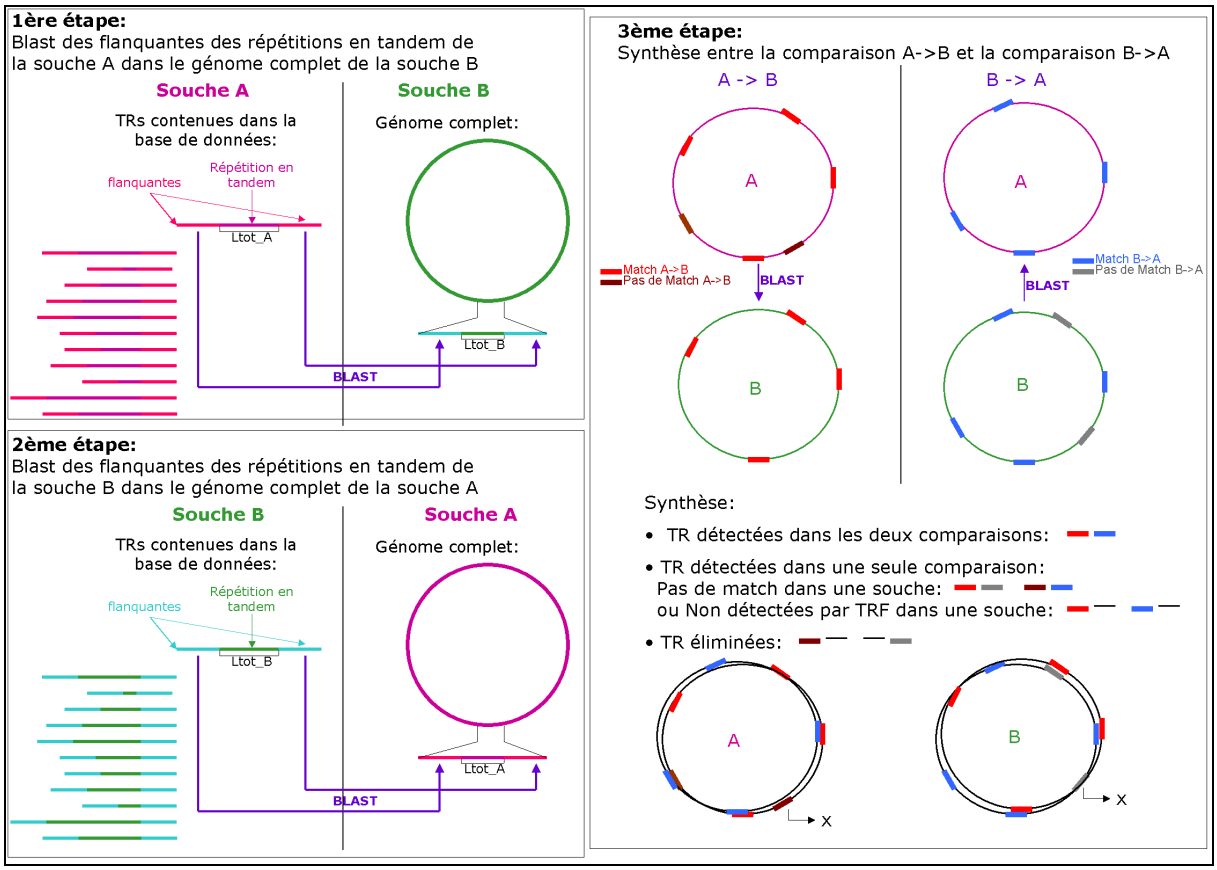


Figure 19 : Méthode de comparaison de souches bactériennes, basée sur le Blast des flanquantes des répétitions en tandem.

La Figure 19 présente la méthode utilisée pour comparer deux souches. Elle repose sur la recherche par BLAST (Altschul 1997) des flanquantes des répétitions en tandem d'une souche dans le génome complet de l'autre souche, et réciproquement. En effet, d'une part, il n'est pas possible d'utiliser la répétition elle-même pour trouver le locus homologue dans la deuxième souche avec un outil tel que BLAST : certains motifs répétés peuvent avoir des similitudes suffisantes pour générer des matchs entre des répétitions en tandem non-apparentées. Ceci avait été illustré par l'hybridation d'une sonde synthétique constituée d'un motif élémentaire aléatoire répété en tandem sur des Southern Blots du génome humain (démarche qui peut s'apparenter au BLAST). De telles sondes identifient plusieurs locus polymorphes indépendants (Vergnaud 1989). D'autre part les BLASTs doivent être effectués de façon réciproque afin de ne pas passer à côté des répétitions en tandem qui n'auraient été détectées

(avec les seuils utilisés lors du traitement par le TRF) que dans une des deux souches, l'autre contenant une répétition trop courte ou insuffisamment conservée. Cette méthode nécessite une homologie suffisante entre les flanquantes, c'est-à-dire que les souches comparées soient peu divergées. Cette contrainte n'en est pas une en pratique, étant donné que le typage par PCR du polymorphisme de répétitions en tandem n'a en général de sens qu'à l'intérieur d'une espèce, pour distinguer des souches bactériennes ou des individus. Les programmes de comparaison ont été implémentés en Perl. Les résultats des différentes comparaisons ont été importés dans la base de données qui peut être interrogée via la page de comparaison de souches, accessible à l'adresse [http://minisatellites.u-psud.fr/comparison], et permet entre-autres d'identifier les répétitions en tandem polymorphes, c'est-à-dire de longueur variable entre les souches comparées (Denoeud 2004). Dans les cas où plus de deux souches sont disponibles, une étape supplémentaire permet d'obtenir les correspondances entre les répétitions en tandem de toutes les souches. Il s'agit d'effectuer une synthèse entre plusieurs comparaisons de deux souches : par exemple, pour comparer A, B, C et D, les comparaisons AB, AC et AD sont réalisées puis synthétisées en utilisant les positions sur le génome A qui sert ici de référence.

2.1.1.5 Page de génotypage de bactéries

La page de génotypage, dédiée à l'identification de souches, repose en grande partie sur le logiciel Bionumerics (Applied Maths), qui gère une base de données de génotypes (les données brutes étant les images de gels, qui sont analysées par Bionumerics). Ce logiciel permet entre-autres de classifier les souches suivant différents « caractères » : répétitions en tandem polymorphes dans notre cas. Le logiciel Bionumerics fournit également des outils pour interroger la base de données de génotypes depuis Internet. Pour cela, l'utilisateur saisit un génotype, et le serveur Bionumerics retourne la liste des souches ayant le génotype le plus proche du génotype saisi parmi ceux présents dans la base de données hébergée au laboratoire. Cette procédure est effectuée par des scripts propres au logiciel Bionumerics. Cependant, certaines améliorations d'interface méritaient d'y être apportées. En effet, la page de saisie utilisée par Bionumerics contient une zone de texte dans laquelle copier-coller les données de génotypage, ce qui présente les inconvénients d'être peu convivial, et surtout, de laisser trop de « liberté » à l'utilisateur. Les programmes de classification utilisés considèrent comme différents des allèles n'ayant qu'une paire de bases d'écart, il est donc prudent de discrétiser les données qui leur sont transmises. Nous aurions pu choisir de traiter les allèles entrés afin de les faire correspondre à l'allèle le plus proche dans notre base de génotypes, mais nous préférons laisser ce soin à l'utilisateur, afin de lui permettre d'analyser de nouveaux allèles, absents de notre base. J'ai donc créé une page de saisie dans laquelle une liste d'allèles possibles est proposée, mais qui permet également de saisir de nouveaux allèles. La plupart des données de génotypage interrogeables par la page d'identification correspondent à des allèles exprimés en nombres de copies et non en longueur totale. La page d'identification permet une saisie en taille (pb) ou en nombre de copies, et effectue la

conversion requise pour permettre la comparaison avec les génotypes de la base. La page d'identification de souches utilise la technologie ASP pour générer la page de requête utilisée par Bionumerics. Le fonctionnement de la page de génotypage de bactéries est décrit et illustré dans les articles (Denoeud 2004) et (Le Flèche 2002) présentés plus loin (chapitres 2.1.2 et 2.2.2).

2.1.2 Fonctionnalités de la base de données

L'article suivant, intitulé « Resources for Bacterial Strain Identification Using Polymorphic Tandem Repeats » (ressources pour l'identification de souches bactériennes grâce aux répétitions en tandem polymorphes) a fait l'objet d'une présentation orale au congrès ECCB2003 : « European Conference on Computational Biology » (27-30 septembre 2003, Paris). Cet article court présente les différentes fonctionnalités de la base de données des répétitions en tandem, pour une application à l'épidémiologie bactérienne.

Résumé :

Chez les bactéries pathogènes, pour lesquelles les études épidémiologiques requièrent une identification précise des souches, les répétitions en tandem se sont avérées être une source de marqueurs très informatifs. Nous avons développé des ressources Internet qui mettent à profit la disponibilité des séquences de génomes bactériens afin de faciliter l'identification de souches basée sur les répétitions en tandem. La première fonctionnalité, la base de données des répétitions en tandem elle-même, permet l'identification de répétitions en tandem au sein de génomes entiers. La deuxième fonctionnalité, la page de comparaison de souches, permet l'identification directe de répétitions en tandem polymorphes pour les espèces dont plusieurs souches ont déjà été séquencées. Afin de faciliter le développement de nouveaux marqueurs, de rechercher des locus similaires, ou d'identifier des locus déjà caractérisés, il est également possible de lancer des requêtes BLAST dans la base de données. Enfin, lorsque qu'une procédure de typage de répétitions en tandem a été développée, les génotypes peuvent être rendus accessibles, comme nous l'avons fait au laboratoire avec la page de génotypage bactérien, un service Internet d'identification de souches.

Resources for Bacterial Strain Identification Using Polymorphic Tandem Repeats

France Denœud [1] and Gilles Vergnaud [1,2]

[1] Laboratoire GPMS, Institut de Génétique et Microbiologie, Bat 400, Université Paris-Sud, 91405 Orsay cedex, France

France.Denoeud@igmors.u-psud.fr

[2] Centre d'Etudes du Bouchet, BP3, 91710 Vert le Petit, France

Keywords. VNTR, tandem repeat, polymorphism, bacteria, strain identification, genotyping, database, epidemiology

Abstract

In pathogenic bacteria, where precise identification at the strain level is essential for epidemiological purposes, tandem repeats have been shown to provide a source of very informative markers. We have developed an internet-based resource which takes advantage of available genome sequences to help develop and run tandem repeats based strain identifications. The first utility, the Tandem Repeats Database *per se*, enables the identification of tandem repeats across entire genomes. The second utility, the Strain Comparison Page, makes direct identification of polymorphic tandem repeats possible for species in which already more than one strain has been sequenced. In order to facilitate the development of new markers, to search for related loci across different genomes, or to identify independently characterised loci, the possibility to run Blast searches in the tandem repeats database has been implemented. Finally, once a tandem repeat typing assay has been developed, genotyping data can be made accessible and directly queried as illustrated for instances developed in our laboratory by the Bacterial Genotyping Page, a web-based service for strain identification.

Introduction

Epidemiological analyses of infections caused by pathogenic bacteria depend on the accurate identification of strains. Tandem repeats have been shown to provide a source of very informative markers in various bacterial species [1], [2]. The tandem repeats database (accessible at http://minisatellites.u-psud.fr) enables the identification of tandem repeats across entire genomes [3], [4], so that all that is required to obtain informative markers is experimental testing for polymorphism (at present at least, the polymorphism associated with a bacterial tandem repeat still cannot be easily inferred from its primary sequence [4][5][6]). In some instances, this still represents a significant task. The Strain Comparison Page takes advantage of the availability of genome sequences from more than one strain from a growing number of species, and directly identifies tandem repeats differing in size between the sequenced strains. The Blast in the Tandem Repeat Database page facilitates the search for a known tandem repeat, the prediction of PCR amplification products, the verification of primer specificity, and more generally the validation of amplification primers choices. Once the informative tandem repeats have been typed on a relevant set of strains, it is relatively easy to set-up databases of genotypes for localuse, or to be queried across the internet. The Bacterial Genotyping Page illustrates a freely accessible, fast, and easy to use, internet-based service for strain identification.

The Strain Comparison Page (http://minisatellites.u-psud.fr/comparison)

The comparison of two strains is based on a pre-computed BLASTing [7] of the flanking sequences of tandem repeats from each strain against the other. The search must run in both ways to avoid missing some tandem repeats that are absent in the tandem repeats database for one strain because they were not detected by the Tandem Repeats Finder [8]. This occurs for instance when there is only one copy of the repeated unit in one strain, or even a deletion, or more generally because in one strain the locus is not identified by TRF as a significant tandem repeat. The database set-up allows queries in the tandem repeats database according to the length difference between the two strains, as well as other tandem repeats characteristics (unit length, copy number, etc.). It is possible to further select tandem repeats with length difference a multiple of the repeat unit length. The query "length difference $\geq$ 9 bp and unit length $\geq$ 9 bp" applied to *Mycobacterium tuberculosis* strains H37Rv and CDC1551 identifies 54 tandem repeats, 31 of which are already known as informative markers [2], meaning that detected length differences between the strains really correspond to variable number of tandem repeats, and not to sequencing errors in one of the strains.

For a number of bacteria, sequences are available for more than two strains: we thus developed a Comparison Page for more than two strains, that synthesizes the results of several 2-strains comparisons.

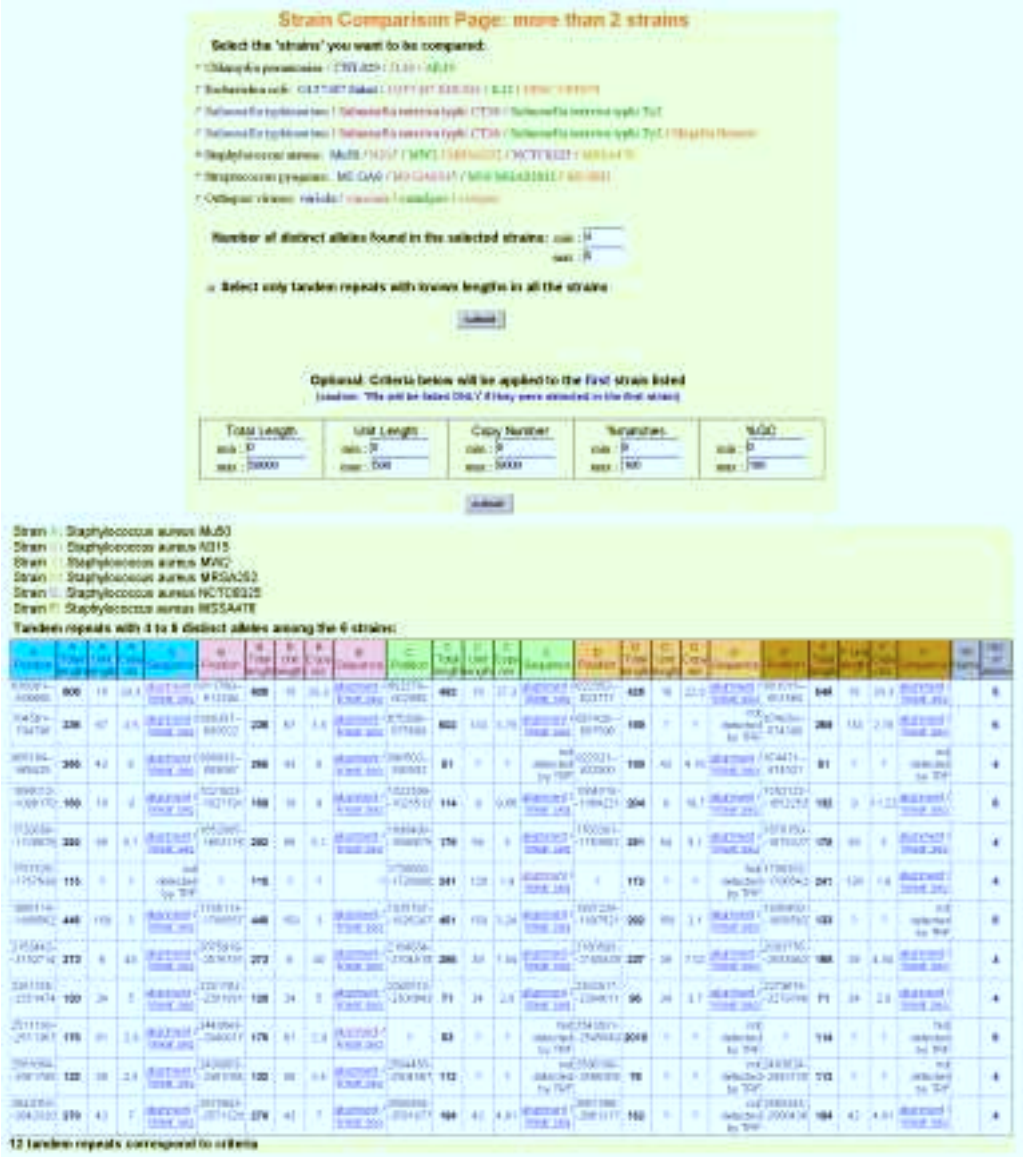


Fig. 1 A query using the Strain Comparison Page

Figure 1 illustrates a query made for *Staphylococcus aureus*, for which the genome sequence from six different strains is available. Twelve tandem repeats with at least four alleles among the six strains are identified. They are likely to be the more informative ones to discriminate *Staphylococcus aureus* strains. Question marks indicate that no tandem repeat was identified in the corresponding strain at that locus.

Of course, the results obtained depend on the phylogenetic distance between the strains compared. If the strains are very closely related, only a small proportion of the tandem repeats polymorphic within the species will be identified, but these tandem repeats will probably be the most polymorphic ones. Conversely, if the strains are distant relatives within the species of interest,, a larger proportion of polymorphic tandem repeats will be found, some of which will be only moderately informative. Obviously, in this respect and because of its major public health significance, *S. aureus* constitutes a perfect model : some strains are closely related (Mu50 and N315 for instance) and some others are more distant (NCTC8325 for instance).

The Blast in the Tandem Repeats Database Page (http://minisatellites.u-psud.fr/Blast)

To facilitate the identification of already studied tandem repeats, we implemented BLAST [7] against the tandem repeats from the database, *i.e.* the tandem repeats themselves and their flanking sequences. All bacteria can be blasted at once, which allows the identification of tandem repeats families conserved in several bacterial species,

even when they share a similar consensus motif but different flanking sequences: insertions/deletions due to tandem repeats polymorphism generally do not affect the recognition. Another page is dedicated to the Blast of PCR primers: it provides the size of the PCR products in all the species/strains where the primers match, and thus facilitates the primer picking step.

The Bacterial Genotyping Page (http://bacterial-genotyping.igmors.u-psud.fr)

Once the tandem repeats identified via the Tandem Repeats Database or the Strain Comparison Page have been typed by PCR, and genotypes have been assigned to a sufficient number of strains, the next step is to build a strain classification to evaluate the proximity of the strains. But, often, such studies remain internal to a laboratory, despite the high portability of tandem repeats typing assay. The Bacterial Genotyping Page aims to remedy this lack : it allows external users to query a number of genotyping data (*Bacillus anthracis*, *Yersinia pestis*, *Mycobacterium tuberculosis* for the moment) in order to identify bacterial strains [2]. For each locus, allele sizes can be selected among a list of possibilities (observed sizes). The results of the query indicate a similarity score and include links to the complete data for each strain listed. If the tandem repeats typing approach is considered to be of use by the community interested in a particular species, and given that the associated data is highly portable, it should then be relatively easy, through collaborative efforts, to significantly expand the available data. It is hoped that this data will constitute an easy-to-use high-resolution classification resource which will then help address medical and epidemiological issues.

References

[1] A. van Belkum, S. Scherer, W. van Leeuwen, D. Willemse, L. van Alphen and H. Verbrugh, Variable number of tandem repeats in clinical strains of *Haemophilus influenzae*. *Infect Immun* 65:5017-27, 1997.

[2] P. Le Flèche, M. Fabre, F. Denœud, J.L. Koeck and G. Vergnaud, High resolution, on-line identification of strains from the *Mycobacterium tuberculosis* complex based on tandem repeat typing. *BMC Microbiol* 2:37, 2002.

[3] G. Vergnaud and F. Denœud, Minisatellites: Mutability and Genome Architecture. *Genome Res* 10:899-907, 2000.

[4] P. Le Flèche, Y. Hauck, L. Onteniente, A. Prieur, F. Denœud, V.Ramisse, P. Sylvestre, G. Benson, F. Ramisse and G. Vergnaud, A tandem repeats database for bacterial genomes: application to the genotyping of *Yersinia pestis* and *Bacillus anthracis*. *BMC Microbiol* 1:2, 2001.

[5] F. Denoeud, G. Vergnaud and G. Benson, Predicting human minisatellite polymorphism. *Genome Res* 13(5):856-67, 2003.

[6] C. Pourcel, Y. Vidgop, F. Ramisse, G. Vergnaud, and C. Tram, Characterization of a tandem repeat polymorphism in Legionella pneumophila and its use for genotyping. *J Clin Microbiol*. 41(5):1819-26, 2003.

[7] S.F. Altschul, T.L Madden, A.A Schaffer, J. Zhang, Z. Zhang, W. Miller and D.J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25:3389-402, 1997.

[8] G. Benson, Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* 27:573-580, 1999.

L'article suivant (Denoeud 2004), intitulé « Identification of polymorphic tandem repeats by direct comparison of genome sequence from different bacterial strains : a Web-based ressource » (identification de répétitions en tandem polymorphes par la comparaison directe des séquences génomiques de différentes souches bactériennes : une ressource web) présente les fonctionnalités de la base de données utiles à l'épidémiologie bactérienne. Il insiste plus particulièrement sur la page de comparaison de souches, qui identifie directement les répétitions en tandem polymorphes entre plusieurs souches séquencées. Cet article aborde également le problème récurrent de la prédiction du polymorphisme d'une répétition en tandem à partir de sa séquence. Nous y montrons que certaines caractéristiques des séquences de répétitions en tandem bactériennes (conservation interne, nombre de copies) sont corrélées au polymorphisme mais ont une valeur prédictive très variable d'une espèce à l'autre. La meilleure approche pour identifier des répétitions en tandem polymorphes chez les bactéries reste donc la comparaison de souches, quand celle-ci est rendue possible par la disponibilité des séquences.

Résumé :

Contexte : Les répétitions en tandem polymorphes sont utilisées pour le typage de souches chez un nombre grandissant de bactéries pathogènes pour l'Homme, dont *Bacillus anthracis*, *Mycobacterium tuberculosis, Yersinia pestis*, *Pseudomonas aeruginosa*, *Legionella pneumophila*, *Salmonella typhi*, *Brucella*, *Francisella tularensis* et *Staphylococcus aureus*. Les répétitions en tandem peuvent facilement être identifiées à partir des nombreuses séquences de génomes bactériens actuellement disponibles. La mise en place d'une procédure de génotypage ne nécessite donc plus que l'évaluation de leur polymorphisme, ce qui est fait en général par des tests systématiques. Ces tests peuvent toutefois représenter une tâche laborieuse. Pour de nombreux pathogènes d'importance, tels que *S. aureus*, plus d'une souche a été séquencée, ce qui permet dorénavant d'identifier *in silico* les répétitions en tandem polymorphes parmi différentes souches.

Résultats : En supplément de la base de données des répétitions en tandem déjà décrite, nous avons développé une page d'identification automatique des répétitions en tandem de longueur différente dans les génomes de plus de deux souches bactériennes proches. Les comparaisons de génomes sont effectuées puis importées dans une base de données, qui peut être interrogée par Internet selon des critères d'intérêt tels que la longueur du motif, la différence de taille prédite, etc. Les comparaisons sont disponibles pour 16 espèces bactériennes et les virus du groupe orthopox, comprenant le virus de la variole et trois de ses voisins proches.

Conclusions : Nous présentons une ressource Internet qui facilite le développement de méthodes de typage de souches bactériennes à partir des répétitions en tandem. Elle comprend actuellement quatre parties, accessibles à partir de l'adresse http://minisatellites.u-psud.fr. La base de données des répétitions en tandem permet d'identifier les répétitions en tandem dans des génomes entiers. La page de comparaison de souches sélectionne les répétitions en tandem différentes entre plusieurs génomes d'une même espèce. La page de Blast dans la base de données facilite la recherche de répétitions en tandem connues et la validation des couples d'amorces PCR. La page de génotypage permet l'identification de souches en ligne.

BMC Bioinformatics

Database

Identification of polymorphic tandem repeats by direct comparison of genome sequence from different bacterial strains : a web-based resource

France Denœud*[1] and Gilles Vergnaud[1,2]

Address: [1]Laboratoire GPMS, Institut de Génétique et Microbiologie, Bat 400, Université Paris-Sud, 91405 Orsay cedex, France and [2]Centre d'Etudes du Bouchet, BP3, 91710 Vert le Petit, France

Email: France Denœud* - France.Denoeud@igmors.u-psud.fr; Gilles Vergnaud - Gilles.Vergnaud@igmors.u-psud.fr

* Corresponding author

Published: 12 January 2004

BMC Bioinformatics 2004, **5**:4

Received: 24 September 2003

Accepted: 12 January 2004

This article is available from: http://www.biomedcentral.com/1471-2105/5/4

Abstract

Background: Polymorphic tandem repeat typing is a new generic technology which has been proved to be very efficient for bacterial pathogens such as *B. anthracis*, *M. tuberculosis*, *P. aeruginosa*, *L. pneumophila*, *Y. pestis*. The previously developed tandem repeats database takes advantage of the release of genome sequence data for a growing number of bacteria to facilitate the identification of tandem repeats. The development of an assay then requires the evaluation of tandem repeat polymorphism on well-selected sets of isolates. In the case of major human pathogens, such as *S. aureus*, more than one strain is being sequenced, so that tandem repeats most likely to be polymorphic can now be selected *in silico* based on genome sequence comparison.

Results: In addition to the previously described general Tandem Repeats Database, we have developed a tool to automatically identify tandem repeats of a different length in the genome sequence of two (or more) closely related bacterial strains. Genome comparisons are pre-computed. The results of the comparisons are parsed in a database, which can be conveniently queried over the internet according to criteria of practical value, including repeat unit length, predicted size difference, etc. Comparisons are available for 16 bacterial species, and the orthopox viruses, including the variola virus and three of its close neighbors.

Conclusions: We are presenting an internet-based resource to help develop and perform tandem repeats based bacterial strain typing. The tools accessible at http://minisatellites.u-psud.fr now comprise four parts. The Tandem Repeats Database enables the identification of tandem repeats across entire genomes. The Strain Comparison Page identifies tandem repeats differing between different genome sequences from the same species. The "Blast in the Tandem Repeats Database" facilitates the search for a known tandem repeat and the prediction of amplification product sizes. The "Bacterial Genotyping Page" is a service for strain identification at the subspecies level.

Background

Molecular epidemiology, the integration of molecular typing and conventional epidemiological studies, is likely to add significant value to analyses of infections caused by pathogenic bacteria (see [1] for review). Multilocus Sequence Typing (MLST) for instance is now a major reference method for the molecular epidemiology of *Neisseria meningitidis* and other human pathogens [2]. In this

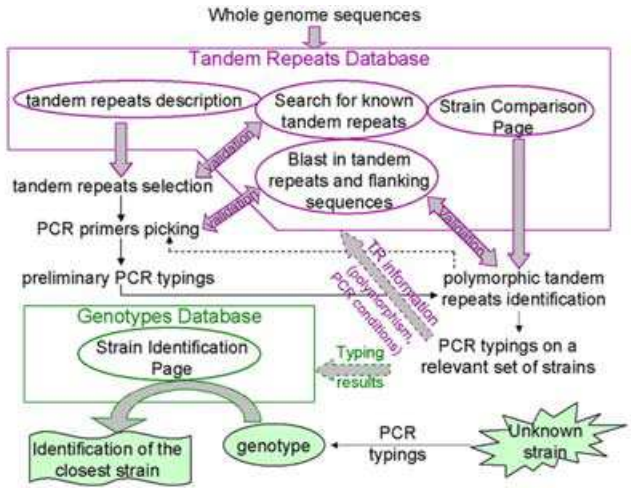
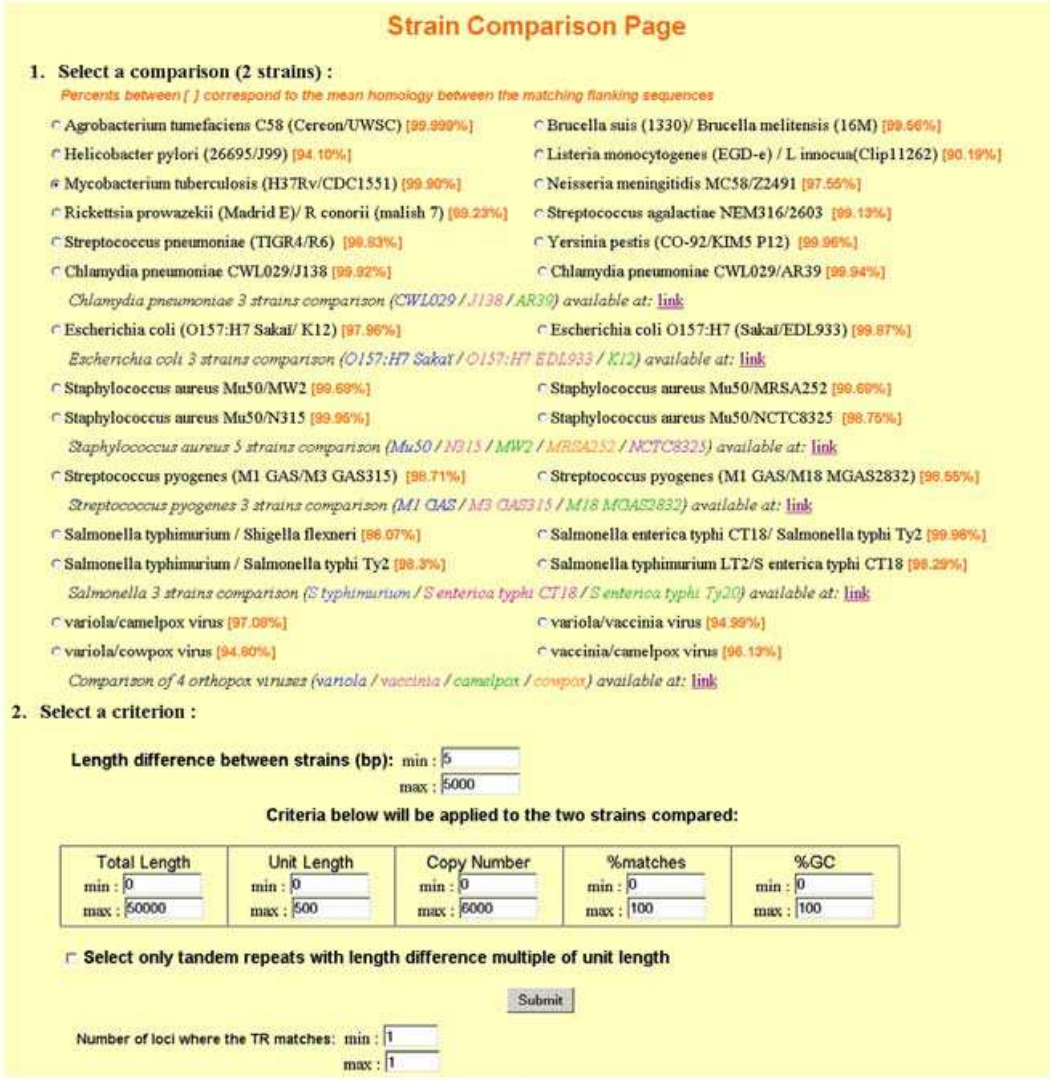


Figure 1
The procedure to find polymorphic tandem repeats for use in strain typing. The steps leading from the release of a complete (or incomplete) genome sequence to the validation of new polymorphic markers are described. The purpose of the web-based tools developed is to facilitate the bioinformatics data-management steps.

kind of assay, a set of typically 7 genes is partially sequenced, and the resulting data is converted into sequence types, which can be easily stored in databases, and compared to others. However a number of significant pathogens, including *M. tuberculosis* [3], *B. anthracis* and *Y. pestis* [4] are not amenable to this approach, because of the recent emergence of these pathogens and the resulting rarity of sequence variations. In these pathogens, tandem repeats (TRs) are a source of very informative markers for strain genotyping [5-10]. Tandem repeats in pathogenic bacteria were initially identified within genes associated with bacterial virulence [11,12]. In other instances, the contribution of tandem repeats to genome polymorphism was established after extensive searches based for instance on AFLP (amplified fragment length polymorphism) profiling. This is well illustrated by *B. anthracis*, in which polymorphic bands in AFLP patterns [13] were subsequently demonstrated by sequencing to be due to tandem repeat variations [14]. Eventually, some of these tandem repeats have been shown to directly contribute to phenotypic variations of the *B. anthracis* exosporium which makes the outer layer of the spores [15]. The frequent observation that tandem repeat-containing genes are often associated with outer membrane proteins suggests that such genes help bacteria adapt to their environment, and may be to some extent mutation hotspots as a result of positive selection.

Comparison	%matches between flanking sequences	% of flanking seqs not rearranged	% of polymorphic tandem repeats	positions of matching tandem repeats
Agrobacterium tumefaciens C58 Cereon/UWSC	99.99	99.95	0	
Salmonella enterica typhi CT18/Ty2	99.96	98.83	2.04	
Yersinia pestis CO-92/KIM5 P12	99.96	98.51	8.47	
Staphylococcus aureus Mu50/N315	99.95	98.05	1.38	
Chlamydia pneumoniae CWL029/AR39	99.94	99.86	1.41	
Chlamydia pneumoniae CWL029/J138	99.92	99.86	2.30	
Mycobacterium tuberculosis H37Rv/CDC1551	99.90	99.50	4.06	
Escherichia coli O157:H7 Sakai/EDL933	99.87	99.49	2.13	
Brucella suis 1330/ *melitensis* 16M	99.56	99.08	7.85	
Streptococcus agalactiae NEM316/2603	99.13	85.44	2.76	
Streptococcus pneumoniae TIGR4/R6	99.83	90.01	16.50	
Staphylococcus aureus Mu50/NCTC8325	98.75	92.35	8.12	
Streptococcus pyogenes M1GAS/M3GAS	98.71	86.30	6.38	
Staphylococcus aureus *Mu50/MW2*	98.68	91.04	8.24	
Staphylococcus aureus Mu50/MRSA252	98.68	90.17	8.81	
Streptococcus pyogenes M1GAS/M18GAS	98.55	89.19	5.82	
Salmonella typhimurium LT2 / typhi CT18	98.29	87.65	5.44	
Escherichia coli O157:H7 Sakai/ K12	97.96	77.23	8.92	
Neisseria meningitidis MC58/Z2491	97.54	92.66	18.88	
Helicobacter pylori 26695/J99	94.10	87.23	18.91	
Listeria monocytogenes EGDe / *innocua* Clip	90.19	74.13	3.99	
Rickettsia prowazekii Madrid E / *conorii malish7*	89.23	61.05	8.23	
Salmonella typhimurium / *Shigella flexneri*	86.06	47.16	6.23	

Figure 2
Comparison of strains using different indexes. The four columns correspond to (from left to right): (1) mean %identity provided by BLAST when the match occurred on more than half the length of the 500 bp of submitted flanking sequence ; (2) proportion (%) of flanking sequences that matched on more than half their length between the two strains ; (3) proportion (%) of tandem repeats of a different size in the two strains ; and (4) plot of the positions of homologous tandem repeat loci in the two genomes which indirectly reflects large scale genome rearrangements. Species are listed according to the first index (mean %identity)



Strain Comparison Page

1. Select a comparison (2 strains) :

Percents between [ ] correspond to the mean homology between the matching flanking sequences

- ○ Agrobacterium tumefaciens C58 (Cereon/UWSC) [99.999%]
- ○ Brucella suis (1330)/ Brucella melitensis (16M) [99.56%]
- ○ Helicobacter pylori (26695/J99) [94.10%]
- ○ Listeria monocytogenes (EGD-e) / L innocua(Clip11262) [90.19%]
- ◉ Mycobacterium tuberculosis (H37Rv/CDC1551) [99.90%]
- ○ Neisseria meningitidis MC58/Z2491 [97.55%]
- ○ Rickettsia prowazekii (Madrid E)/ R conorii (malish 7) [89.23%]
- ○ Streptococcus agalactiae NEM316/2603 [99.13%]
- ○ Streptococcus pneumoniae (TIGR4/R6) [99.83%]
- ○ Yersinia pestis (CO-92/KIM5 P12) [99.96%]
- ○ Chlamydia pneumoniae CWL029/J138 [99.92%]
- ○ Chlamydia pneumoniae CWL029/AR39 [99.94%]

Chlamydia pneumoniae 3 strains comparison (CWL029 / J138 / AR39) available at: link

- ○ Escherichia coli (O157:H7 Sakai/ K12) [97.96%]
- ○ Escherichia coli O157:H7 (Sakai/EDL933) [99.87%]

Escherichia coli 3 strains comparison (O157:H7 Sakai / O157:H7 EDL933 / K12) available at: link

- ○ Staphylococcus aureus Mu50/MW2 [99.69%]
- ○ Staphylococcus aureus Mu50/MRSA252 [99.68%]
- ○ Staphylococcus aureus Mu50/N315 [99.96%]
- ○ Staphylococcus aureus Mu50/NCTC8325 [98.75%]

Staphylococcus aureus 5 strains comparison (Mu50 / N315 / MW2 / MRSA252 / NCTC8325) available at: link

- ○ Streptococcus pyogenes (M1 GAS/M3 GAS315) [98.71%]
- ○ Streptococcus pyogenes (M1 GAS/M18 MGAS2832) [96.55%]

Streptococcus pyogenes 3 strains comparison (M1 GAS / M3 GAS315 / M18 MGAS2832) available at: link

- ○ Salmonella typhimurium / Shigella flexneri [86.07%]
- ○ Salmonella enterica typhi CT18/ Salmonella typhi Ty2 [99.96%]
- ○ Salmonella typhimurium / Salmonella typhi Ty2 [98.3%]
- ○ Salmonella typhimurium LT2/S enterica typhi CT18 [98.29%]

Salmonella 3 strains comparison (S typhimurium / S enterica typhi CT18 / S enterica typhi Ty20) available at: link

- ○ variola/camelpox virus [97.08%]
- ○ variola/vaccinia virus [94.99%]
- ○ variola/cowpox virus [94.60%]
- ○ vaccinia/camelpox virus [96.13%]

Comparison of 4 orthopox viruses (variola / vaccinia / camelpox / cowpox) available at: link

2. Select a criterion :

Length difference between strains (bp): min : 5 max : 5000

Criteria below will be applied to the two strains compared:

Total Length	Unit Length	Copy Number	%matches	%GC
min : 0	min : 0	min : 0	min : 0	min : 0
max : 50000	max : 500	max : 6000	max : 100	max : 100

☐ **Select only tandem repeats with length difference multiple of unit length**

Submit

Number of loci where the TR matches: min : 1 max : 1

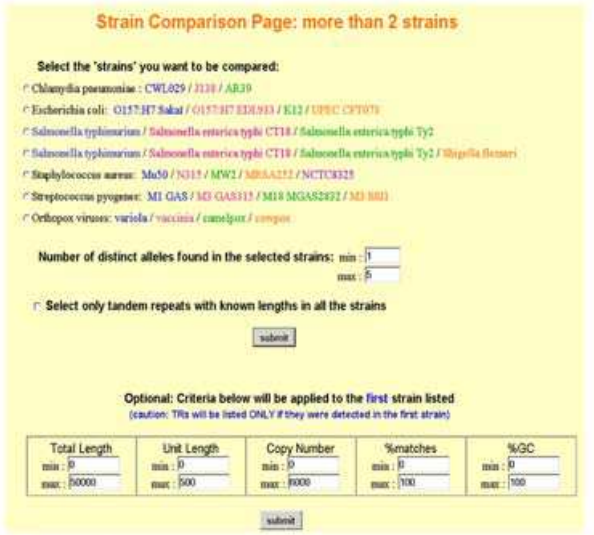
Mycobacterium tuberculosis (H37Rv) / Mycobacterium tuberculosis (CDC1551)

Tandem repeats with length difference between the two strains >= 5 bp and <= 5000 bp

Mycobacterium tuberculosis (H37Rv)									Mycobacterium tuberculosis (CDC1551)										
Position	Total length	Contig	Unit length	Copy nbr	% match	% GC	Sequence	TR name	Position	Total length	Contig	Unit length	Copy nbr	% match	% GC	Sequence	TR name	Length diff	Nbr of matches	Match orientation
24648-24825	**178**	MT_H37Rv	18	10	70%	72%	alignment	H37Rv_0024	24648-24807	**160**	MT_CDC1551	18	9	67%	73%	alignment	H37Rv_0024	**18**	1	+/+
79503-79562	**60**	MT_H37Rv	9	6	96%	77%	alignment	H37Rv_0079	79482-79550	**69**	MT_CDC1551	9	7	96%	77%	alignment	H37Rv_0079	**9**	1	+/+
149881-150864	**984**	MT_H37Rv	9	113	53%	80%	alignment		149870-151033	**1164**	MT_CDC1551	9	134	53%	80%	alignment		**180**	1	+/+
424010-424141	**132**	MT_H37Rv	51	3	100%	75%	alignment	H37Rv_0424	424065-424298	**234**	MT_CDC1551	51	5	100%	74%	alignment	H37Rv_0424	**102**	1	+/+
424871-427298	**2428**	MT_H37Rv	15	268	53%	61%	alignment		425029-427441	**2413**	MT_CDC1551	30	132	59%	61%	alignment		**15**	1	+/+
577284-577494	**211**	MT_H37Rv	58	4	100%	65%	alignment	ETR-C	578726-578878	**153**	MT_CDC1551	58	3	100%	65%	alignment	ETR-C	**58**	1	+/+
802426-802498	**73**						not detected by TRF		804387-804675	**289**	MT_CDC1551	54	5	100%	67%	alignment	MIRU40	**216**	1	+/+
960165-960321	**157**	MT_H37Rv	53	3	100%	68%	alignment	MIRU10	960053-960315	**263**	MT_CDC1551	53	5	99%	67%	alignment	MIRU10	**106**	1	+/+

Figure 3

Example of a query in the Strain Comparison Page. On the top, the query page shows the 28 comparisons currently available (others will be added as new genome sequences are finished and released). Bottom, the result of a query performed for *Mycobacterium tuberculosis* strains H37Rv and CDC1551 is summarized.



Strain Comparison Page: more than 2 strains

Select the 'strains' you want to be compared:

- Chlamydia pneumoniae : CWL029 / J138 / AR39
- Escherichia coli: O157:H7 Sakai / O157:H7 EDL933 / K12 / UPEC CFT073
- Salmonella typhimurium / Salmonella enterica typhi CT18 / Salmonella enterica typhi Ty2
- Salmonella typhimurium / Salmonella enterica typhi CT18 / Salmonella enterica typhi Ty2 / Shigella flexneri
- Staphylococcus aureus: Mu50 / N315 / MW2 / MRSA252 / NCTC8325
- Streptococcus pyogenes: M1 GAS / M3 GAS315 / M18 MGAS8232 / M3 SSI1
- Orthopox viruses: variola / vaccinia / camelpox / cowpox

Number of distinct alleles found in the selected strains: min : 1 max : 5

Select only tandem repeats with known lengths in all the strains

submit

Optional: Criteria below will be applied to the first strain listed
(caution: TRs will be listed ONLY if they were detected in the first strain)

Total Length	Unit Length	Copy Number	%matches	%GC
min : 0 max : 50000	min : 0 max : 500	min : 0 max : 9000	min : 0 max : 100	min : 0 max : 100

submit

Strain A: Escherichia coli o157:H7 (Sakai)
Strain B: Escherichia coli 0157:H7 (EDL933)
Strain C: Escherichia coli K-12
Strain D: Escherichia coli UPEC CFT073

Tandem repeats with 2 to 4 distinct alleles among the 4 strains:

A Position	A Total length	A Unit length	A Copy nbr	A Sequence	B Position	B Total length	B Unit length	B Copy nbr	B Sequence	C Position	C Total length	C Unit length	C Copy nbr	C Sequence	D Position	D Total length	D Unit length	D Copy nbr	D Sequence	TR Name	Nbr of alleles
201--253	**53**	6	10	alignment / linear seq	198--253	**56**	6	10	alignment / linear seq	201--235	**35**	12	3	alignment / linear seq	201--947	**747**	12	3,2	alignment / linear seq		**4**
46590--46724	**135**	57	2,3	alignment / linear seq	46591--46725	**135**	57	2,3	alignment / linear seq	42183--42311	**129**	57	2	alignment / linear seq	44835--44989	**135**	57	2,3	alignment / linear seq		**2**
71495--71683	**189**	85	2,2	alignment / linear seq	71496--71684	**189**	85	2,2	alignment / linear seq	66564--66837	**274**	85	3	alignment / linear seq	69078--69351	**274**	85	3,1	alignment / linear seq		**2**
133896--134273	**378**	96	3,9	alignment / linear seq	133897--134274	**378**	96	3,9	alignment / linear seq	129524--129901	**378**	96	4	alignment / linear seq	139926--140207	**282**	96	2,9	alignment / linear seq		**2**
134629--134668	**40**	21	1,9	alignment / linear seq	134630--134669	**40**	21	1,9	alignment / linear seq	130257--130296	**40**	9	7	alignment / linear seq	140563--140626	**64**	24	3,08	alignment / linear seq		**2**
391255--391705	**451**	93	4,9	alignment / linear seq	391257--391707	**451**	93	4,9	alignment / linear seq	338982--339339	**358**	93	4	alignment / linear seq	430253--430612	**360**	93	3,9	alignment / linear seq		**2**
411014--411748	**735**	94	8	alignment / linear seq	411016--411750	**735**	94	8	alignment / linear seq	353825--353993	**171**	?	?	not detected by TRF	443971--444137	**167**	?	?	not detected by TRF		**3**
861213--861641	**429**	15	29	alignment / linear seq	862867--863295	**429**	15	29	alignment / linear seq	775930--776439	**510**	15	34	alignment / linear seq	799377--799886	**510**	15	34,46	alignment / linear seq		**2**
865089--865244	**156**	78	2	alignment / linear seq	866743--866898	**156**	78	2	alignment / linear seq	779887--779937	**51**	?	?	not detected by TRF	803335--803490	**156**	78	2	alignment / linear seq		**2**
865293--865449	**157**	79	2	alignment / linear seq	866947--867103	**157**	79	2	alignment / linear seq	780191--780445	**256**	79	2	alignment / linear seq	803539--803597	**59**	79	2	alignment / linear seq		**3**
1065508--1065583	**76**	39	1,9	alignment / linear seq	1154731--1154808	**76**	39	1,9	alignment / linear seq	933654--933729	**76**	39	2	alignment / linear seq	988559--988673	**115**	39	3,1	alignment / linear seq		**2**
1066515--1066790	**276**	39	7	alignment / linear seq	1155738--1156013	**276**	39	7	alignment / linear seq	934661--934897	**237**	39	6	alignment / linear seq	989605--989856	**252**	3	84	alignment / linear seq		**3**
1520733--1520758	**26**	6	4,3	alignment / linear seq	1605820--1605852	**33**	6	6	alignment / linear seq	1160861--1160874	**14**	?	?	not detected by TRF	1305139--1305152	**14**	?	?	not detected by TRF		**3**
1678733--1678808	**76**	33	2,3	alignment / linear seq	1770139--1770214	**76**	33	2,3	alignment / linear seq	1236469--1236511	**43**	?	?	not detected by TRF	1474673--1474715	**43**	?	?	not detected by TRF		**2**
3157084--3157367	**284**	91	3,1	alignment / linear seq	3227038--3227321	**284**	91	3,1	alignment / linear seq	3428933--3429034	**102**	?	?	not detected by TRF	2712440--2712814	**375**	91	4,1	alignment / linear seq		**3**
3255152--3255318	**167**	12	16	alignment / linear seq	3324996--3325150	**155**	12	8	alignment / linear seq	2528793--2528947	**155**	21	8	alignment / linear seq	2811652--2811818	**167**	9	29,44	alignment / linear seq		**2**
3297758--3297794	**37**	14	2,6	alignment / linear seq	3387590--3387626	**37**	14	2,6	alignment / linear seq	2579654--2579690	**37**	14	3	alignment / linear seq	2852392--2852436	**45**	?	?	not detected by TRF		**2**
3297804--3297841	**38**	19	2	alignment / linear seq	3387636--3387673	**38**	19	2	alignment / linear seq	2579700--2579737	**38**	19	2	alignment / linear seq	2852438--2852483	**46**	?	?	not detected by TRF		**2**

Figure 4

Example of a query in the Strain Comparison Page for more than two strains. Top, the query page shows the 6 comparisons currently available (others will be added as new genome sequences are finished and released). Bottom, the result of a query performed for *Escherichia coli* strains O157:H7 Sakaï, O157:H7 EDL933, K12 and UPEC-CFT073 is summarized. In several loci, the size of the repeat is listed differently for the different strains, which is due to different detections by the Tandem Repeats Finder, usually as a result of internal variations within the tandem array. Total length is calculated from positions of matching flanking sequences in the different strains, and does not necessarily correspond to the length of the tandem repeat detected by TRF in the locus. "Number of alleles" refers to the number of predicted sizes differing by at least 5 bp among the strains compared.

Polymorphic tandem repeats (VNTRs, for Variable Number of Tandem Repeats), once identified, provide convenient tools requiring ordinary molecular biology equipment and the data can be easily exchanged and compared. The resulting assay, called MLVA (for multiple locus VNTR analysis) can even be automated [16]. We have developed tools which facilitate the bioinformatics step of genome analysis required to start a project. A previously described Tandem Repeats Database enables the identification of tandem repeats across entire genomes [9,10,17-19]. It has been constantly updated, with now more than a hundred bacterial genomes available, compared to 35 at the onset of the database. We present here a new and major development of this resource which takes advantage of the fact that more than two different strains from the same species have now been sequenced at least for a number of major human pathogens. As a result, the tools accessible over the Internet at http://minisatellites.u-psud.fr now comprise four complementary parts. The newly added resource, the Strain Comparison Page, takes advantage of the availability of genome sequences from more than one strain from a growing number of species to directly identify tandem repeats differing between the sequenced strains. This is of interest because the vast majority of tandem repeats is often not polymorphic [19]. The "Blast in the Tandem Repeats Database" page facilitates the search for a known tandem repeat, the prediction of PCR amplification products size, and the verification of primer specificity. Once an MLVA assay has been set up, and carefully validated by typing collections of isolates, it is relatively easy to construct databases of genotypes to be used locally or which can be queried across the Internet. The "Bacterial Genotyping Page" illustrates a freely accessible, fast and easy to use internet-based service for strain comparisons, in which a user can compare a genotype produced for one of his isolates to the existing data.

Construction and content

The Tandem Repeats Database main page

Tandem repeats were identified from finished microbial genome sequences (as listed by the Genome OnLine Database [20]) using the tandem repeats finder (TRF) software [21,22] with the following options: alignment parameters, "2,3,5" (these parameters are the less stringent ones), minimum alignment score to report repeat, 50 (this score allows to detect short structures), maximum period size, 500 base-pairs. When the program reported redundant (overlapping) repeats, the redundancy was eliminated as described in [23], before import in the database. The database uses Microsoft Access 2000 and the querying process uses Active Server Pages (ASP, Microsoft) with Perlscripts or VBscripts. Perl was obtained from the ActiveState Programmer Network [24]. The database is hosted on a server running under Windows 2000 server (Microsoft). The tandem repeats database main page is described in more detail in [9].

The Strain Comparison page

Sequence comparisons used BLAST [25]. The BLAST software was obtained from the NCBI FTP site [26]. The flanking sequences of TRs from one strain were compared to the whole sequence of the other strain (and reciprocally, to avoid missing some tandem repeats that would not appear in the tandem repeats database for one strain because they were not detected by the Tandem Repeats Finder [21] -for instance because there is only one copy of the repeated unit in the considered strain). The resulting list of matching tandem repeats was then imported in the database, where it can be queried. The comparison of more than two strains was made possible through a supplemental step before import in the database: the synthesis of several 2-strains comparisons, of the same "reference" strain against each of the others (matching between TRs of the different strains was deduced from the positions on the reference strain).

The Blast page

The Blast Page allows users to run BLAST [25] in the tandem repeats and flanking sequences from the database via Perlscripts. The Blast outputs are linked to the database, in order to easily obtain the description of identified tandem repeats.

The Bacterial Genotyping page

The web-page site performing identifications was developed using the BNserver application (version 3.0, Applied-Maths, Belgium) and ASP (Microsoft) using Perlscript. The typing results (gel images and resulting data) were managed using the Bionumerics software package as described in [10]. The output of a query is a list of strains and genotypes from the database together with similarity scores.

Utility

The procedure to find polymorphic tandem repeats (TRs) for use in strain typing

Figure 1 shows the steps leading from a genome sequence to the exploitation of polymorphic tandem repeats for bacterial strain genotyping. Although Tandem Repeats are easily identified using the Tandem Repeats Database, TR polymorphism must be evaluated by typing across a set of relevant strains. If the sequences of several strains of the species of interest are available, the Strain Comparison Page can be used to directly identify tandem repeats predicted to be polymorphic in size between the two (or more) sequenced strains. However, it is important to keep in mind that the tandem repeats predicted as being polymorphic will depend on the sequenced strains and well-planned surveys of isolates will still be necessary. The

available tools do not replace this validation step, as the value of each marker must be carefully established on an appropriate set of isolates. The definition of an appropriate set of isolates depends upon the question which is being addressed, *i.e.* large scale or local epidemiology. The Blast Page has been implemented in the tandem repeats database in order to easily determine the size of the expected PCR amplification products. The database is also manually updated to contain PCR conditions as well as polymorphism index, and links to the original reports [27] (input from users is welcome). Eventually, when an MLVA assay has been fully developed and validated, typing data can be made accessible so that individual queries can be run. The Bacterial Genotyping Page illustrates how this could work. The genotyping data for a strain can be entered and submitted via this page. The output is the description of the closest strains. The data which has been submitted is not incorporated in the database itself, since this would require stringent data validation steps. In the following sections, we are presenting the web-based resources associated with this procedure.

The "Strain Comparison" pages

The strain comparison pages are available via [28]. The comparison of two strains is based on a pre-computed BLAST [25] analysis of the flanking sequences of tandem repeats from one strain against the other, and vice-versa. Figure 2 summarizes the results of this first step for 23 comparisons. Three indexes are scored (see figure legend): (1) the "mean %identity" between the flanking sequences is a measure of single nucleotide polymorphism (SNPs) frequency (not insertions-deletions), (2) the proportion (%) of flanking sequences that matched the flanking sequence of its homologue in the other strain on more than half of the 500 bp assayed here – *i.e.* that were not rearranged, by insertion of mobile elements for instance - , (3) the proportion (%) of tandem repeats that were found to be of a different length between the two strains being compared. In addition, the positions of matching tandem repeats in the two genomes is plotted to reveal large-scale genome rearrangements. A number of situations are observed: for instance *Yersinia pestis* orientalis strain CO-92 [29], and medievalis strain KIM5 P12 [30] show a very high "mean %identity" (99.96 %), in agreement with the recent emergence of *Yersinia pestis* [4]. In spite of this, the two strains differ by a high number of large rearrangements (as seen on the plot), which reflects the high genome plasticity observed in this species [31], together with a relatively high rate of polymorphic tandem repeats (8.47%). In contrast, *Listeria monocytogenes* strain EGD-e and *Listeria innocua* strain Clip 11262 have a lower homology (90.19%) and only 3.99% of polymorphic tandem repeats in spite of the evolutionary distance (see Figure 2).

The strain comparison page allows queries in the tandem repeats database according to the tandem repeat length difference between the two strains compared, and also to other tandem repeats characteristics (unit length, copy number, etc...). Figure 3 illustrates a query done for *Mycobacterium tuberculosis* strains H37Rv and CDC1551 [32]: the query "length difference ≥ 5 bp" identifies 58 tandem repeats (8 are shown on Figure 3). This prediction has been tested for the 30 loci amenable to PCR analysis and polymorphism has been confirmed in all cases [10].

When more than two strains have been sequenced, a synthesis of the results of several 2-strains comparisons is also available. Figure 4 illustrates a query made for *Escherichia coli* strains O157:H7 Sakaï, O157:H7 EDL933, K12, and UPEC-CFT073 [33-35]: 87 tandem repeats were found with 2 to 4 alleles among the 4 strains (18 of which are listed in Figure 4).

The "Blast in the Tandem Repeats Database" page

To facilitate the identification of already studied tandem repeats, we implemented BLAST [25] against the tandem repeats from the database, *i.e.* the tandem repeats themselves and their flanking sequences. The Blast page is available at [36]. All bacteria can be queried at once, which allows the identification of tandem repeats families, conserved in several bacterial species. Another page is dedicated to the Blast of PCR primers and provides the size of the PCR products in all the species/strains where the primers match. Figure 5 shows the results of searching the PCR primer pair from tandem repeat H37Rv_0024_18 bp [10] in all bacteria: as expected, the PCR primer pair matches *Mycobacterium tuberculosis* strains H37Rv and CDC1551, providing different PCR product lengths.

The Bacterial Genotyping page

The Bacterial Genotyping page [37] provides one illustration on how tandem repeat typing data can be made available via internet to allow external users to query genotyping data (*Bacillus anthracis*, *Yersinia pestis*, *Mycobacterium tuberculosis*, *Pseudomonas aeruginosa* for the moment) and compare a new strain to existing data as previously described in [10]. For each locus, allele sizes can be selected among a list of possibilities (observed sizes). The results of the query indicate a similarity score and include links to the complete data recorded for each strain listed. This page is just meant as an illustration and prototype. MLVA reference data could also be made available for downloading as tabular data files, or can be copied from published datasets, which can then be complemented by in-house data, and analyzed by the appropriate clustering software.

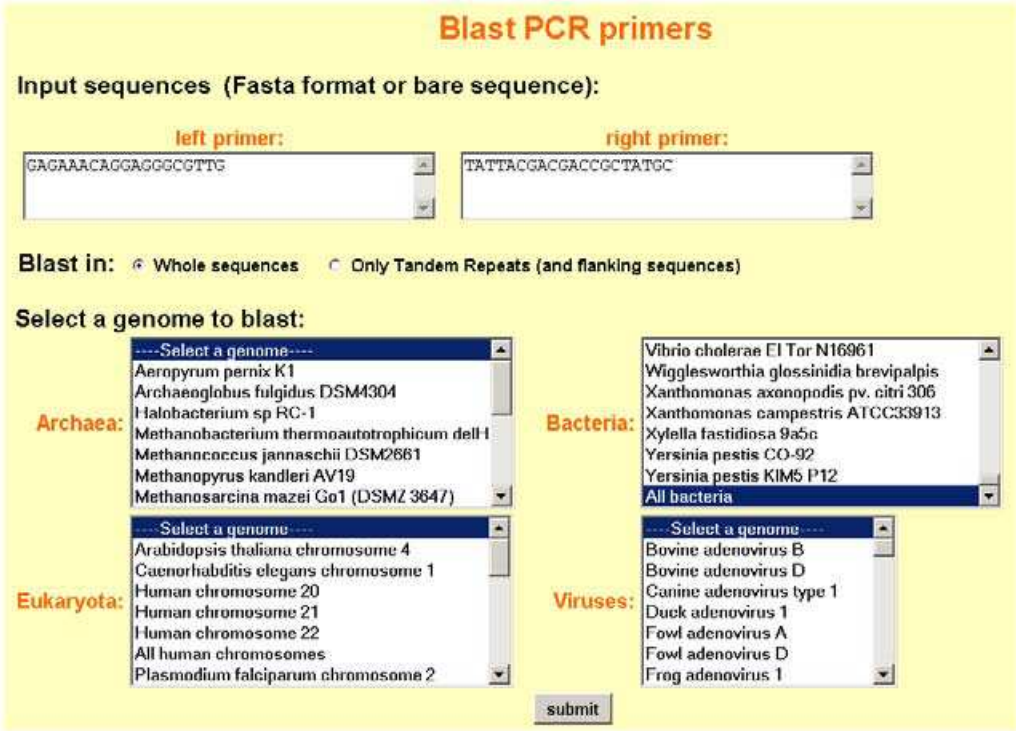


Figure 5
Example of a query in the "Blast of PCR primers" page, providing the length of the PCR products in the strains/species where the primer pair matches, and links to the corresponding tandem repeats descriptions.

Discussion

Bacterial genomes evolution

As shown by the indexes from Figure 2, there are different ways to represent the divergence/similarity between two strains. They are not correlated, suggesting independent evolution processes. First, the "mean %identity" between two genomes reflects point mutations, and is an indicator of the time passed since the two strains diverged. For instance, *Yersinia pestis* is known to be of recent emergence [4] and shows a high "mean %identity" between strains CO-92 (orientalis) and KIM5 P12 (medievalis). In contrast, and as shown by the dot plot, large genome rearrangements occurred in this genome, which is representative of a high genome plasticity [31]. The index "% of flanking sequences not rearranged" is an indicator of small-scale genome rearrangements, such as the inser-

tions of mobile elements. This index is low for genomes rich in mobile elements, like *Streptococcus agalactiae*, in which such elements significantly contribute to strain diversity [38]. Finally, the index "% of polymorphic tandem repeats" between two strains represents the tandem repeats evolution rate. For the moment, the mechanisms of bacterial VNTRs mutations have not been precisely investigated, but it seems likely to be independent of the other processes mentioned, as there are no correlations between the indexes. Figure 2 provides clues to assess which typing method(s) will be efficient in the different species. For instance, the two bacterial species *Salmonella typhimurium* strain LT2 [39] and *Shigella flexneri* strain 2a301 [40] share only 86.06% of sequence identity, clearly making the identification of matching tandem repeats between the two species difficult and of low significance. MLVA analysis appears to be of highest interest for the subspecies typing of highly monomorphic species including *Yersinia pestis*, *Bacillus anthracis*, *Mycobacterium tuberculosis* and *Brucella* [9,10,41].

Strain comparison efficiency

The sequencing of more than one strain for some bacterial species allows direct identification of polymorphic tandem repeats, assuming that no sequencing errors occurred. Earlier investigations provide good reasons to believe that tandem repeats in the size range considered here (a few hundred base-pairs) are correctly sequenced, and consequently, that the strain comparison data is reliable. As a negative control, the comparison of two independent sequences from the same strain of *Agrobacterium tumefaciens* strain (C58), one from Cereon genomics [42] and the other from Washington University [43], shows that no length polymorphism is detected among tandem repeats (Figure 2) between the two independent sequences. As a positive control, the tandem repeats predicted to be polymorphic by genome sequence comparison between the two strains of *M. tuberculosis* have indeed been proved polymorphic by PCR typing of isolates [10].

Selection based on comparison of sequence data from two strains will miss some polymorphic loci. Indeed, the results provided by the approach rely upon the phylogenetic distance between the two strains being compared. If the strains are very closely related, only a few TRs will be found different between them, but these tandem repeats will probably be the most polymorphic ones. Conversely, if the strains are distant in the phylogenetic tree, a larger number of polymorphic TRs will be found, some of them will be only moderately polymorphic. Obviously, when a few well-selected strains have been sequenced, it is likely that very few polymorphic tandem repeats are undetected in the Strain Comparison pages.

It is of course still going to be very important to determine the TR allele frequency for isolates carefully selected to be representative of the global diversity of a given pathogen before suggesting the configuration of an MLVA assay to use in subsequent studies. In addition, those TR markers that are highly polymorphic in diverse test panels of isolates may be monomorphic when applied to isolates responsible for local outbreaks. The configuration of TR markers used to make up an assay needs to be determined empirically with representative local isolates and tailored to the study population and study questions.

Polymorphic tandem repeats selection for species with only one sequenced strain

The identification of simple criteria able to predict tandem repeat polymorphism when genome sequence data is available for only one strain would indeed greatly facilitate the development of MLVA assays. It would seem reasonable for instance to expect that the number of copies and the internal homogeneity of tandem arrays are strong predictors [23]. We take advantage here of the many strain comparisons which are made available via the strain comparison pages to evaluate such criteria.

We have analyzed bacteria with at least three sequenced genomes (*Staphylococcus aureus*: 6 strains, *Escherichia coli*: 4 strains, *Streptococcus pyogenes*: 4 strains and *Salmonella typhi* and *typhimurium*: 3 strains). We assume that in such cases, only a few polymorphic tandem repeats are missed in the comparisons. We compared the distribution of tandem repeats sequence characteristics among the group of "polymorphic" loci (differing in at least two of the strains compared, excluding length differences between strains that resulted from microdeletions in the flanking sequences) and the others. Comparisons were performed for the following sequence characteristics: unit length, copy number, total length, %GC, GC bias (=|%G-%C|/(%G+%C)), %matches, and HistoryR (a score derived from tandem repeat history reconstruction algorithm [44] as described in [23]). None of the variables were normally distributed, as tested with Kolmogorov-Smirnov test, so a non-parametric Wilcoxon test was used to compare the distributions, which were judged significantly different at the .05 level of the statistic (2 tailed). Distributions were significantly different for all 4 species studied for %matches, total length and copy number. As shown on Figure 6, polymorphic TRs have a higher internal conservation and total length than monomorphic ones. Copy number, which is correlated with total length, is also higher among polymorphic TRs.

Selecting the longest and most conserved tandem repeats should thus improve polymorphic TRs identification. Table 1 illustrates the query "total length ≥ 80 bp and %matches ≥ 80%" applied to the four species used to find

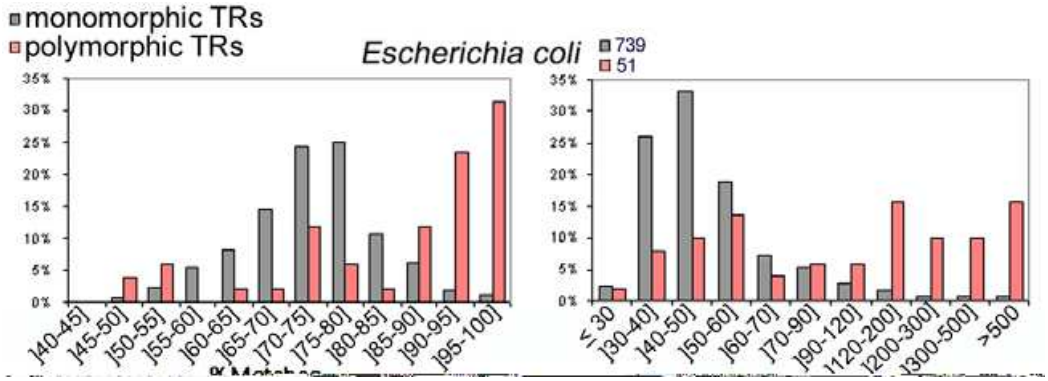


Figure 6
Proportion of predicted polymorphic (pink) and monomorphic (grey) tandem repeats according to different parameters (internal homogeneity of the repeat array (%matches) or total length). P-values obtained for the non-parametric Wilcoxon tests appear below each histogram.

Table 1: Use of the criterion "total length ≥ 80 bp and %matches ≥ 80%" on 4 species for which 3 strains or more were compared. The number of monomorphic, polymorphic (2 alleles or more) and highly polymorphic (3 alleles or more) TRs in whole set, and positive and negative groups are listed. (a) "criterion" refers to the selection of TRs with L ≥ 80 bp and %M ≥ 80%

Comparison (total number of TRs)	Whole set (proportion of total number)			Tandem repeats with L≥80 bp AND %M≥80% (proportion among the set)			Tandem repeats with L<80 bp OR %M<80% (proportion among the set)			% of the polymorphic TRs (2 alleles or more) that were detected by criterion[a]	% of the TRs with 3 alleles or more that were detected by criterion[a]	% of all TRs that fulfil the criterion[a]
	1 allele	2 alleles or more	3 alleles or more	1 allele	2 alleles	3 alleles or more	1 allele	2 alleles or more	3 alleles or more			
S aureus (833 TRs)	762 (91.5%)	71 (8.5%)	38 (4.5%)	5 (13%)	8 (20%)	25 (66%)	757 (95%)	25 (3.5%)	13 (1.5%)	46%	66%	7.23%
E coli (790 TRs)	739 (93.5%)	51 (6.5%)	12 (1.5 %)	12 (38%)	13 (40%)	7 (22%)	727 (96%)	26 (3.5%)	5 (0.5%)	39%	58%	4.86%
S typhi / typhimurium (641 TRs)	625 (97.5%)	16 (2.5%)	2 (0.3%)	13 (68%)	4 (22%)	2 (10%)	612 (98%)	10 (2%)	0 (0%)	37.5%	100%	3.27%
S pyogenes (292 TRs)	276 (94.5%)	16 (5.5%)	3 (1%)	4 (67%)	0 (0%)	2 (33%)	272 (95%)	14 (4.7%)	1 (0.3%)	12.5%	67%	2.71%

Table 2: Use of the criterion "total length ≥ 80 bp and %matches ≥ 80%" on 10 species for which 2 strains were compared. The numbers of tandem repeats with equal lengths and different lengths between the two strains in the whole set, and positive and negative groups are listed.

Comparison (total number of TRs loci)	Whole set (proportion)		Criterion + (L≥80 bp, %M≥80%)		Criterion -		Sensitivity (% of the TRs with different lengths that were detected by criterion)	Specificity (% of the TRs predicted by the criterion that have different length)	% of all TRs that fulfil the criterion
	equal length	different length	equal length	different length	equal length	different length			
H pylori 26695/J99 (624 TRs)	506 (81%)	118 (19%)	0	11	506	107	9%	100%	2%
N meningitidis MC58/Z2491 (642 TRs)	528 (82%)	114 (18%)	10	23	518	91	20%	70%	5%
M tuberculosis H37Rv/ CDC1551(1502 TRs)	1441 (96%)	61 (4%)	35	27	1406	34	44%	44%	4%
L monocytogenes EGD-e/*L innocua* Clip11262 (576 TRs)	553 (96%)	23 (4%)	2	3	551	20	13%	60%	1%
S agalactiae NEM316/2603 (398 TRs)	387 (97%)	11 (3%)	2	1	385	10	9%	33%	1%
S pneumoniae TIGR4/R6 (406 TRs)	339 (83%)	67 (17%)	14	29	325	38	43%	67%	10%
Y pestis CO-92/KIM5 P12 (1499 TRs)	1372 (92%)	127 (8%)	44	19	1328	108	15%	30%	4%
R prowazekii Madrid E/*R conorii* malish 7 (316 TRs)	290 (92%)	26 (8%)	0	2	290	24	8%	100%	1%
Brucella suis 1330/ *Brucella melitensis* 16 M (739 TRs)	681 (92%)	58 (8%)	2	4	679	54	7%	67%	1%
X fastidiosa 9a5c/grape Temecula1 (573 TRs)	440 (77%)	133 (23%)	2	28	438	105	21%	93%	5%

predictive criteria. For all four species, the group fulfilling the criterion is, as expected, enriched in polymorphic (at least two alleles) tandem repeats: in *Staphyloccocus aureus*, polymorphic tandem repeats represent only 8.5% of the whole population of tandem repeat loci but are predominant (87%) in the criterion positive group. The enrichment is even greater for highly polymorphic TRs, *i.e.* with 3 alleles or more: for example from 4.5% in the whole set to 66% in the positive group for *Staphylococcus aureus*. However this simple criterion misses more than half of the polymorphic loci. In addition, the efficiency of the criterion is highly variable in the different species: it is relatively satisfying in *Staphylococcus aureus* (54% of polymorphic tandem repeats would be missed) but very inefficient in *Streptococcus pyogenes* (almost 90% are missed). The results for highly polymorphic loci (3 alleles or more) are more consistent (the proportion of TRs with 3 alleles or more detected by the criterion ranges from 58% for *Escherichia coli* to 100% for *Salmonella*).

It is tempting to speculate that these observations are applicable to other species. Subsequently, we applied the criterion to ten of the 2-strains comparisons available on the Strain Comparison Page (Table 2). In all ten instances,

the criterion positive group is enriched in TRs with different lengths between the two strains, compared to the whole set. This proportion varies from less than 3% in *Streptococcus agalactiae* to more than 20% in *Xylella fastidiosa* in the whole set. It is increased to 33% and 93% respectively among the set of loci which satisfy the criterion (these percentages correspond to the predictor's specificity), but the vast majority of polymorphic loci will be missed (90% and 80% respectively). Sensitivity, that is % of the TRs with different lengths that were detected by criterion varies from 6.90% for *Brucella* to 44.26% for *Mycobacterium tuberculosis*.

The finding that polymorphic tandem repeats have, on average, a higher internal conservation, total length, and copy number than monomorphic ones is in agreement with previous observations that TR polymorphism is correlated with conservation in *Yersinia pestis* and with total length in *Bacillus anthracis* [9]. It is also reminiscent of the behavior of microsatellites (also called short sequence repeats: SSR, see [45] for review), which are stabilized by internal variations [46] and by reduction of the number of repeats [47]. Unfortunately, we show here that such simple prediction criteria may miss a very large proportion of polymorphic tandem repeats, and provide highly variable results in different species. This indicates that, in the absence of sequence data from two strains or more, the systematic testing of tandem repeats polymorphism across a set of relevant strains remains the most appropriate way to develop an MLVA assay. Consequently, the Strain Comparison page is of great use when two strains or more have been sequenced.

Conclusions

Bacterial strain typing at the subspecies level is essential for epidemiological issues in the context of disease control. This can be used to determine if an *S. aureus* or *P. aeruginosa* infection for instance has been acquired in an hospital environment or not. On a larger scale, it can be used to trace the emergence of new, more virulent or drug resistant *M. tuberculosis* strains. It is also of interest in the field of bioterrorism and bioweapons control, as was shown by the investigations following the 2001 *B. anthracis* attacks. Tandem repeats typing has recently emerged as one way to address this issue. Indeed, in the case of a number of highly monomorphic bacterial species, including *B. anthracis* and *Y. pestis*, tandem repeats typing is the method of choice for subspecies typing. In addition to the fact that these loci represent an important fraction of the existing polymorphism, it offers a number of practical advantages, including the ease of typing, and of data exchanges among different countries. It is hoped that the tools which are described here will help evaluate the potential of tandem repeats typing assays for a larger range of pathogens.

Availability

All the tools presented are freely available from http://minisatellites.u-psud.fr.

List of abbreviations used

ASP: active server pages

MLVA: multiple locus VNTR analysis

PCR: polymerase chain reaction

TR: tandem repeat

TRF: tandem repeats finder

Authors contributions

FD is the developer of the database and web site, and the curator of the database. GV participated in the development of the initial procedure for the tandem repeat size comparisons between two genomes. The two authors contributed equally to the writing.

Acknowledgments

This work was funded by grants from Délégation Générale de l'Armement (DGA, France) aimed at facilitating the typing of dangerous pathogens.

References

1. van Belkum A: **High-throughput epidemiologic typing in clinical microbiology.** *Clin Microbiol Infect* 2003, **9:**86-100.
2. Enright MC, Spratt BG: **Multilocus sequence typing.** *Trends Microbiol* 1999, **7:**482-7.
3. Gutacker MM, Smoot JC, Migliaccio CA, Ricklefs SM, Hua S, Cousins DV, Graviss EA, Shashkina E, Kreiswirth BN, Musser JM: **Genome-wide analysis of synonymous single nucleotide polymorphisms in *Mycobacterium tuberculosis* complex organisms: resolution of genetic relationships among closely related microbial strains.** *Genetics* 2002, **162:**1533-43.
4. Achtman M, Zurth K, Morelli G, Torrea G, Guiyoule A, Carniel E: ***Yersinia pestis*, the cause of plague, is a recently emerged clone of *Yersinia pseudotuberculosis*.** *Proc Natl Acad Sci U S A* 1999, **96:**14043-8.
5. van Belkum A, Scherer S, van Leeuwen W, Willemse D, van Alphen L, Verbrugh H: **Variable number of tandem repeats in clinical strains of *Haemophilus influenzae*.** *Infect Immun* 1997, **65:**5017-27.
6. Frothingham R, Meeker-O'Connell WA: **Genetic diversity in the *Mycobacterium tuberculosis* complex based on variable numbers of tandem DNA repeats.** *Microbiology* 1998, **144:**1189-1196.
7. Supply P, Mazars E, Lesjean S, Vincent V, Gicquel B, Locht C: **Variable human minisatellite-like regions in the *Mycobacterium tuberculosis* genome.** *Mol Microbiol* 2000, **36:**762-771.
8. Adair DM, Worsham PL, Hill KK, Klevytska AM, Jackson PJ, Friedlander AM, Keim P: **Diversity in a variable-number tandem repeat from *Yersinia pestis*.** *J Clin Microbiol* 2000, **38:**1516-9.
9. Le Flèche P, Hauck Y, Onteniente L, Prieur A, Denoeud F, Ramisse V, Sylvestre P, Benson G, Ramisse F, Vergnaud G: **A tandem repeats database for bacterial genomes: application to the genotyping of *Yersinia pestis* and *Bacillus anthracis*.** *BMC Microbiol* 2001, **1:**2.
10. Le Flèche P, Fabre M, Denoeud F, Koeck JL, Vergnaud G: **High resolution, on-line identification of strains from the *Mycobacterium tuberculosis* complex based on tandem repeat typing.** *BMC Microbiol* 2002, **2:**37.
11. Spanier JG, Jones SJ, Cleary P: **Small DNA deletions creating avirulence in *Streptococcus pyogenes*.** *Science* 1984, **225:**935-8.
12. Hollingshead SK, Fischetti VA, Scott JR: **Size variation in group A streptococcal M protein is generated by homologous recom-**

bination between intragenic repeats. *Mol Gen Genet* 1987, **207:**196-203.
13. Keim P, Kalif A, Schupp J, Hill K, Travis SE, Richmond K, Adair DM, Hugh-Jones M, Kuske CR, Jackson P: **Molecular evolution and diversity in *Bacillus anthracis* as detected by amplified fragment length polymorphism markers.** *J Bacteriol* 1997, **179:**818-24.
14. Keim P, Price LB, Klevytska AM, Smith KL, Schupp JM, Okinaka R, Jackson PJ, Hugh-Jones ME: **Multiple-Locus Variable-Number Tandem Repeat Analysis Reveals Genetic Relationships within *Bacillus anthracis*.** *J Bacteriol* 2000, **182:**2928-2936.
15. Sylvestre P, Couture-Tosi E, Mock M: **Polymorphism in the collagen-like region of the *Bacillus anthracis* BclA protein leads to variation in exosporium filament length.** *J Bacteriol* 2003, **185:**1555-63.
16. Supply P, Lesjean S, Savine E, Kremer K, van Soolingen D, Locht C: **Automated high-throughput genotyping for study of global epidemiology of *Mycobacterium tuberculosis* based on mycobacterial interspersed repetitive units.** *J Clin Microbiol* 2001, **39:**3563-3571.
17. Vergnaud G, Denoeud F: **Minisatellites: Mutability and Genome Architecture.** *Genome Res* 2000, **10:**899-907.
18. Pourcel C, Vidgop Y, Ramisse F, Vergnaud G, Tram C: **Characterization of a Tandem Repeat Polymorphism in *Legionella pneumophila* and Its Use for Genotyping.** *J Clin Microbiol* 2003, **41:**1819-1826.
19. Onteniente L, Brisse S, Tassios PT, Vergnaud G: **Evaluation of the polymorphisms associated with tandem repeats for *Pseudomonas aeruginosa* strain typing.** *J Clin Microbiol* 2003, **41:**4991-7.
20. **GOLD Genomes OnLine Database** [http://ergo.integratedgenomics.com/GOLD/]
21. Benson G: **Tandem repeats finder: a program to analyze DNA sequences.** *Nucleic Acids Res* 1999, **27:**573-580.
22. **The Tandem Repeats Finder** [http://tandem.bu.edu/trf/trf.html]
23. Denoeud F, Vergnaud G, Benson G: **Predicting Human Minisatellite Polymorphism.** *Genome Res* 2003, **13:**856-867.
24. **The ActiveState Programmer Network (ASPN) ActivePerl download page** [http://www.activestate.com/products/ActivePerl]
25. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25:**3389-402.
26. **The NCBI BLAST ftp site** [ftp://ftp.ncbi.nih.gov/blast/]
27. **The tandem repeats database** [http://minisatellites.u-psud.fr]
28. **The Strain Comparison Page** [http://minisatellites.u-psud.fr/comparison]
29. Parkhill J, Wren BW, Thomson NR, Titball RW, Holden MT, Prentice MB, Sebaihia M, James KD, Churcher C, Mungall KL, Baker S, Basham D, Bentley SD, Brooks K, Cerdeno-Tarraga AM, Chillingworth T, Cronin A, Davies RM, Davis P, Dougan G, Feltwell T, Hamlin N, Holroyd S, Jagels K, Karlyshev AV, Leather S, Moule S, Oyston PC, Quail M, Rutherford K, Simmonds M, Skelton J, Stevens K, Whitehead S, Barrell BG: **Genome sequence of *Yersinia pestis*, the causative agent of plague.** *Nature* 2001, **413:**523-7.
30. Deng W, Burland V, Plunkett G 3rd, Boutin A, Mayhew GF, Liss P, Perna NT, Rose DJ, Mau B, Zhou S, Schwartz DC, Fetherston JD, Lindler LE, Brubaker RR, Plano GV, Straley SC, McDonough KA, Nilles ML, Matson JS, Blattner FR, Perry RD: **Genome sequence of *Yersinia pestis* KIM.** *J Bacteriol* 2002, **184:**4601-11.
31. Radnedge L, Agron PG, Worsham PL, Andersen GL: **Genome plasticity in *Yersinia pestis*.** *Microbiology* 2002, **148:**1687-98.
32. Fleischmann RD, Alland D, Eisen JA, Carpenter L, White O, Peterson J, DeBoy R, Dodson R, Gwinn M, Haft D, Hickey E, Kolonay JF, Nelson WC, Umayam LA, Ermolaeva M, Salzberg SL, Delcher A, Utterback T, Weidman J, Khouri H, Gill J, Mikula A, Bishai W, Jacobs WR Jr, Venter JC, Fraser CM: **Whole-Genome Comparison of *Mycobacterium tuberculosis* Clinical and Laboratory Strains.** *J Bacteriol* 2002, **184:**5479-5490.
33. Hayashi T, Makino K, Ohnishi M, Kurokawa K, Ishii K, Yokoyama K, Han CG, Ohtsubo E, Nakayama K, Murata T, Tanaka M, Tobe T, Iida T, Takami H, Honda T, Sasakawa C, Ogasawara N, Yasunaga T, Kuhara S, Shiba T, Hattori M, Shinagawa H: **Complete genome sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K-12.** *DNA Res* 2001, **8:**11-22.
34. Perna NT, Plunkett G 3rd, Burland V, Mau B, Glasner JD, Rose DJ, Mayhew GF, Evans PS, Gregor J, Kirkpatrick HA, Posfai G, Hackett J, Klink S, Boutin A, Shao Y, Miller L, Grotbeck EJ, Davis NW, Lim A, Dimalanta ET, Potamousis KD, Apodaca J, Anantharaman TS, Lin J, Yen G, Schwartz DC, Welch RA, Blattner FR: **Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7.** *Nature* 2001, **409:**529-33.
35. Welch RA, Burland V, Plunkett G 3rd, Redford P, Roesch P, Rasko D, Buckles EL, Liou SR, Boutin A, Hackett J, Stroud D, Mayhew GF, Rose DJ, Zhou S, Schwartz DC, Perna NT, Mobley HL, Donnenberg MS, Blattner FR: **Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*.** *Proc Natl Acad Sci U S A* 2002, **99:**17020-4.
36. **The Blast in the tandem repeats database page** [http://minisatellites.u-psud.fr/Blast]
37. **The Bacterial Genotyping Page** [http://bacterial-genotyping.igmors.u-psud.fr]
38. Tettelin H, Masignani V, Cieslewicz MJ, Eisen JA, Peterson S, Wessels MR, Paulsen IT, Nelson KE, Margarit I, Read TD, Madoff LC, Wolf AM, Beanan MJ, Brinkac LM, Daugherty SC, DeBoy RT, Durkin AS, Kolonay JF, Madupu R, Lewis MR, Radune D, Fedorova NB, Scanlan D, Khouri H, Mulligan S, Carty HA, Cline RT, Van Aken SE, Gill J, Scarselli M, Mora M, Iacobini ET, Brettoni C, Galli G, Mariani M, Vegni F, Maione D, Rinaudo D, Rappuoli R, Telford JL, Kasper DL, Grandi G, Fraser CM: **Complete genome sequence and comparative genomic analysis of an emerging human pathogen, serotype V *Streptococcus agalactiae*.** *Proc Natl Acad Sci U S A* 2002, **99:**12391-6.
39. McClelland M, Sanderson KE, Spieth J, Clifton SW, Latreille P, Courtney L, Porwollik S, Ali J, Dante M, Du F, Hou S, Layman D, Leonard S, Nguyen C, Scott K, Holmes A, Grewal N, Mulvaney E, Ryan E, Sun H, Florea L, Miller W, Stoneking T, Nhan M, Waterston R, Wilson RK: **Complete genome sequence of *Salmonella enterica* serovar Typhimurium LT2.** *Nature* 2001, **413:**852-6.
40. Jin Q, Yuan Z, Xu J, Wang Y, Shen Y, Lu W, Wang J, Liu H, Yang J, Yang F, Zhang X, Zhang J, Yang G, Wu H, Qu D, Dong J, Sun L, Xue Y, Zhao A, Gao Y, Zhu J, Kan B, Ding K, Chen S, Cheng H, Yao Z, He B, Chen R, Ma D, Qiang B, Wen Y, Hou Y, Yu J: **Genome sequence of *Shigella flexneri* 2a: insights into pathogenicity through comparison with genomes of *Escherichia coli* K12 and O157.** *Nucleic Acids Res* 2002, **30:**4432-41.
41. Bricker BJ, Ewalt DR, Halling SM: **Brucella 'Hoof-Prints': strain typing by multi-locus analysis of variable number tandem repeats (VNTRs).** *BMC Microbiol* 2003, **3:**15.
42. Goodner B, Hinkle G, Gattung S, Miller N, Blanchard M, Qurollo B, Goldman BS, Cao Y, Askenazi M, Halling C, Mullin L, Houmiel K, Gordon J, Vaudin M, Iartchouk O, Epp A, Liu F, Wollam C, Allinger M, Doughty D, Scott C, Lappas C, Markelz B, Flanagan C, Crowell C, Gurson J, Lomo C, Sear C, Strub G, Cielo C, Slater S: **Genome sequence of the plant pathogen and biotechnology agent *Agrobacterium tumefaciens* C58.** *Science* 2001, **294:**2323-8.
43. Wood DW, Setubal JC, Kaul R, Monks DE, Kitajima JP, Okura VK, Zhou Y, Chen L, Wood GE, Almeida NF Jr, Woo L, Chen Y, Paulsen IT, Eisen JA, Karp PD, Bovee D Sr, Chapman P, Clendenning J, Deatherage G, Gillet W, Grant C, Kutyavin T, Levy R, Li MJ, McClelland E, Palmieri A, Raymond C, Rouse G, Saenphimmachak C, Wu Z, Romero P, Gordon D, Zhang S, Yoo H, Tao Y, Biddle P, Jung M, Krespan W, Perry M, Gordon-Kamm B, Liao L, Kim S, Hendrick C, Zhao ZY, Dolan M, Chumley F, Tingey SV, Tomb JF, Gordon MP, Olson MV, Nester EW: **The genome of the natural genetic engineer *Agrobacterium tumefaciens* C58.** *Science* 2001, **294:**2317-23.
44. Benson G, Dong L: **Reconstructing the duplication history of a tandem repeat.** *Proc Int Conf Intell Syst Mol Biol* 1999:44-53.
45. van Belkum A, Scherer S, van Alphen L, Verbrugh H: **Short-sequence DNA repeats in prokaryotic genomes.** *Microbiol Mol Biol Rev* 1998, **62:**275-93.
46. Schumacher S, Fuchs RP, Bichara M: **Two distinct models account for short and long deletions within sequence repeats in *Escherichia coli*.** *J Bacteriol* 1997, **179:**6512-7.
47. De Bolle X, Bayliss CD, Field D, van de Ven T, Saunders NJ, Hood DW, Moxon ER: **The length of a tetranucleotide repeat tract in *Haemophilus influenzae* determines the phase variation rate of a gene with homology to type III DNA methyltransferases.** *Mol Microbiol* 2000, **35:**211-22.

2.2 Utilisation de la base de données pour l'épidémiologie bactérienne

2.2.1 Application au génotypage de *Yersinia pestis* et *Bacillus anthracis*

L'article suivant (Le Flèche 2001) intitulé « A tandem repeats database for bacterial genomes : application to the genotyping of *Yersinia pestis* and *Bacillus anthracis* » (une base de données des répétitions en tandem des génomes bactériens : application au génotypage de *Yersinia pestis* et *Bacillus anthracis*), a constitué la première application de la base de données en épidémiologie moléculaire. Ma contribution a été la constitution de la première version de la base, avec à l'époque 35 génomes analysés. J'ai également effectué une comparaison de différentes espèces bactériennes entièrement séquencées, selon les caractéristiques de leurs répétitions en tandem (densité en répétitions en tandem, proportion de répétitions en tandem avec une unité multiple de 3…), qui révèle une grande hétérogénéité du règne bactérien. Enfin, j'ai constitué des fiches descriptives des marqueurs analysés. Le travail est orienté sur des bactéries considérées comme étant des menaces « bioterroristes ». Quelques mois plus tard d'ailleurs, aux Etats-Unis, des enveloppes contenant des spores de *Bacillus anthracis* ont été envoyées par courrier (Jernigan 2002). L'enquête qui a suivi a recherché l'origine des spores utilisées : les différentes enveloppes contenaient-elles la même souche, et quelle souche ? L'approche utilisée est identique à celle décrite ici, c'est-à-dire le typage de répétitions en tandem, et elle a fourni l'essentiel des réponses attendues en quelques jours. Le laboratoire qui a réalisé le travail aux Etats-Unis a comparé le génotype obtenu à un fichier préalablement constitué grâce à l'analyse de souches provenant du monde entier.

Résumé :

Contexte : Certaines espèces de bactéries pathogènes sont génétiquement très homogènes, ce qui rend difficile la distinction entre les souches. Ces dernières années, les répétitions en tandem ont été mises en avant comme des marqueurs de choix pour le génotypage d'un certain nombre de pathogènes. L'évolution rapide de ces structures semble contribuer à la flexibilité phénotypique des bactéries pathogènes. La disponibilité de séquences de génomes entiers a ouvert la voie à l'évaluation systématique de la diversité des répétitions en tandem et à leur utilisation pour des études épidémiologiques.

Résultats : Cet article présente une base de données (http://minisatellites.u-psud.fr) des répétitions en tandem de génomes bactériens, d'accès public, qui facilite l'identification et la sélection des répétitions en tandem. Nous illustrons son utilisation par la caractérisation de minisatellites de deux importants pathogènes humains, *Yersinia pestis* et *Bacillus anthracis*.

Afin d'éviter les locus de contingence, qui sont probablement de faible valeur en tant que marqueurs épidémiologiques, et de proposer des outils de génotypage exploitables par électrophorèse sur des gels d'agarose classiques, seules les répétitions en tandem d'unité répétée d'au moins 9 pb ont été évaluées. *Yersinia pestis* contient 64 minisatellites de ce type, dans lesquels l'unité est répétée au moins 7 fois. Un lot de 12 locus supplémentaires, contenant 6 unités répétées et ayant une forte conservation interne a également été testé. Parmi 5 souches de *Yersinia*, 49 locus sont polymorphes (25 parmi 3 souches de *Yersinia pestis*). *Bacillus anthracis* contient 30 structures comparables, dans lesquelles l'unité est répétée au moins 10 fois. La moitié de ces répétitions en tandem est polymorphe parmi les souches testées.

Conclusions: L'analyse des séquences des génomes bactériens actuellement disponibles montre que *Bacillus anthracis* et *Yersinia pestis* ont une densité intermédiaire en répétitions en tandem de plus de 100 pb (environ 30 par Mb) par rapport aux autres génomes bactériens analysés jusqu'à présent. Dans les deux cas, tester le polymorphisme d'une fraction seulement de ces séquences a suffi pour développer rapidement un lot de plus de 15 marqueurs informatifs, certains montrant un très fort degré de polymorphisme. Par exemple, pour le marqueur BAMS7 de *Bacillus anthracis*, l'indice de polymorphisme (PIC) atteint 0.82, avec des allèles couvrant une large plage de tailles (600 à 1950 pb), et 9 allèles sont distingués parmi le nombre restreint de souches typées dans cette étude.

Research article

A tandem repeats database for bacterial genomes: application to the genotyping of *Yersinia pestis* and *Bacillus anthracis*

Philippe Le Flèche[1,2], Yolande Hauck[2], Lucie Onteniente[2], Agnès Prieur[1,2], France Denoeud[2], Vincent Ramisse[1], Patricia Sylvestre[1], Gary Benson[3], Françoise Ramisse[1] and Gilles Vergnaud*[1,2]

Address: [1]Centre d'Etudes du Bouchet, BP3, 91710 Vert le Petit, France, [2]Génomes et Minisatellites, Institut de Génétique et Microbiologie, Bat 400, Université Paris XI, 91405 Orsay cedex, France and [3]Department of Biomathematical Sciences, Box 1023, Mount Sinai School of Medicine, One Gustave L. Levy Place, New York, USA

E-mail: Philippe Le Flèche - lefleche@igmors.u-psud.fr; Yolande Hauck - Yolande.Hauck@igmors.u-psud.fr; Lucie Onteniente - Lucie.Onteniente@igmors.u-psud.fr; Agnès Prieur - Agnes.Prieur@igmors.u-psud.fr; France Denoeud - France.Denoeud@igmors.u-psud.fr; Vincent Ramisse - Vincent.Ramisse@ceb.etca.fr; Patricia Sylvestre - psylvest@pasteur.fr; Gary Benson - benson@ecology.biomath.mssm.edu; Françoise Ramisse - f.ramisse@freesurf.fr; Gilles Vergnaud* - Gilles.Vergnaud@igmors.u-psud.fr
*Corresponding author

Published: 30 March 2001

BMC Microbiology 2001, **1**:2

Received: 19 February 2001
Accepted: 30 March 2001

This article is available from: http://www.biomedcentral.com/1471-2180/1/2

Abstract

Background: Some pathogenic bacteria are genetically very homogeneous, making strain discrimination difficult. In the last few years, tandem repeats have been increasingly recognized as markers of choice for genotyping a number of pathogens. The rapid evolution of these structures appears to contribute to the phenotypic flexibility of pathogens. The availability of whole-genome sequences has opened the way to the systematic evaluation of tandem repeats diversity and application to epidemiological studies.

Results: This report presents a database ([http://minisatellites.u-psud.fr]) of tandem repeats from publicly available bacterial genomes which facilitates the identification and selection of tandem repeats. We illustrate the use of this database by the characterization of minisatellites from two important human pathogens, *Yersinia pestis* and *Bacillus anthracis*. In order to avoid simple sequence contingency loci which may be of limited value as epidemiological markers, and to provide genotyping tools amenable to ordinary agarose gel electrophoresis, only tandem repeats with repeat units at least 9 bp long were evaluated. *Yersinia pestis* contains 64 such minisatellites in which the unit is repeated at least 7 times. An additional collection of 12 loci with at least 6 units, and a high internal conservation were also evaluated. Forty-nine are polymorphic among five *Yersinia* strains (twenty-five among three *Y. pestis* strains). *Bacillus anthracis* contains 30 comparable structures in which the unit is repeated at least 10 times. Half of these tandem repeats show polymorphism among the strains tested.

Conclusions: Analysis of the currently available bacterial genome sequences classifies *Bacillus anthracis* and *Yersinia pestis* as having an average (approximately 30 per Mb) density of tandem repeat arrays longer than 100 bp when compared to the other bacterial genomes analysed to date. In both cases, testing a fraction of these sequences for polymorphism was sufficient to quickly develop a set of more than fifteen informative markers, some of which show a very high degree of polymorphism. In one instance, the polymorphism information content index reaches 0.82 with allele length covering a wide size range (600-1950 bp), and nine alleles resolved in the small number of independent *Bacillus anthracis* strains typed here.

Background

The polymorphism associated with tandem repeats has been instrumental in mammalian genetics for the construction of genetic maps and still is the basis of DNA fingerprinting in forensic applications. Tandem repeats are usually classified among satellites (spanning megabases of DNA, associated with heterochromatin), minisatellites (repeat units in the range 6-100 bp, spanning hundreds of base-pairs) and microsatellites (repeat units in the range 1-5 bp, spanning a few tens of nucleotides).

More recently, a number of studies have supported the notion that tandem repeats reminiscent of mini and microsatellites are likely to be a highly significant source of very informative markers for the identification of pathogenic bacteria even when these pathogens are recently emerged, highly monomorphic species [1–5]. This probably reflects the important contribution of tandem repeats to the adaptation of the pathogen to its host. Tandem repeats appear to contribute to phenotypic variation in bacteria in at least two ways. Tandem repeats located within the regulatory region of a gene can constitute an on/off switch of gene expression at the transcriptional level [6,7]. Similarly, tandem repeats within coding regions with repeat units length not a multiple of three can induce a reversible premature end of translation when a mutation changes the number of repeats (reviewed in [8–10]). In other instances, the repeated unit length is a multiple of three, and the tandem repeat contributes to a coding region. In such cases, variations in the number of copies modify the gene product itself [11].

Mutation mechanisms of micro and minisatellites have been studied in some detail in eukaryotes, essentially human and yeast (reviewed in [12]). In brief, the data obtained so far suggest that microsatellites mutate by replication slippage processes; mutation rates depend upon the efficiency of mismatch repair mechanisms and an internal heterogeneity within the array strongly stabilizes the tandem repeat. In contrast, minisatellites mutate predominantly as the result of the repair of a double strand break initiated within, or very close to, the tandem repeat. In eukaryotes at least, these events can be of replicative origin [13], or can be genetically controlled, and specifically induced, during meiosis, at double strand breaks hot-spots. Minisatellite mutation rate in eukaryotes appears to be insensitive to mismatch repair efficiency, and internal heterogeneity is compatible with a high mutation rate [12, 14].

In bacteria, loci containing a tandem repeat from the microsatellite class (repeat unit sizes of 1-8 bp) have been called simple sequence contingency loci [8]. Altered number of repeats allows for reversible on and off states of expression for the corresponding gene. The mutation rate of a tetranucleotide (microsatellite) tract in *Haemophilus influenzae* is higher than 10^{-4} and contributes to the adaptation of the pathogen to its hosts as the infection progresses [15]. In such an extreme situation, the microsatellite is of limited value for strain identification, epidemiological and phylogenetic studies. The tandem repeat array is composed of perfect copies of the elementary unit, and different alleles are observed in a single culture. In contrast, the phylogenetic identity of minisatellite alleles of identical size can usually be further checked by DNA sequencing, since the repeated units are often not perfect [16]. The pattern of variants along the array provides an additional level of allele identification and phylogenetic information. In addition, tandem repeats with longer repeat unit length can be relatively easily typed in the size range of a few hundred base-pairs using ordinary horizontal gel electrophoresis.

In this report, we will first describe the use of a tandem repeats database for bacterial genomes ([http://minisatellites.u-psud.fr]) and briefly compare the general characteristics of tandem repeats in a number of bacterial genomes for which the sequence has been determined and made publicly available. We will then show how this tool can easily be applied to the rapid characterization of new highly polymorphic markers in two pathogens, *Y. pestis* and *B. anthracis*.

Both *Y. pestis* (causative agent of plague) and *B. anthracis* (causative agent of anthrax) are recently emerged clones of respectively Y. pseudotuberculosis [17] and *B. cereus* [18]. In the case of *Y. pestis*, a high resolution typing tool based on RFLP (Restriction Fragment Length Polymorphism) analysis of IS100 locations has already been developed [17]. However this technology is more demanding than PCR typing, which justifies the development of such an assay. In the case of *B. anthracis*, polymorphisms were initially identified essentially using AFLP (Amplified Fragment Length Polymorphism) typing [19]. Subsequent analyses demonstrated that the most informative fragments in AFLP patterns resulted from tandem repeat array length variations (five minisatellite loci were characterized in this way [2]).

Results and discussion

Use of the tandem repeats database

To date, 36 bacterial genome sequences from 32 species have been released in the public domain and are included in the database (Figure 1A; the nine archaebacteria genomes sequenced to date are presented in an other page, which can be accessed from [http://minisatellites.u-psud.fr/]). As many other sequencing projects are under way ([http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/bact.html] ; [http://www.tigr.org/tdb/mdb/

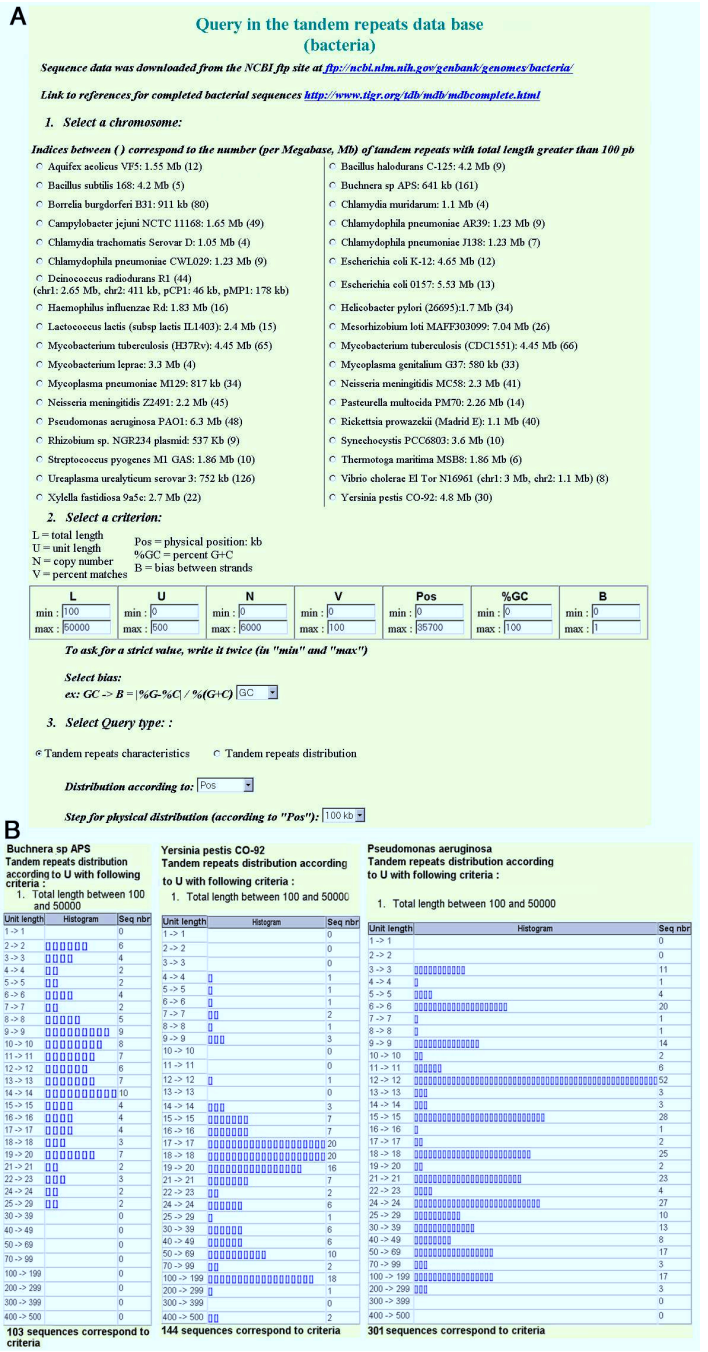


Figure 1

Querying the tandem repeats database 1A: bacterial tandem repeats main page Bacteria species are listed in alphabetical order. The name of the strain used for sequencing is indicated after the species name and before the genome size (expressed in megabase). The rightmost figure indicates the density (per Mb) of tandem repeat arrays longer than 100 bp. The search for tandem repeats can be restricted according to a combination of criteria, including total array length (L), repeat unit length (U), number of repeats (N), internal conservation of the repeats (V), position (expressed in kilobase) on the genome (Pos), GC content of the array (%GC), strand bias (B). Three different biases can be evaluated, GC bias, AT bias and Purine-Pyrimidine bias. The bias reflects strand asymmetry of the repeat sequence. The search output can either present a list of characteristics of the tandem repeats fulfilling criteria, ordered according to their position on the genome, or classify the tandem repeats according to a selected structural parameter. 1B: examples of queries in three genomes All tandem repeat arrays spanning more than 100 base-pairs are classified according to repeat unit length. The query was run on *Buchnera sp.* (left panel), *Yersinia pestis* (middle panel) and *Pseudomonas aeruginosa* (right panel).

mdbinprogress.html] ; [http://www.sanger.ac.uk/Projects/Microbes/]), the database will be regularly updated. The collection of tandem repeats present in a given genome can be queried according to a combination of criteria, total tandem repeat array length (L), repeat unit length (U), number of repeats (N), percentage of conservation of the repeats along the array (V), position on the genome (Pos), average GC percent of the repeats (%GC), strand bias in nucleotide composition (B) (these values have been precomputed using the Tandem Repeats Finder software described in [20]). The results shown on Figure 1B use the "Tandem Repeats Distribution according to repeat unit length" option (Figure 1A). Three genomes were searched for tandem repeat arrays longer than 100 base-pairs (L ≥ 100). The genomes selected illustrate three different behaviors. On the right panel, *Pseudomonas aeruginosa* shows a very striking bias towards minisatellites with a motif length multiple of three. On the left and middle panels of Figure 1B, B*uchnera sp* and Y. *pestis*, show no such bias. The overall density of tandem repeat arrays longer than 100 base-pairs varies in the different genomes. *Buchnera sp.* contains 103 such loci, for a total genome size of 641 kb, which corresponds to a density per megabase of 161. *Pseudomonas aeruginosa*, with a total genome length of 6.3 Mb, has a density of 48. *Y. pestis* has an intermediate value of 30. Figure 2 summarizes the values observed in the 32 species. Ten non pathogenic species are presented in the upper part, 22 pathogenic species on the lower part. The species are ordered from top to bottom according to increasing genome size. The dark bars indicate for each genome the density per megabase of tandem repeat arrays longer than 100 bp. The clear bars reflect the excess of tandem repeats with unit length a multiple of three. A wide range of situations is observed, with a remarkable excess of tandem repeats multiples of three in *Mycobacterium tuberculosis* and *Pseudomonas aeruginosa*, presumably reflecting a significant contribution of tandem repeats to coding regions in these two bacteria.

As a quick illustration of the use of this database to facilitate the development of genotyping tools for bacterial genomes, we have evaluated the polymorphism associated with tandem repeats from *Y. pestis* on one hand and *B. anthracis* on the other (in this second instance, the genome sequence has not been completed yet and does not appear on the publicly accessible Tandem Repeats Database page, Figure 1A).

Application to Y. pestis

Figure 3A presents the result of a query run on *Y. pestis*, to identify tandem repeats with repeat units longer than 9 base-pairs repeated at least 7 times in the strain which has been sequenced (CO-92 biovar Orientalis).Sixty-four tandem repeats fulfill these criteria (an additional group

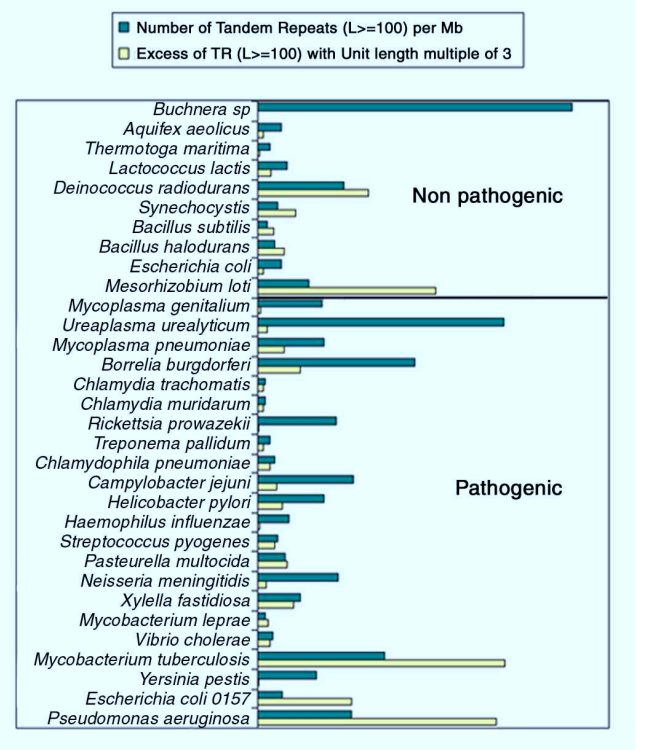


Figure 2
Relative frequency of tandem repeats within bacterial genomes The ten non-pathogen species are listed on top. Within each category, species are ordered according to genome size (smallest genome on top). The density of tandem repeat arrays longer than 100 bp is plotted for each species (dark bars). The clear bars reflect the excess (χ^2 values) of tandem repeats with a repeat unit length multiple of three.

of forty-nine have 6 copies of the motif; the twelve loci with the highest internal conservation were also included in this study). The output includes links to individual alignment files, as produced by the Tandem Repeat Finder software [20]. The alignment file also includes 200 base-pairs of flanking sequence from each side of the tandem repeat, from which primers can be selected for PCR amplification. Figure 3B shows an annotated extract of one alignment file. The positions of the primers selected for subsequent PCR amplification are underlined. Three *Y. pestis* (representing the Antiqua, Medievalis, and Orientalis biovars [17]) and two *Y. pseudotuberculosis* strains were used for the initial identification of minisatellites sufficiently polymorphic to be of interest for further studies. Table 1 summarizes the PCR conditions used for each polymorphic locus and the results obtained. A total of 76 tandem repeats were tested. PCR amplification failed in 6 cases. Twenty one loci are monomorphic in the five *Yersinia* strains typed here. Forty-nine of the loci are polymorphic (Table 1). Twenty-five of these are polymorphic among the *Y. pestis* strains.

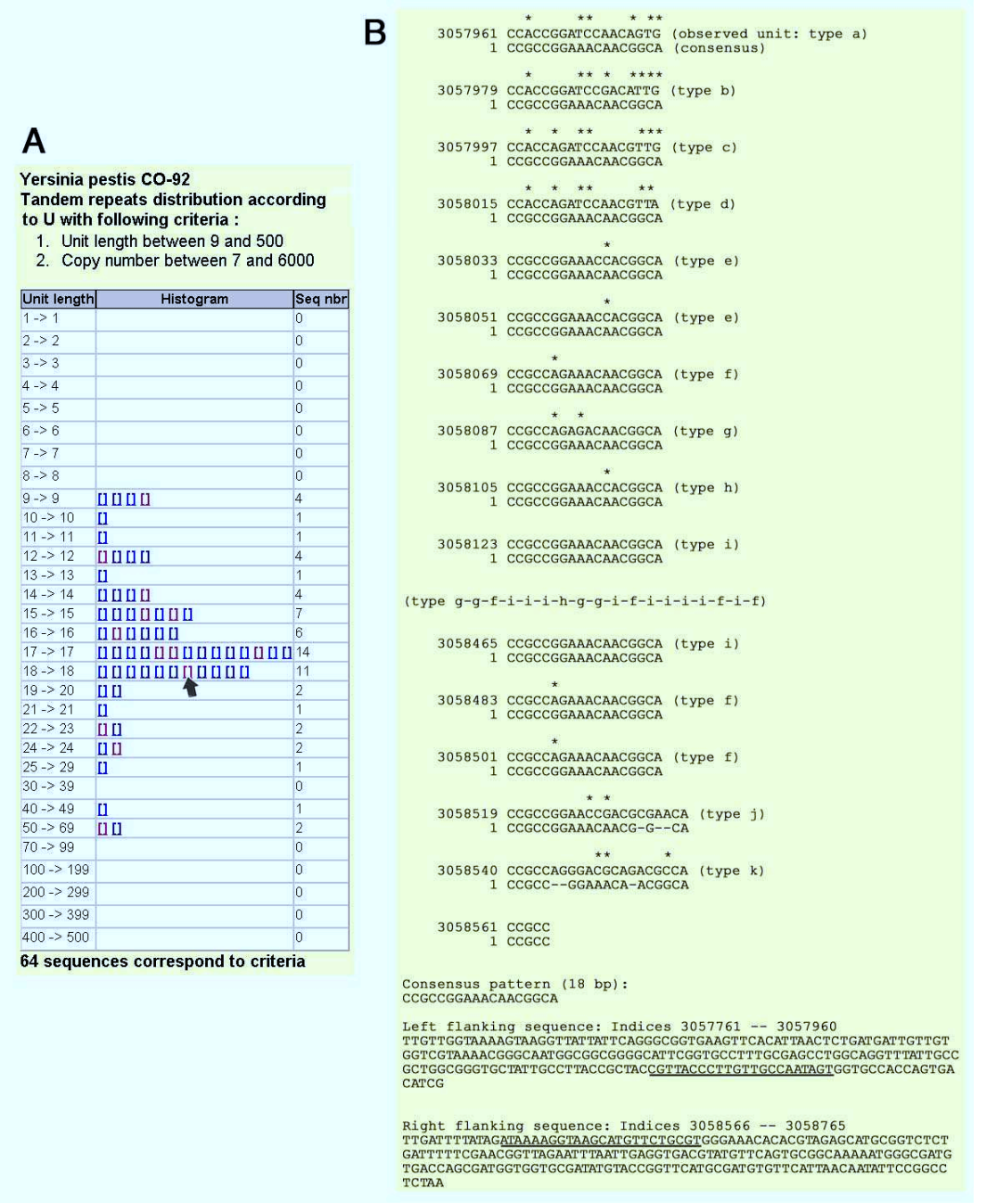


Figure 3
Selection procedure of minisatellites for *Y. pestis* 3A: Sixty-four tandem repeats have at least 7 units longer than 9 base-pairs. Panel A presents the distribution of these 64 loci according to repeat unit length. Each rectangle is an hyperlink to an alignment file. The rectangle indicated by the arrow is linked to the file illustrated in panel B. 3B: This is an annotated alignment file. The file corresponds to Yp3057ms09 (Table 1 and Figure 4; Yp : *Yersinia pestis*; 3057 : position on the genome, expressed in kilobases; MS09 : MiniSatellite index). The consensus pattern of 18 base-pairs is aligned to each motif. Annotations of the file are inserted within brackets. Although this minisatellite is very polymorphic, eleven different motifs (labeled a-k) are observed in the sequenced allele. The first four and last two copies are most diverged and rare. Four types of motifs (f, g, h, i) constitute most of the array. For convenience, 18 motifs have been removed from the alignment file and replaced by their letter code. The last two copies are 21 base-pair long instead of 18. The end of the alignment file (panel B, bottom) provides sequence data flanking the tandem repeat array. The positions of the primers chosen for PCR amplification of this locus (Table 1) are shown underlined.

Seven present a different allele in each of the five *Yersinia* strains, thirteen have a different allele in each of the three *Y. pestis* strains. Gel images for the 25 loci polymorphic among *Y. pestis* are shown in Figure 4. As can be seen, the repeat unit size and the overall length of the PCR products are such that tandem repeats differing by a single repeat unit can be distinguished by simple agarose gel electrophoresis.

Application to B. anthracis

Given the relatively low overall size of most bacterial tandem repeats, tandem repeat search can be run even on unfinished sequences. Tandem Repeats Finder was applied to *B. anthracis* sequence obtained from The Institute for Genomic Research through the website at [http://www.tigr.org] . The sequence was recovered as approximately 1000 contigs, for a total amount of slightly more than 5 Mb. Thirty tandem repeats have at least 10 copies of a repeat unit longer than 9 base-pairs. Fourteen of them are polymorphic among the 31 *B. anthracis* strains typed here (Table 2). Twenty-seven different genotypes are identified. Polymorphism information content (PIC) indexes based on the 27 genotypes vary from 0.07 to 0.82. Nine PIC values are above 0.5. Eight alleles are identified for CEB-Bams30, in a size range 270-900 base-pairs (Figure 5). In this case, the resolution of the largest alleles would probably be improved by using an automated DNA sequencer, and more alleles might be resolved. There are clear gaps in the size range coverage shown in Figure 5, and it is likely that the typing of additional strains would uncover new alleles. The genotyping data obtained was used to construct a phylogenetic tree based upon the Neighbor-Joining method ([http://www.infobiogen.fr]). In order to be able to correlate the tree obtained here with earlier studies [2], 5 minisatellites and one microsatellite reported previously were also typed. Figure 6 presents the data obtained and the resulting tree, using the nomenclature previously proposed [2]. Six *Bacillus cereus* strains have also been included and used as an outgroup in the analysis. Occasionally *B. cereus* strains will not amplify (scored as 0 in Figure 6) or will give weak amplification signals (Figure 5, last six lanes on the right). The proposed tree is in good agreement with earlier results. In particular, the A and B clusters are well defined. We have apparently no representatives for the A1b and A3a group, whereas strains 9533 and 9502 to 9505 appear to define a new branch. The correspondence between allele numbering and allele size is indicated in Table 3.

Correlations between polymorphism and structural characteristics of minisatellites

We have looked for correlations between on one hand the number of alleles and polymorphism of the minisatellites, and on the other, simple structural characteristics of the tandem repeats in the sequenced strain : motif size, number of motifs, total length, conservation of the motifs along the array (percent identity), GC content, strand bias. In the case of *B. anthracis*, a highly significant correlation (0.01 level) is observed between polymorphism and both total length and GC content. This is not true for *Y. pestis* in which a strong correlation is seen between the number of alleles and the conservation of the motifs (Figure 7).

Conclusions

We limited here our investigation of tandem repeats to minisatellites, i.e. repeat units longer than 9 base-pairs, so as to avoid simple sequence contingency loci [8] of limited epidemiological value, and to facilitate the typing of alleles with agarose gel electrophoresis. However, simple sequence contingency loci are also represented in the database and are of great interest for molecular pathogenicity studies [6–8]. The use of the tandem repeats database was demonstrated here on two of the most genetically homogeneous human pathogenes, *Y. pestis* and *B. anthracis*. There is consequently a possibility that a common database format for identification and epidemiological analyses of pathogens amenable to minisatellite typing be developed. As more data becomes available on polymorphism associated with tandem repeats, it will be added to the database presented here in order to avoid duplication of work and nomenclature.

Bacterial species differ very significantly in the density of tandem repeats within their genome, and also in their use of tandem repeats. Some species have a very strong excess of tandem repeats with repeat units length which are multiple of three, the most striking examples being *M. tuberculosis* and *P. aeruginosa*. Polymorphism in such tandem repeats is likely to modulate the protein structure rather than gene activity. In *M. tuberculosis*, all tandem repeats with total length (L) higher than 100 bp and 9 or 15 base-pairs long units are located with ORFs [21]. An important proportion of these tandem repeats correspond to the so-called PE and PPE multigene families [21].

In the two species studied here, tandem repeat polymorphism is strongly correlated with one or more of the sequenced allele characteristics, as illustrated in Figure 7. In *Yersinia pestis* a strong correlation is observed between number of alleles observed and homogeneity of the tandem array. In *Bacillus anthracis*, the strongest correlations are with total array length and GC content. It appears that the correlations are not the same in the two species, so that at present at least, the polymorphism associated with a tandem repeat cannot be inferred from its primary sequence. In particular, and in contrast to what is known for microsatellites (1-5 bp repeat units),

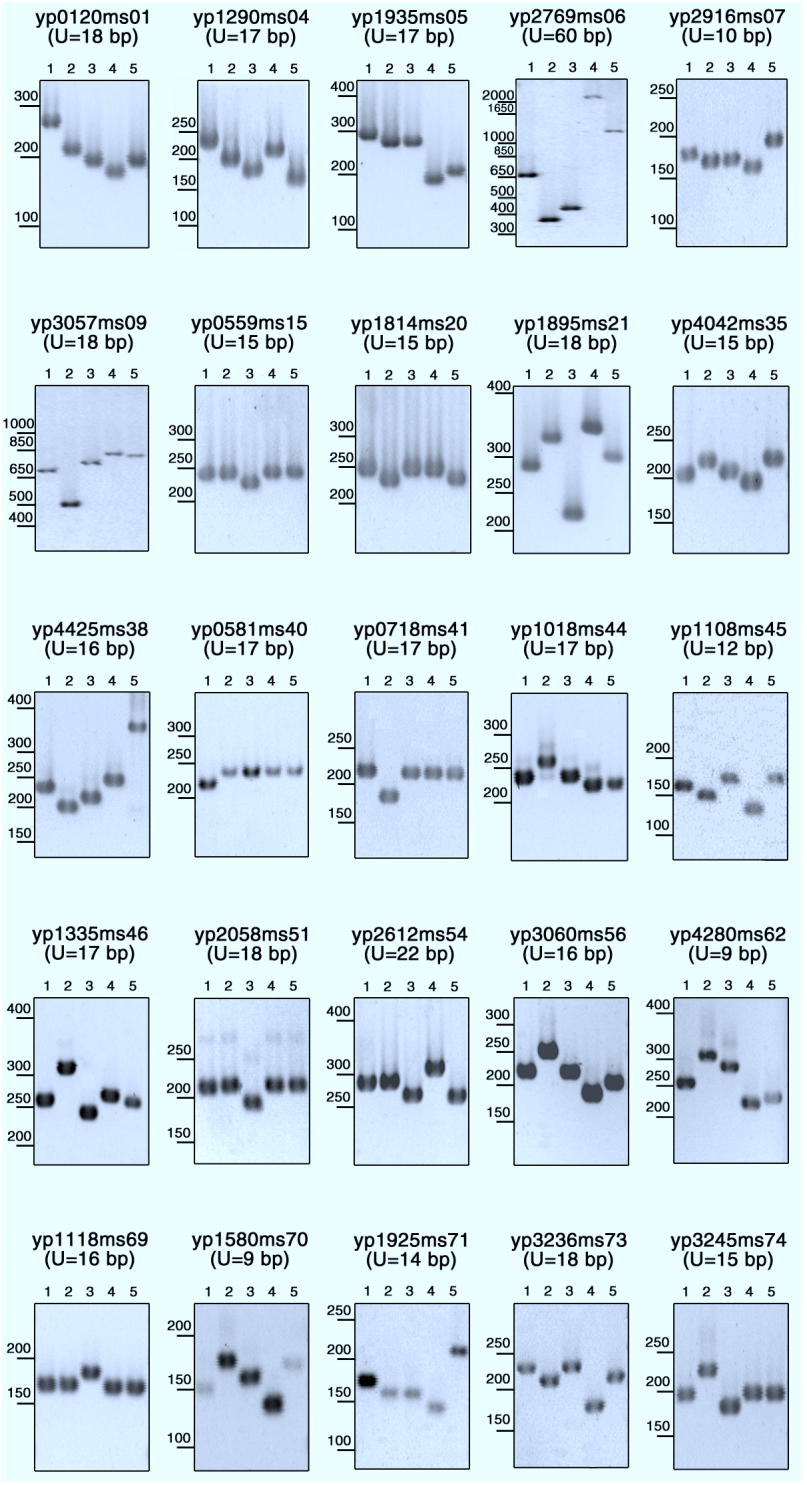


Figure 4

Images of PCR amplification of the twenty-five minisatellites polymorphic in the *Y. pestis* strains DNA from three reference *Y. pestis* strains representing each of the main biovars, antiqua (lane 1), medievalis (lane 2) and orientalis (lane 3) and two *Y. pseudotuberculosis* strains (lanes 4 and 5) have been PCR amplified and an aliquot of the products has been run on 2% horizontal agarose gels as described. The length of the minisatellite motifs (U) and the size range is indicated on each panel. Yp2916ms07 has one of the shortest (10 bp) unit. Four alleles are clearly distinguished between the 150 and 200 bp marker fragments.

some of the minisatellites are highly polymorphic in spite of a poor internal homogeneity of the sequenced allele, as is also the case for minisatellites in the human genome [12]. However, more systematic allele sequencing will be required to demonstrate that polymorphism is not associated with a subclass of alleles showing a higher internal homogeneity. Similarly, allele sequencing will be required to formally establish that the allele size variations observed are indeed (as is likely) the consequence of variations in the number of repeats.

Five among the *B. anthracis* markers described here (Ceb-Bams1, 3, 7, 13 and 30) are highly polymorphic with PIC values (or Nei's index) above 0.7. In this respect, it is important to observe that the length of the allele observed for Ceb-Bams1 in the Ames strain is not of the size expected from the sequence data (Table 2). This may result either from a high mutation rate at Ceb-Bams1 or from a sequencing error. The expected allele size corresponds to allele 4 (Table 3), which is unlikely for the Ames strain because Ceb-Bams1 allele 4 is observed only in cluster B strains (Figure 6) and Ames is well apart of cluster B [2]. A similar situation is observed for Ceb-Bams28, for which the expected product does not correspond to any existing allele in the collection of strains typed. In this case however, the locus is moderately polymorphic, with a PIC value of 0.26 and only three alleles observed (Table 2), so that a sequencing error is the most likely interpretation. This issue could be easily solved by typing with Ceb-Bams1 and Ceb-Bams28 the very strain which has been used for the sequencing project.

It is interesting to observe that, although the magnitude of allele size difference has not been taken into account when building the distance matrix, the resulting phylogenetic tree proposed in Figure 6 tends to group together strains with alleles of similar size at these most variable loci. This is reminiscent of observations made in *H. influenzae* [1] and suggest that mutation events are predominantly small size changes. Here again, more detailed studies involving full allele sequencing should now help understand the succession of events producing a population of alleles.

Materials and methods

Bacterial genomes DNA sequences

Finished sequences in the public domain were recovered by ftp from the NCBI or the Sanger center sites ([http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/bact.html] ; [http://www.sanger.ac.uk/Projects/Microbes/]). Preliminary sequence data for *B. anthracis* was obtained from The Institute for Genomic Research through the website at [http://www.tigr.org] .

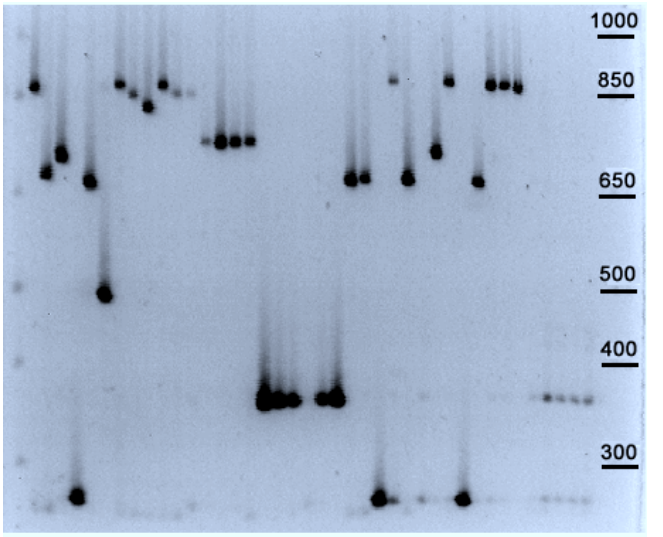


Figure 5
PCR amplification of *B. anthracis* minisatellite CEB-Bams30 DNA from *B. anthracis* and *B. cereus* (six rightmost lanes) was amplified using primers for CEB-Bams30 (Table 2). The PCR products were run on a 40 cm long 2% ordinary agarose gel.

DNA preparation

All strains used here are part of the collection maintained by the Centre d'Etudes du Bouchet (CEB). They originate either from the CIP (Collection Institut Pasteur, [http://www.pasteur.fr/]) or from AFSSA (Agence Française de Sécurité Sanitaire des Aliments, [http://www.afssa.fr/] , Dr Josée Vaissaire). DNA from each isolate was obtained by large-batch procedures or by the simplified procedure as described in [2]. In addition, 15 µg of DNA from the *B. anthracis* Ames strain were kindly provided by Dr Mats Forsman, FOA, Sweden.

Minisatellite PCR amplification and genotyping

PCR reactions were performed in 15 µl containing 1 ng of DNA, 1x Long Range Reaction Buffer 3 (Roche-Boehringer), 1 unit of Taq DNA polymerase, 200 µM of each dNTP, 0.3 µM of each flanking primer. The Taq DNA polymerase was either prepared essentially as described in [22] or purchased from Qbiogen or Roche-Boehringer. The 1x LongRange Buffer 3 is 1.75 mM $MgCl_2$, 50 mM Tris-HCl pH9.2, 16 mM $(NH_4)2SO_4$.

PCR reactions were run on a Perkin-Elmer 9600 or a MJResearch PTC200 thermocycler. An initial denaturation at 96°C for five minutes was followed by 34 cycles of denaturation at 96°C for 20 seconds, annealing at 60°C for 30 seconds, elongation at 65°C for 1 minute, followed by a final extension step of 5 minutes at 65°C. In few cases, other annealing temperatures and/or elongation times were used (see tables 1 and 2). Five microliters of

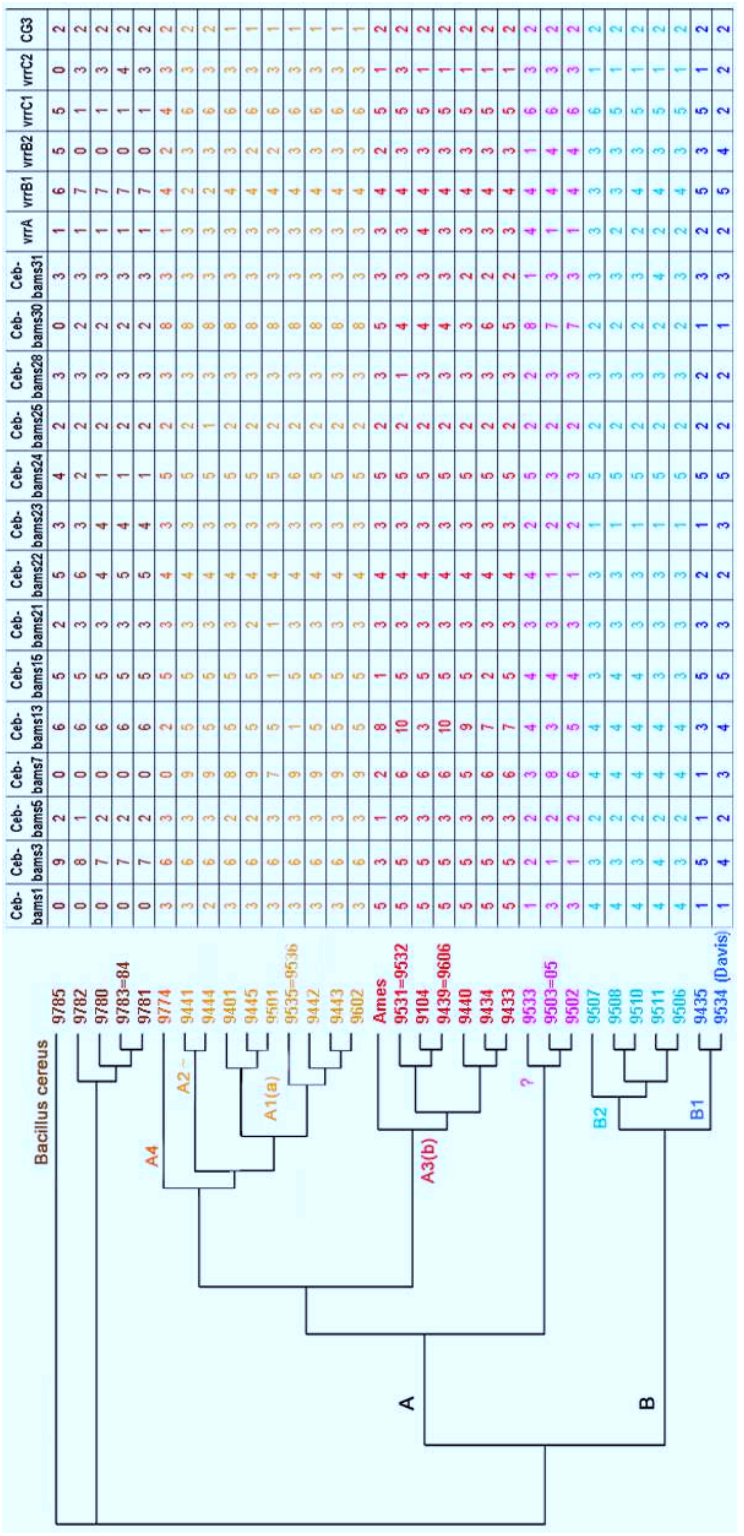


Figure 6

Bacillus anthracis **phylogenetic tree** The genotype of each strain for the polymorphic minisatellites is given (size estimates for each allele are given in Table 3). "0" indicates a failure of the PCR amplification. This is most often associated with *B. cereus* strains, and probably reflects in these cases sequence divergence in the flanking sequence. The phylogenetic tree was produced using the Neighbor-Joining method as available on-line at [http://www.infobiogen.fr.]

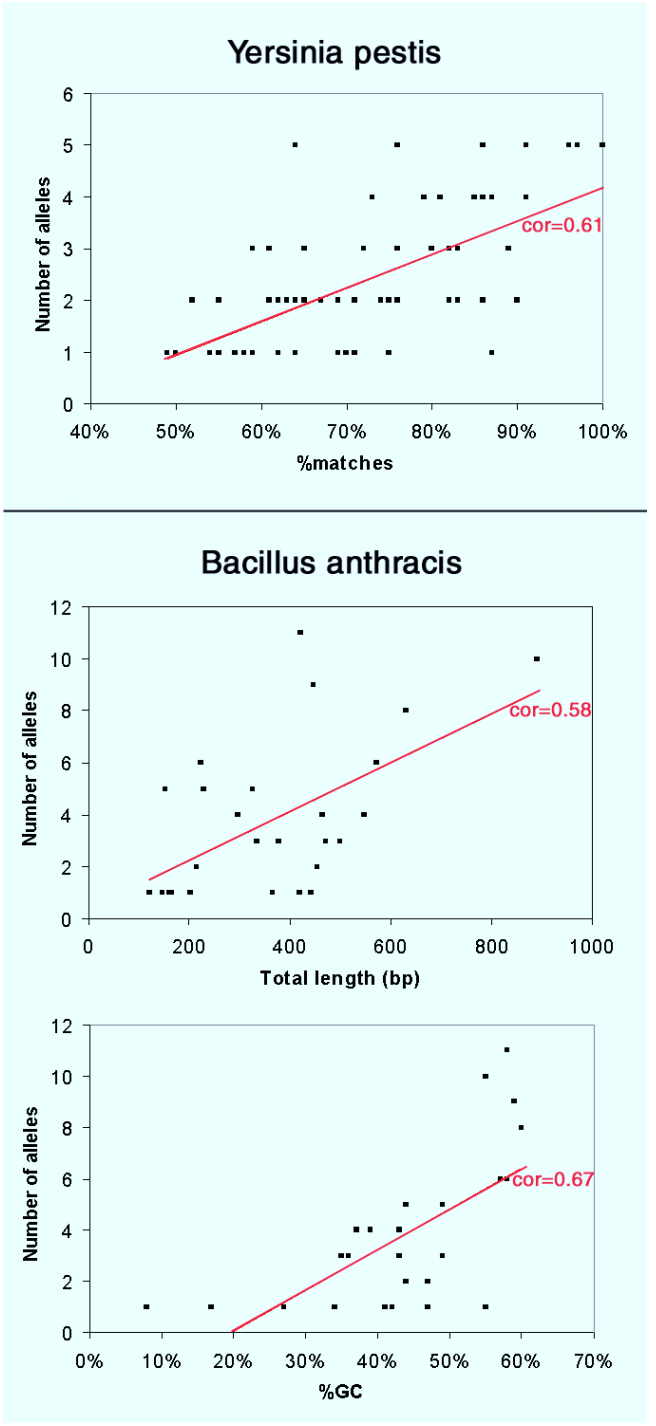


Figure 7
Significant correlation between number of alleles and minisatellites structural characteristics The number of alleles is plotted as a function of Total length and %GC for *Bacillus anthracis*, and %matches for *Yersinia pestis* (the correlations are highly significant at the 0.01 level). Number of alleles for each locus is the total number detected (i.e. *Bacillus anthracis* and *B. cereus*; *Yersinia pestis* and *Y. pseudotuberculosis*).

the PCR products where run on standard 1% or 2% agarose gel (Qbiogen) in 0.5 x TBE buffer at a voltage of 10 V/cm as indicated in Tables 1 and 2. Gel length of 10 to 40 cm were used according to PCR product size and motif length. Gels were stained with ethidium bromide and visualized under UV light. Allele sizes were estimated using as size markers the 1 kb ladder plus (Gibco-BRL which also includes a 100 bp ladder between 100 bp and 500 bp, plus 650, 850 and 1000 bp bands) or the 50 bp ladder (Euromedex) which provides a 50 bp ladder between 50 and 300 bp and a 100 bp ladder from 300 bp to 1000 bp.

Data analysis

Tandem Repeats Finder analysis:

Sequences were processed using the Tandem Repeats Finder software ([http://c3.biomath.mssm.edu/trf.html]). The output was processed to eliminate duplicates before being imported in a database (running under Access2000, Microsoft Corp.) as described previously [12]. The *B. anthracis* preliminary sequence data file uses FASTA type of headers (i.e. >sequenceId) to separate the independent contigs. The headers were replaced by runs of 10 Ns before running Tandem Repeats Finder.

Blast queries against the M. tuberculosis genome:

The identifications of the open reading frames containing a given tandem repeat from M. tuberculosis were done by running a BLAST search on the dedicated web page at [http://www.sanger.ac.uk/Projects/M_tuberculosis/blast_server.shtml] .

Estimation of the excess of tandem repeats with motif length multiple of three:

A χ^2 test was calculated for the difference between the observed number of tandem repeats with motif length multiple of 3 and the expected number of tandem repeats with motif length multiple of 3 (expected value in the absence of bias being the total number of tandem repeats divided by 3). The χ^2 values vary from 0.01 to 253.5. There is a significant excess ($\chi^2 > 3.841$) for all species but 6 (*Buchnera sp*, *T. maritima*, *H. influenzae*, *M. genitalium*, *R. prowazekii*, *Y. pestis*).

Polymorphism index:

Polymorphism Information Index (PIC) or Nei's diversity index is calculated as $1 - \Sigma (\text{allele frequency})^2$ based upon the unique genotypes.

Phylogenetic reconstruction:

A phenetic approach, based on a distance matrix was used. Distance matrix between strains was obtained by counting the number of differences between the corresponding genotypes. Then, Neighbor Joining cluster

analysis was performed with Phylip [23] accessed at [http://www.infobiogen.fr/] . An outgroup was arbitrary chose among *Bacillus cereus* strains (9785) and input order of species was randomised.

Data (genotypes, distance matrix, phylogenetic tree) are available at [http://minisatellites.u-psud.fr/ASPSamp/Phylogenie/data.htm]

Correlation analysis

Correlations were calculated with the statistical program SPSS: Pearson correlation, and non-parametric correlations (Kendall's tau and Spearman's rho) show similar results.

Table 1: Description of *Yersinia* polymorphic markers

Marker	U	N	% GC	V	Primer sequences	PCR	Expected product length (bp)	Estimated size range (bp)	Number of Alleles in *Y.pestis*	Total number of alleles
Markers polymorphic in *Yersinia pestis* strains										
yp0120ms01	18	8	34	86	**L:** CTAAGCACAATTGTTATGCTGAACC **R:** TACTGAATCTGCTTCATTGTTCAAA		228	180 - 280	3	4
yp1290ms04	17	8	27	96	**L:** CGCTGTTGAAGTTTTAGTGTAAGAA **R:** AAATGTAACCTGCCAAACGTG		230	160 - 240	3	5
yp1935ms05	17	11	36	87	**L:** CCTCAGTTCATTGTGTAAAATCTCA **R:** GTATTAGCGAGATCACAGATGAGC		291	190 - 300	2	4
yp2769ms06	60	9	48	64	**L:** AATTTTGCTCCCCAAATAGCAT **R:** TTTTCCCCATTAGCGAAATAAGTA	90 s	606	370 - 2500	3	5
yp2916ms07	10	9	44	85	**L:** ATACCGCTACGATCAGCCTCTAT **R:** ATTTAATATTGATTTTGGGACTTGC		184	150 - 200	2	4
yp3057ms09	18	33	65	91	**L:** CGTTACCCTTGTTGCCAATAGT **R:** ACGCAGAACATGCTTACCTTTTAT	90 s	682	500 - 820	3	5
yp0559ms15	15	10	30	62	**L:** TTGACCAAGTGTAAAAGGCATAAAT **R:** AAACTATCGCCAGCCATTTTAGTA		237	225 - 250	2	2
yp1814ms20	15	9	47	74	**L:** ACAACCTCAGTTTGCCCTTG **R:** GTAAAGAGCGCAATGATCGTACT		253	230 - 250	2	2
yp1895ms21	18	9	51	76	**L:** GCTTAAAGCAGATTGATACTCACG **R:** CTGCATGTTCACCGGTTCAG		278	220 - 350	3	5
yp4042ms35	15	8	41	59	**L:** CTGTTACCGGTCAAAGTGGATATT **R:** AGGCTCTCCTTATCATTATTTGGTC		204	195 - 225	2	3
yp4425ms38	16	8	41	86	**L:** GTGAGGTATAGCTAAACGGTGATGT **R:** CGCCGTAGATTATTTGTCACTTTAT		233	200 - 380	3	5
yp0581ms40	17	7	28	76	**L:** GCAATCATTCACCTAACCATATCTC **R:** GTGCAATAGGCGTTGTTGTGTA		214	220 - 240	2	2
yp0718ms41	17	7	41	75	**L:** GAAGAAAGCCAGCTAATCTGATG **R:** TAATGAATAGCAACGACAACCAATA		217	180 - 220	2	2
yp1018ms44	17	7	38	61	**L:** CAATTCCAACAGCTATTAATGCAA **R:** GAATTTTCATAACACGTTCTTCCTG		233	220 - 260	2	3
yp1108ms45	12	7	65	79	**L:** GCATCGGAGACTGGGTAAAC **R:** TTTCTGAGGATTTATCGGTGTGAT		161	120 - 170	3	4
yp1335ms46	17	7	33	73	**L:** CAGGTTTTACGTTATTTTCTGAAGG **R:** CAGCATGAAGTATGACGGGTATATTA		252	230 - 310	3	4
yp2058ms51	18	7	37	65	**L:** GGTTTTTACCGATATAAATCCTGAG **R:** GACCAAGAAGTTAAGTTGCTTATCG		207	190 - 210	2	2
yp2612ms54	22	7	28	82	**L:** GTCCACCATTTTCATACTGTCACTT **R:** GCTCTTTGTTCGATTTTATTGAATG		281	250 - 300	2	3
yp3060ms56	16	7	21	81	**L:** AACCGACTGACTCACTTTATATTGG **R:** TTCTTTTCCATTACTCAGCCTGTT		220	180 - 250	2	4
yp4280ms62	9	7	33	60	**L:** TTTAGTCTTGATTAAGCTGCGTTTT **R:** ACGGAAGACAACCTTATTATTGATG		240	220 - 310	3	5
yp1118ms69	16	6	39	82	**L:** GACGTTGCAACTGCAAAAATAAG **R:** ACTTGTTGTGAAGACCATCACTCT		179	165 - 180	2	2
yp1580ms70	9	6	32	97	**L:** AAACCAACGGTTCATATTGAATAAA **R:** CTTCTTCCGCTATTTTCCTACAGA		146	140 - 170	3	5
yp1925ms71	14	6	45	91	**L:** GCTACTCGAATATGAGTTAGCCAAA **R:** ATTGCCATATTGGATGCTAAAATAA		171	145 - 210	2	4
yp3236ms73	18	6	40	89	**L:** AATACCCTGTGGGTGATAATGAAC **R:** ATCGATTTAGGTACCACCAATTCA		225	175 - 230	2	3
yp3245ms74	15	6	44	83	**L:** CCCCGACTTATATCAAGCACTG		195	180 - 225	3	3

Table 1: Description of *Yersinia* polymorphic markers

					R: AACTGACGATCTTTTTCACTGAGTT					
Markers polymorphic in 5 *Yersinia* strains (monomorphic in *pestis*)										
yp0802ms02	**18**	**12**	**49**	**86**	**L:** CTGACACAAAACGAGAGCCTATTT **R:** AGCGTGAGTGGGCTATCAATAC	53°C 1 min	**314**	**240 - 315**	**1**	**2**
yp2925ms08	**15**	**12**	**39**	**63**	**L:** AGCCTTTTTGTTGATTATCAGTCAT **R:** CGATAATAATACTGAATTACCGGATG		**270**	**270 - 290**	**1**	**2**
yp4411ms10	**14**	**8**	**32**	**69**	**L:** ATCATGCTTTTGCCTCAATATAATC **R:** GAAACGCAGTCCCTGTTGTAG		**191**	**190 - 210**	**1**	**2**
yp0813ms16	**17**	**8**	**39**	**64**	**L:** GTTGGTTATCCGACAGTCTTCAATA **R:** GCAATTCGTTATGGCTTAGTAAAAA		**235**	**230 - 270**	**1**	**2**
yp1269ms18	**27**	**9**	**54**	**55**	**L:** GCAAAAGCTGAAGCAGATAAAATAG **R:** AAACCACCAAACAAATCATCAAC		**303**	**220 - 250**	**1**	**2**
yp2196ms22	**20**	**8**	**12**	**55**	**L:** AAACCAACAAGAAAACTGTAACCAC **R:** CATTCACCATTGATGTCCTTAGAC	90 s	**265**	**270 - 1500**	**1**	**2**
yp2324ms24	**19**	**8**	**34**	**65**	**L:** TTCACCGGGTTACCTTAATTACATA **R:** CTACCTTGCTGTCAACACTCGAC		**255**	**215 - 255**	**1**	**2**
yp2331ms25	**17**	**9**	**36**	**76**	**L:** AACGCGTTAATAAAACAATAAAGTG **R:** CAATATCCTTTTACTCAGCCGATG		**181**	**190 - 230**	**1**	**3**
yp2679ms27	**16**	**8**	**20**	**76**	**L:** ATGATTACTGGCAAGAGCACTATGT **R:** AACAAGATCACCTGGTCGTTAAA		**217**	**200 - 220**	**1**	**2**
yp2908ms28	**18**	**8**	**40**	**69**	**L:** GCAGAAATAATCTTCAGGAGAAACA **R:** AGATCGTCGTTAGTCCATGTCAG		**242**	**190 - 290**	**1**	**2**
yp2958ms29	**16**	**8**	**23**	**61**	**L:** AAAATAGTCTGTGTTTCAGCAAAGC **R:** CCTTAAAAACCCTAAGTGGGTAAAA		**215**	**215 - 245**	**1**	**2**
yp3225ms30	**54**	**11**	**51**	**52**	**L:** CAATAATACCATCGTGCGTGATAC **R:** TATTAATGGTGGTGTTAGTCGCTGT		**683**	**680 - 900**	**1**	**2**
yp3532ms31	**14**	**8**	**30**	**67**	**L:** GTTATTTATTTTTGCCCCAACTTGT **R:** TTAGCCTGTTTGTTCTTCAAATAGC		**217**	**215- 245**	**1**	**2**
yp3787ms32	**18**	**8**	**49**	**65**	**L:** CGATAACGTTAATGCCATCAGTAG **R:** GCGCCGGTAAGTTTTGTTTATTA		**218**	**190 - 240**	**1**	**3**
yp3795ms33	**15**	**8**	**43**	**67**	**L:** CCCTTCTTTTTATGCTTGAAGATACT **R:** GTTGAACCACAGGCTGTTGAG		**210**	**210 - 225**	**1**	**2**
yp4371ms37	**18**	**8**	**35**	**82**	**L:** TACTTAGGCATTGTCTCTTCACTCC **R:** CTGAAATTATCAGTAGGTGGTTCGT		**235**	**235 - 255**	**1**	**2**
yp0999ms43	**17**	**7**	**38**	**80**	**L:** ATTCCACCACCAACAATTATCAC **R:** GGTATTCGTATTGAAGATGACATTG		**211**	**220 - 300**	**1**	**3**
yp1962ms50	**18**	**7**	**34**	**71**	**L:** TACCGAGGTATTCCCTGGTCTAAT **R:** AGTTGACTCCCAGTCACTTTTCC		**225**	**225 - 240**	**1**	**2**
yp3734ms59	**16**	**7**	**36**	**69**	**L:** ATTATCATGACCCTTCCAGTGCTAT **R:** CATCAAAATGCCAGGAGAGTAAC		**218**	**200 - 220**	**1**	**2**
yp4338ms63	**17**	**7**	**38**	**72**	**L:** ATTAACGATTTCTTGTCGCTCAGT **R:** AATCAGTAACGGCATGTGTCAGTA		**194**	**190 - 275**	**1**	**3**
yp0549ms66	**18**	**6**	**41**	**83**	**L:** TAAAAGCGTCAACAAAGTAGGTCAT **R:** GTTCCTGTTGTTGAAAATGCTG		**212**	**200 - 220**	**1**	**2**
yp0782ms67	**18**	**6**	**40**	**90**	**L:** TTCCAGGCTAAAGATATTGACTTTG **R:** CTCGGCTTGTTCTACGTTTAATG		**248**	**250 - 270**	**1**	**2**
yp1053ms68	**18**	**6**	**32**	**82**	**L:** CCGTTATCTGGTGAAAGTGAACAG **R:** GTCCGGTAGCCTGATTGTTTATT		**182**	**175 - 205**	**1**	**3**
yp3634ms75	**15**	**6**	**36**	**80**	**L:** ATGTGAGCTTGATTGCTGAGTAGT **R:** TCATATTTTAGGTGTTTTGCCTTTG		**210**	**180 - 210**	**1**	**3**

Some structural characteristics of the tandem repeats are presented : U (unit length), N (number of repeats), %GC, V (% of conservation). PCR and electrophoresis conditions are as described in the material and methods section : annealing temperature is 60°C, elongation time is 60 seconds and gels are 2% agarose except when indicated otherwise. Total number of alleles means number of alleles in 3 *Y. pestis* and 2 *Y. pseudotuberculosis* strains.

Table 2: Description of *Bacillus anthracis* polymorphic markers

Marker	U	N	% GC	V	Primer sequences	PCR	Expected product length in bp (observed)	Estimated size range (bp)	Number of alleles in B. *anthracis*	Total number of alleles	PIC index
Ceb-Bams 1	21	16	44	88	**L:** GTTGAGCATGAGAGGTACCTTGTCCTTTTT **R:** AGTTCAAGCGCCAGAAGGTTATGAGTTATC		485 (520)	410-520	5	5	0.72
Ceb-Bams 3	15	25	59	73	**L:** GCAGCAACAGAAAACTTCTCTCCAATAACA **R:** TCCTCCCTGAGAACTGCTATCACCTTTAAC	1%	544	480-860	6	9	0.75
Ceb-Bams 5	39	10	49	92	**L:** GCAGGAAGAACAAAAGAAACTAGAAGAGCA **R:**ATTATTAGCAGGGGCCTCTCCTGCATTACC	53°C 60s	307	305-385	3	3	0.56
Ceb-Bams 7	18	49	55	69	**L:** GAATATTCGTGCCACCTAACAAAACAGAAA **R:** TGTCAGATCTAGTTGGCCCTACTTTTCCTC	65°C 1%	1017	600-1950	9	9	0.82
Ceb-Bams 13	9	70	60	79	**L:** AATTGAGAAATTGCTGTACCAAACT **R:** CTAGTGCATTTGACCCTAATCTTGT	120s 1%	814	330-850	8	11	0.79
Ceb-Bams 15	18	12	57	77	**L:** GTATTTCCCCCAGATACAGTAATCC **R:** GTGTACATGTTGATTCATGCTGTTT		409	410-610	5	5	0.59
Ceb-Bams 21	45	11	43	75	**L:** TGTAGTGCCAGATTTGTCTTCTGTA **R:** CAAATTTTGAGATGGGAGTTTTACT		676	540-680	3	3	0.14
Ceb-Bams 22	36	15	39	81	**L:** ATCAAAAATTCTTGGCAGACTGA **R:** ACCGTTAATTCACGTTTAGCAGA		735	590-950	4	6	0.51
Ceb-Bams 23	42	11	37	85	**L:** CGGTCTGTCTCTATTATTCAGTGGT **R:** CCTGTTGCTCCTAGTGATTTCTTAC		653	570-820	3	4	0.49
Ceb-Bams 24	42	11	44	80	**L:** CTTCTACTTCCGTACTTGAAATTGG **R:** CGTCACGTACCATTTAATGTTGTTA		630	335-670	3	6	0.2
Ceb-Bams 25	15	14	45	60	**L:** CCGAATACGTAAGAAATAAATCCAC **R:** TGAAAGATCTTGAAAAACAAGCATT		391	375-390	2	2	0.07
Ceb-Bams 28	24	14	36	70	**L:** CTCTGTTGTAACAAAATTTCCGTCT **R:** TATTAAACCAGGCGTTACTTACAGC		493 (400)	300-400	3	3	0.26
Ceb-Bams 30	27	16	58	78	**L:** AGCTAATCACCTACAACACCTGGTA **R:** CAGAAAATATTGGACCTACCTTCC	120s 1%		200-890	11	11	0.77
Ceb-Bams 31	9	64	58	57	**L:** GCTGTATTTATCGAGCTTCAAAATCT **R:** GGAGTACTGTTTGTTGAATGTTGTTT	 1%	772	300-850	4	4	0.32

Some structural characteristics of the tandem repeats are presented : U (unit length), N (number of repeats), %GC, V (% of conservation). PCR and electrophoresis conditions are as described in the material and methods section : annealing temperature is 60°C, elongation time is 60 seconds and gels are 2% agarose except when indicated otherwise. The expected product length is deduced from the sequencing data corresponding to the Ames strain. When the Ames strains typing does not fit with the expected value, the observed value is indicated between (). Only one side of the Ceb-Bams30 minisatellite can be identified in the available Ames sequence. The other side was identified in the course of the independent, partial sequencing of *B. anthracis* strains (Vergnaud and col., unpublished data). Total number of alleles includes alleles observed in the *B. cereus* strains. Polymorphism Information Index (PIC) or Nei's diversity index is calculated as $1 - \Sigma (\text{allele frequency})^2$.

Table 3: Correspondence between *B. anthracis* allele sizes and allele numbering

allele nb marker name	1	2	3	4	5	6	7	8	9	10
Ceb-Bams1	~ 410	~ 430	~ 450	~ 480	~ 520					
Ceb-Bams3	484	514	544	559	574	589	704	734	857	
Ceb-Bams5	307	346	385							
Ceb-Bams7	603	1017	1305	1503	1557	1647	1809	1899	1953	
Ceb-Bams13	328	382	454	481	490	652	742	787	814	850
Ceb-Bams15	409	535	571	589	607					
Ceb-Bams21	541	631	676							
Ceb-Bams22	591	627	699	735	~ 900	~ 950				
Ceb-Bams23	569	611	653	821						
Ceb-Bams24	336	420	462	504	630	672				
Ceb-Bams25	376	391								
Ceb-Bams28	~ 300	~ 375	~ 400							
Ceb-Bams30	266	375	500	660	695	730	760	850 to		

Table 3: Correspondence between *B. anthracis* allele sizes and allele numbering

							900
Ceb-Bams31	304	700	772	853			
vrrA	289	301	313	325	337		
vrrB1	184	193	220	229	256	~ 280	~ 290
vrrB2	~ 135	153	162	171	~ 180		
vrrC1	400	502	520	538	583	613	685
vrrC2	532	568	607	660			
CG3	153	158					

Alleles have been numbered in increasing size order. When the allele size (in base-pairs) observed in the Ames strain was in agreement with the size expected according to Ames sequence data, the values indicated in the table assume that alleles differ in size by a multiple of the motif length. These likely values will have to be confirmed by more accurate size estimation tools and allele sequencing. When the allele size in Ames is not as expected (Ceb-Bams1 and Ceb-Bams28), the estimated values are preceded by a ~. The Vrr and CG3 allele sizes were described in [2]; new alleles are indicated by a ~.

Acknowledgements

Minisatellite investigations in the laboratory are supported by grants from Délégation Générale de l'Armement (DGA/DSA/STTC and DGA/DSA/SP-Nuc). Preliminary sequence data for *B. anthracis* was obtained from The Institute for Genomic Research through the website at [http://www.tigr.org] . Sequencing of *B. anthracis* was accomplished with support from Office of Naval Research, Department of Energy, and National Institute of Allergy and Infectious diseases. We wish to thank the referees for the significant improvements they have suggested.

References

1. van Belkum A, Scherer S, van Leeuwen W, Willemse D, van Alphen L, Verbrugh H: **Variable number of tandem repeats in clinical strains of Haemophilus influenzae.** *Infect Immun* 1997, **65**:5017-27
2. Keim P, Price LB, Klevytska AM, Smith KL, Schupp JM, Okinaka R, Jackson PJ, Hugh-Jones ME: **Multiple-Locus Variable-Number Tandem Repeat Analysis Reveals Genetic Relationships within Bacillus anthracis.** *J Bacteriol* 2000, **182**:2928-2936
3. Frothingham R, Meeker-O'Connell WA: **Genetic diversity in the Mycobacterium tuberculosis complex based on variable numbers of tandem DNA repeats.** *Microbiology* 1998, **144**:1189-96
4. Supply P, Mazars E, Lesjean S, Vincent V, Gicquel B, Locht C: **Variable human minisatellite-like regions in the Mycobacterium tuberculosis genome.** *Mol Microbiol* 2000, **36**:762-71
5. Adair DM, Worsham PL, Hill KK, Klevytska AM, Jackson PJ, Friedlander AM, Keim P: **Diversity in a variable-number tandem repeat from Yersinia pestis.** *J Clin Microbiol* 2000, **38**:1516-9
6. van Ham SM, van Alphen L, Mooi FR, van Putten JP: **Phase variation of H. influenzae fimbriae: transcriptional control of two divergent genes through a variable combined promoter region.** *Cell* 1993, **73**:1187-96
7. Weiser JN, Love JM, Moxon ER: **The molecular mechanism of phase variation of H. influenzae lipopolysaccharide.** *Cell* 1989, **59**:657-65
8. Bayliss CD, Field D, Moxon ER: **The simple sequence contingency loci of Haemophilus influenzae and Neisseria meningitidis.** *J Clin Invest* 2001, **107**:657-66
9. Henderson IR, Owen P, Nataro JP: **Molecular switches - the ON and OFF of bacterial phase variation.** *Mol Microbiol* 1999, **33**:919-32
10. Wang G, Ge Z, Rasko DA, Taylor DE: **Lewis antigens in Helicobacter pylori: biosynthesis and phase variation.** *Mol Microbiol* 2000, **36**:1187-96
11. Wilton JL, Scarman AL, Walker MJ, Djordjevic SP: **Reiterated repeat region variability in the ciliary adhesin gene of Mycoplasma hyopneumoniae.** *Microbiology* 1998, **144**:1931-43
12. Vergnaud G, Denoeud F: **Minisatellites: Mutability and Genome Architecture.** *Genome Res* 2000, **10**:899-907
13. Kokoska RJ, Stefanovic L, Tran HT, Resnick MA, Gordenin DA, Petes TD: **Destabilization of yeast micro- and minisatellite DNA sequences by mutations affecting a nuclease involved in Okazaki fragment processing (rad27) and DNA polymerase delta (pol3-t).** *Mol Cell Biol* 1998, **18**:2779-88
14. Debrauwère H, Buard J, Tessier J, Aubert D, Vergnaud G, Nicolas A: **Meiotic instability of human minisatellite CEB1 in yeast requires DNA double-strand breaks.** *Nat Genet* 1999, **23**:367-71
15. De Bolle X, Bayliss CD, Field D, van de Ven T, Saunders NJ, Hood DW, Moxon ER: **The length of a tetranucleotide repeat tract in Haemophilus influenzae determines the phase variation rate of a gene with homology to type III DNA methyltransferases.** *Mol Microbiol* 2000, **35**:211-22
16. van Belkum A, Scherer S, van Alphen L, Verbrugh H: **Short-sequence DNA repeats in prokaryotic genomes.** *Microbiol Mol Biol Rev* 1998, **62**:275-93
17. Achtman M, Zurth K, Morelli G, Torrea G, Guiyoule A, Carniel E: **Yersinia pestis, the cause of plague, is a recently emerged clone of Yersinia pseudotuberculosis [published erratum appears in Proc Natl Acad Sci U S A 2000 Jul 5;97(14):8192].** *Proc Natl Acad Sci U S A* 1999, **96**:14043-8
18. Helgason E, Okstad OA, Caugant DA, Johansen HA, Fouet A, Mock M, Hegna I, Kolsto : **Bacillus anthracis, Bacillus cereus, and Bacillus thuringiensis - one species on the basis of genetic evidence.** *Appl Environ Microbiol* 2000, **66**:2627-30
19. Keim P, Kalif A, Schupp J, Hill K, Travis SE, Richmond K, Adair DM, Hugh-Jones M, Kuske CR, Jackson P: **Molecular evolution and diversity in Bacillus anthracis as detected by amplified fragment length polymorphism markers.** *J Bacteriol* 1997, **179**:818-24
20. Benson G: **Tandem repeats finder: a program to analyze DNA sequences.** *Nucleic Acids Res* 1999, **27**:573-80
21. Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, Gordon SV, Eiglmeier K, Gas S, Barry CE, Tekaia F, Badcock K, Basham D, Brown D, Chillingworth T, Connor R, Davies R, Devlin K, Feltwell T, Gentles S, Hamlin N, Holroyd S, Hornsby T, Jagels K, Barrell BG, *et al*: **Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence.** *Nature* 1998, **393**:537-44
22. Engelke DR, Krikos A, Bruck ME, Ginsburg D: **Purification of Thermus aquaticus DNA polymerase expressed in Escherichia coli.** *Anal. Biochem.* 1990, **191**:396-400
23. Felsenstein J: **PHYLIP - Phylogeny Inference Package (Version 3.2).** *Cladistics* 1989, **5**:164-166

2.2.2 Application à l'identification de souches du complexe *Mycobacterium tuberculosis*

L'article suivant (Le Flèche 2002), intitulé « High resolution, on-line identification of strains from the *Mycobacterium tuberculosis* complex based on tandem repeat typing » (identification en ligne, à haute résolution, de souches du complexe *Mycobacterium tuberculosis* basée sur le typage de répétitions en tandem) présente l'application de la base de données des répétitions en tandem à l'identification de marqueurs polymorphes chez *Mycobacterium tuberculosis*. D'autres études avaient démontré la validité du typage par les répétitions en tandem dans le complexe tuberculosis. Cet article effectue une synthèse des travaux antérieurs, et étudie la plupart des répétitions en tandem non encore explorées. La collection de souches utilisées est particulièrement riche en souches du type Africanum. Cette étude est donc relativement unique, à la fois pour le grand nombre de marqueurs utilisés, et pour la représentation du complexe tuberculosis. Elle montre que la classification obtenue par le typage des répétitions en tandem est pertinente d'un point de vue phylogénétique. En outre, un service d'identification de souches en ligne (http://bacterial-genotyping.igmors.u-psud.fr) a été élaboré : l'utilisateur saisit un génotype, et obtient la liste des souches les plus proches, parmi celles qui figurent dans la base de données de génotypes hébergée au laboratoire. Cet article marque une nouvelle évolution de la base. Ma contribution a été la comparaison des trois souches du complexe tuberculosis pour lesquelles la séquence complète avait été déterminée. Bien que dans ce cas la conservation entre les souches soit telle que la comparaison des répétitions en tandem puisse se faire aisément avec la base de données existante, ce projet a été l'occasion du développement de l'outil automatique permettant la comparaison de souches. Cet outil a été décrit plus en détail dans la paragraphe 2.1.1.4. J'ai également participé à l'élaboration de la page d'identification de souches, décrite dans le paragraphe 2.1.1.5.

Résumé :

Contexte : Les méthodes de référence actuellement disponibles pour l'épidémiologie moléculaire du complexe *Mycobacterium tuberculosis* manquent de sensibilité ou sont encore trop lentes et fastidieuses pour être applicables en routine. Le typage de répétitions en tandem est récemment apparu comme une alternative potentielle. Cet article contribue au développement du typage de répétitions en tandem chez *M. tuberculosis* : une synthèse des données existantes a été effectuée, de nouveaux marqueurs polymorphes ont été développés, et un service Internet gratuit, rapide, et facile d'utilisation a été élaboré pour permettre l'identification de souches.

Résultats: Un lot de 21 VNTRs comprenant 13 locus déjà décrits et 8 nouveaux marqueurs a été utilisé pour génotyper 90 souches du complexe *M. tuberculosis* (*M. tuberculosis* : 64 souches ; *M. bovis* : 9 souches dont 4 BCG ; *M. africanum* : 17 souches). 84 génotypes

différents ont été définis. Une analyse de classification montre que les souches de *M. africanum* tombent dans trois grands groupes, l'un étant plus proche des souches de *M. tuberculosis*, et un autre plus proche des souches de *M. bovis.* Les résultats sont publiquement accessibles sur Internet [http://bacterial-genotyping.igmors.u-psud.fr/bnserver] pour permettre des requêtes d'identification de souches.

Conclusions: Le typage de répétitions en tandem, basé sur la technique de PCR, pourrait se révéler être un puissant complément aux outils épidémiologiques existants pour le complexe *M. tuberculosis*. Le nombre de marqueurs à typer dépend de la précision d'identification requise : l'identification peut être effectuée rapidement et à moindre coût en termes de consommables, d'expertise scientifique, et d'équipement.

BMC Microbiology

Research article

High resolution, on-line identification of strains from the *Mycobacterium tuberculosis* complex based on tandem repeat typing

Philippe Le Flèche[1,2], Michel Fabre[3], France Denoeud[2], Jean-Louis Koeck[4] and Gilles Vergnaud*[1,2]

Address: [1]Centre d'Etudes du Bouchet BP3, 91710 Vert le Petit, France, [2]GPMS, Bât. 400, Institut de Génétique et Microbiologie, Université Paris Sud, 91405 Orsay cedex, France, [3]Laboratoire de Biologie Clinique, HIA Percy, 92141 Clamart, France and [4]Département de biologie médicale, HIA Val-de-Grâce, 75230 Paris, France

E-mail: Philippe Le Flèche - lefleche@igmors.u-psud.fr; Michel Fabre - mfabre@free.fr; France Denoeud - France.Denoeud@igmors.u-psud.fr; Jean-Louis Koeck - jlkoeck@filnet.fr; Gilles Vergnaud* - Gilles.Vergnaud@igmors.u-psud.fr

*Corresponding author

Published: 27 November 2002

BMC Microbiology 2002, **2**:37

Received: 17 September 2002
Accepted: 27 November 2002

This article is available from: http://www.biomedcentral.com/1471-2180/2/37

Abstract

Background: Currently available reference methods for the molecular epidemiology of the *Mycobacterium tuberculosis* complex either lack sensitivity or are still too tedious and slow for routine application. Recently, tandem repeat typing has emerged as a potential alternative. This report contributes to the development of tandem repeat typing for *M. tuberculosis* by summarising the existing data, developing additional markers, and setting up a freely accessible, fast, and easy to use, internet-based service for strain identification.

Results: A collection of 21 VNTRs incorporating 13 previously described loci and 8 newly evaluated markers was used to genotype 90 strains from the *M. tuberculosis* complex (*M. tuberculosis* (64 strains), *M. bovis* (9 strains including 4 BCG representatives), *M. africanum* (17 strains)). Eighty-four different genotypes are defined. Clustering analysis shows that the *M. africanum* strains fall into three main groups, one of which is closer to the *M. tuberculosis* strains, and an other one is closer to the *M. bovis* strains. The resulting data has been made freely accessible over the internet [http://bacterial-genotyping.igmors.u-psud.fr/bnserver] to allow direct strain identification queries.

Conclusions: Tandem-repeat typing is a PCR-based assay which may prove to be a powerful complement to the existing epidemiological tools for the *M. tuberculosis* complex. The number of markers to type depends on the identification precision which is required, so that identification can be achieved quickly at low cost in terms of consumables, technical expertise and equipment.

Background

The precise identification of bacterial pathogens at the strain level is essential for epidemiological purposes. Consequently, constant efforts are undertaken to develop easy to use, low cost and standardized methods which can eventually be applied routinely in a clinical laboratory. Newer developments are usually genetic methods based on PCR (Polymerase Chain Reaction) to type variations directly at the DNA level. The development of polymorphic markers is now further facilitated by the availability of whole genome sequences for bacterial genomes. Recently, it has been shown that tandem repeat (usually

called minisatellites or VNTRs for Variable Number of Tandem Repeats) loci provide a source of very informative markers not only in humans where some are still in use for identification purposes (paternity analyses, forensics) but also in bacteria. Tandem repeats are easily identified from genome sequence data, the typing of tandem repeat length is relatively straight forward, and the resulting data can be easily coded and exchanged between laboratories independently of the technology used to measure PCR fragment sizes. Furthermore, the resolution of tandem repeats typing is cumulative, i.e. the inclusion of more markers in the typing assay can, when necessary, increase the identification resolution. However, the density of tandem repeats in bacterial genomes varies from species to species, and not all tandem repeats are polymorphic [1]. In addition, some tandem repeats are so unstable that they have no or little long-term epidemiological value [2]. This indicates that for each species under consideration, tandem repeats must be evaluated using representative collections of strains before they can be used. Tandem repeats for bacterial identification have already proved their utility for the typing of the highly monomorphic pathogens *Bacillus anthracis*, *Yersinia pestis*, [1] and *M. tuberculosis*. In this last case, the value of tandem repeat based identification was recognised very early [3]. The so-called DR (direct repeat) locus is a relatively large tandem repeat locus of unknown biological significance. The motif is 72 bp long, one half is highly conserved, whereas the other half (called the spacer element) is highly diverged. The spoligotyping method [4] takes advantage of these internal variations to distinguish the hundreds of different alleles at this locus, which have been reported in the *M. tuberculosis* complex among the thousands of strains typed so far [5]. Although it is quite powerful, with many advantages, spoligotyping suffers from a lack of resolution compared to the current gold-standard in *M. tuberculosis* genetic identification, IS*6110* typing [6]. IS*6110* typing is an RFLP (Restriction Fragment Length Polymorphism) method using the mobile element IS*6110* as a probe. Strains with a low-copy number of IS*6110* elements (such as most *M. bovis* strains) are poorly resolved by this method. The so-called PGRS (polymorphic GC-rich sequence) method is an other RFLP approach in which the probe used is a GC-rich tandem repeat. The polymorphisms which are scored at multiple loci simultaneously on the Southern blot are variations in the tandem repeats length (and not internal variations at a single locus as assayed by spoligotyping). The profiles generated are very informative, but in comparison with IS*6110* typing, PGRS results are more difficult to score, because the intensity of the bands are highly variable (alleles with a small tandem array yield a lower hybridisation signal) [6]. Both PGRS and IS*6110* typing are hindered by the requirement for relatively large amounts of high quality DNA which is an issue for slow-growing mycobacteria.

More recently, and owing to the release of genome sequence data, the allele-length polymorphism of tandem repeat loci has been evaluated by PCR. Essentially three complementary sets of markers have been developed [7–9]. In the first report, exact tandem repeats (ETRs) were identified by searching the existing literature as well as early versions of the *M. tuberculosis* genome sequence data [7]. The resolution provided by this first set of five loci is lower than both IS*6110* RFLP typing and spoligotyping according to a comparative study [6]. In the second report, a family of tandem repeats characterized by similar repeat units was identified by sequence similarity search in the genome sequence data. A set of 12 loci was selected (including two of the five ETR loci) and the resulting panel has a resolution close to IS*6110* typing according to [10]. In the third report tandem repeats with highly conserved (>95%) motifs longer than 50 bp identified in the *M. tuberculosis* genome sequence have been investigated. Altogether, the currently available collection of polymorphic tandem repeats for the typing of *M. tuberculosis* comprises 27 loci (taking into account duplicates) (Table 1). Fifteen have a polymorphism index above 0.5.

This collection of markers should already provide a typing resolution comparable to the current reference methods. Given that not all tandem repeats present in *M. tuberculosis* have been evaluated for polymorphism, it is likely that the typing resolution of minisatellites could further be improved. Eventually, normalisation work will have to be done in order to promote the use of tandem repeats. A number of the loci analysed are known under different names in different studies, (for instance, ETRD [7] is also known as MIRU4 in [10]; and VNTR 0580 in [11]) and the coding (number of motifs in an allele) of alleles can also be different in different studies, for reasons explained in [11]. This is due in part to the fact that the number of repeats is not necessarily an integer value (Table 1). Furthermore, because the repeats in an array are not necessarily exact repeats, there can be ambiguities in the definition of the first and last base pair of the array. Finally, in addition to length variations due to the addition or deletion of an exact number of units, microdeletions or insertions within some repeat units are sometimes observed (MIRU4 is one such instance [12]).

One purpose of the present report is to contribute to the development of Multiple Loci VNTR Analysis (MVLA) through the evaluation of new markers and the setting up of an on-line identification tool for the *M. tuberculosis* complex which can be queried very easily with the user's personal data. In the present report, we first take advantage of the availability of genome sequence from two *M. tuberculosis* strains to complement the current collection of polymorphic tandem repeat markers. We identified *in silico* tandem repeats showing a different length in the two

Table 1: Polymorphic minisatellite markers for the *M. tuberculosis* complex

Locus name	"MIRU" alias [8]	"ETR" alias [7]	"QUB" alias [9,11]	Other alias	Reference	TR location on H37Rv genome	Expected length in H37Rv (copy number)	Expected length in CDC1551 (copy number)	Expected length in M bovis AF2122 (copy number)	N° of strains	Size range observed (copy number)	N° of alleles observed	Polymorphism index
H37Rv_0024_18 bp				Mtub01	This report	24648	328 (10)	310 (9)	310 (9)	92	274–328 bp (7–10)	4	0.48
H37Rv_0079_9 bp				Mtub02	This report	79503	230 (6)	239 (7)	239 (7)	92	221–275 bp (5–11)	7	0.76
H37Rv_0154_53 bp	MIRU2				[8]	154111	508 (2)	508 (2)	508 (2)	92	455–561 bp (1–3)	3	0.09
H37Rv_0424_51 bp				Mtub04	This report	424010	269 (2.6)	371 (4.6)	269 (2.6)	28	218 – 320 (1.6–3.6)	3	0.52
H37Rv_0531_15 bp				MPTR-A	[7]	531430	328 (16)	328 (16)	328 (16)	48	(15–17)	3	0.23
H37Rv_0577_58 bp		ETR-C			[8]	577172	346 (4)	288 (3)	404 (5)	92	230–404 bp (2–5)	4	0.63
H37Rv_0580_77 bp	MIRU4	ETR-D			[7]	580546	353 (3.3)	330 (3)	483 (5)	92	253–715 bp (2–8)	7	0.35
H37Rv_0802_54 bp	MIRU40				[8]	802194	199 (1)	415 (5)	253 (2)	92	199–469 bp (1–6)	5	0.71
H37Rv_0959_53 bp	MIRU10				[8]	959868	643 (3)	750 (5)	590 (2)	92	537–1014 bp (1–10)	9	0.76
H37Rv_1121_15 bp				Mtub12	This report	1121658	215 (4)	230 (5)	215 (4)	92	200–230 bp (3–5)	3	0.19
H37Rv_1443_56 bp				Mtub16	This report	1443417	291 (1)	347 (2)	347 (2)	11	291–515 (1–5)	3	0.56
H37Rv_1451_57 bp			QUB-1451c		[9]	1451778	305 (3.8)	305 (3.8)	305 (3.8)	56	(2–4) (bovis)	2	0.12
H37Rv_1612_21 bp			QUB-23		[11]	1612529	141 (5)	162 (6)	162 (6)	20	141–203 (5–8)	3	0.18
H37Rv_1644_53 bp	MIRU16				[8]	1644026	671 (2)	724 (3)	671 (2)	92	618–777 bp (1–4)	4	0.59
H37Rv_1895_57 bp			QUB-1895		[9]	1895344	319 (4)	205 (2)	319 (4)	56	(2–4) (bovis)	3	0.35
H37Rv_1955_57 bp				Mtub21	This report	1955580	206 (2)	263 (3)	263 (3)	92	149–491 bp (1–7)	7	0.76
H37Rv_1982_78 bp			QUB-18		[11]	1982873	621 (5)	777 (7)	465 (3)	24	387–1167 (2–12)	9	0.74
H37Rv_2059_77 bp	MIRU20				[8]	2059429	591 (2)	591 (2)	591 (2)	53	(1–2)	2	0.29
H37Rv_2074_56 bp*				Mtub24	This report	2074431	805 (3.6)	693 (1.6)	693 (1.6)	44	637–749 (0.6–2.6)	3	0.52
H37Rv_2163_a_69 bp			QUB-11a	pUCD1	[11]	2163607	305 (3)	581 (7)	788 (10)	92	305–1832 bp (3–26)	15	0.88
H37Rv_2163_b_69 bp			QUB-11b	pUCD1	[11]	2163729	412 (5)	274 (3)	343 (4)	52	136–826 (1–11)	8	0.82
H37Rv_2165_75 bp		ETR-A			[7]	2165223	397 (3)	322 (2)	847 (9)	92	322–847 bp (2–9)	8	0.73
H37Rv_2347_57 bp				Mtub29	This report	2347393	350 (4)	292 (3)	293 (3)	92	236–350 bp (2–4)	3	0.55
H37Rv_2401_58 bp				Mtub30	This report	2401815	319 (2)	435 (4)	435 (4)	92	261–435 bp (1–4)	3	0.55
H37Rv_2461_57 bp		ETR-B			[7]	2461279	292 (3)	235 (2)	406 (5)	92	178–406 bp (1–5)	6	0.51
H37Rv_2531_53 bp	MIRU23				[8]	2531560	873 (6)	820 (5)	767 (4)	92	608–979 bp (1–8)	7	0.60
H37RV_2387_54 bp	MIRU24				[8]	2684427	447 (1)	447 (1)	447 (1)	53	(1–2)	2	0.24
H37Rv_2990_55 bp				Mtub31	This report	2990582	257 (2)	312 (3)	312 (3)	49	202–312 bp (1–3)	3	0.15
H37Rv_2996_51 bp	MIRU26				[8]	2996002	614 (3)	716 (5)	716 (5)	57	563–818 (2–7)	5	0.61
H37Rv_3006_53 bp	MIRU27		QUB-5		[8]	3006875	657 (3)	657 (3)	657 (3)	92	551–710 bp (1–4)	4	0.25

Table 1: Polymorphic minisatellite markers for the *M. tuberculosis* complex *(Continued)*

H37Rv_3171_54 bp				Mtub34	This report	3171465	279 (3)	225 (2)	279 (3)	11	171–225 (1–2)	2	0.3
H37Rv_3192_53 bp	MIRU31	ETR-E			[7]	3192168	651 (3)	651 (3)	651 (3)	92	545–810 bp (1–6)	6	0.67
H37Rv_3232_56 bp			QUB-3232		[9]	3232649	591 (3)	760 (6)	703 (5)	56	(4–22) (bovis)	10	0.65
H37Rv_3239_79 bp		ETR-F			[7]	3239469	476 (2.8)	476 (2.8)	421 (2.1)	48	(1–3)	3	0.49
H37Rv_3336_59 bp			QUB-3336		[9]	3336499	407 (5)	466 (6)	289 (3)	56	(3–21) (bovis)	8	0.55
*H37Rv_3663_63 bp***				Mtub38	This report	3663751	373 (2.7)	310 (1.7)	310 (1.7)	92	247–400 bp (0.7–3.1)	5	0.35
*H37Rv_3690_58 bp**				Mtub39	This report	3690947	341 (2.6)*	397 (3.6)	341 (2.6)	92	247–1349 bp (1–20)*	11	0.64
H37Rv_4052_111 bp			QUB-26		[11]	4052969	708 (5)	819 (6)	597 (4)	100	(4–14) (bovis)	5	0.41
H37Rv_4156_59 bp			QUB-4156c		[9]	4156797	224 (2)	283 (3)	165 (1)	52	106–283 (0–3)	4	0.69
H37Rv_4348_53 bp	MIRU39				[8]	4348401	646 (2)	646 (2)	646 (2)	92	593–699 bp (1–3)	3	0.31

The markers are listed according to their position in the H37Rv genome. The proposed reference name includes the size of the repeat unit. The twenty-one markers used in the present report are italicised and underlined. Alias names identified in the literature are indicated. QUB11a, QUB11b, and ETR-A (position 2163–2165) are located within the gene PPE34 [19]. The expected length assumes that the primers listed in Table 2 were used. * : the observed size (Table 3) is not the expected size. ** : the repeat unit is not easily defined, size variations do not correspond to a multiple of 63 base-pairs. Polymorphism index is calculated as $1 - \sum$ (allele frequency)2 among the 86 distinct genotypes. The values are deduced from the original report in nine cases (indicated by the absence of size range in the "size range" column). In some instances [9,11], the population of strains used is biased (*M. bovis* strains).

strains using the previously described tandem repeat database [http://minisatellites.u-psud.fr][1]. Thirteen loci with a different predicted length in the two genomes and which have not been previously investigated have been tested for polymorphism and ease of typing.

Eight among the 13 polymorphic loci were used together with 13 among the previously described markers to genotype a collection of different *M. tuberculosis* complex strains. The data produced clusters the strains as suggested by morphological observations and biochemical analyses. The resulting data can be queried from a dedicated web page [http://bacterial-genotyping.igmors.u-psud.fr/bn-server].

Table 2: Set of primers for MLVA analysis

Locus name	forward primer	reverse primer
H37Rv_0024_18 bp	GAGAAACAGGAGGGCGTTG	TATTACGACGACCGCTATGC
H37Rv_0079_9 bp	CGTGCACAGTTGGGTGTTTA	TTCGTTCAGGAACTCCAAGG
H37Rv_0154_53 bp	TGGACTTGCAGCAATGGACCAACT	TACTCGGACGCCGGCTCAAAAT
H37Rv_0424_51 bp	GTCCAGGTTGCAAGAGATGG	GGCATCCTCAACAACGGTAG
H37Rv_0531_15 bp	GGTTACCACTTCGATGCGTCTGCG	AGCCGCCGAAACCCATC
*H37Rv_0577_58 bp**	GACTTCAATGCGTTGTTGGA*	GTCTTGACCTCCACGAGTGC*
H37Rv_0580_77 bp	CAGGTCACAACGAGAGGAAGAGC	GCGGATCGGCCAGCGACTCCTC
*H37Rv_0802_54 bp**	AAGCGCAAGAGCACCAAG*	GTGGGCTTGTACTTGCGAAT*
H37Rv_0959_53 bp	GTTCTTGACCAACTGCAGTCGTCC	GCCACCTTGGTGATCAGCTACCT
H37Rv_1121_15 bp	CTCCCACACCCAGGACAC	CGGCCTACCCAACATTCC
H37Rv_1443_56 bp	GGTAATCCTGGTCGCTTGTC	ACCCAAATTGCCCTGGTC
H37Rv_1451_57 bp	GGTAGCCGTCGTCGAGAAGC	CGCCACCACCGCACTGGC
H37Rv_1612_21 bp	GCTGCACCGGTGCCCATC	CACCGGAGCCGGAACGGC
H37Rv_1644_53 bp	TCGGTGATCGGGTCCAGTCCAAGTA	CCCGTCGTGCAGCCCTGGTAC
H37Rv_1895_57 bp	GGTGCACGGCCTCGGCTCC	AAGCCCCGCCGCCAATCAA
H37Rv_1955_57 bp	AGATCCCAGTTGTCGTCGTC	CAACATCGCCTGGTTCTGTA
H37Rv_1982_78 bp*	ATCGTCAGCTGCGGAATAGT*	AATACCGGGGATATCGGTTC*
H37Rv_2059_77 bp	TCGGAGAGATGCCCTTCGAGTTAG	GGAGACCGCGACCAGGTACTTGTA
H37Rv_2074_56 bp	AAATTCAAAGAGTTTCTCGACAGTG	GATCTTGAGAACCAAGATGTCCTT
H37Rv_2163_a_69 bp	CCCATCCCGCTTAGCACATTCGTA	TTCAGGGGGGATCCGGGA
H37Rv_2163_b_69 bp	CGTAAGGGGGATGCGGGAAATAGG	CGAAGTGAATGGTGGCAT
*H37Rv_2165_75pb**	ATTTCGATCGGGATGTTGAT*	TCGGTCCCATCACCTTCTTA*
H37Rv_2347_57 bp	AACCCATGTCAGCCAGGTTA	ATGATGGCACACCGAAGAAC
H37Rv_2401_58 bp	AGTCACCTTTCCTACCACTCGTAAC	ATTAGTAGGGCACTAGCACCTCAAG
H37Rv_2461_57 bp	GCGAACACCAGGACAGCATCATG	GGCATGCCGGTGATCGAGTGG
H37Rv_2531_53 bp	CAGCGAAACGAACTGTGCTATCAC	CGTGTCCGAGCAGAAAAGGGTAT
H37RV_2387_54 bp	CGACCAAGATGTGCAGGAATACAT	GGGCGAGTTGAGCTCACAGAA
H37Rv_2990_55 bp	GTGACGTTTACCGTGCTCTATTTC	GTCGTCGGACAGTTCTAGCTTT
H37Rv_2996_51 bp	CCCGCCTTCGAAACGTCGCT	TGGACATAGGCGACCAGGCGAATA
H37Rv_3006_53 bp	TCGAAAGCCTCTGCGTGCCAGTAA	GCGATGTGAGCGTGCCACTCAA
H37Rv_3171_54 bp	GCAGATAACCCGCAGGAATA	GGAGAGGATACGTGGATTTGAG
*H37Rv_3192_53 bp**	ACTGATTGGCTTCATACGGCTTTA*	GTGCCGACGTGGTCTTGAT*
H37Rv_3232_56 bp	CAGACCCGGCGTCATCAAC	CCAAGGGCGGCATTGTGTT
H37Rv_3239_79 bp	CTCGGTGATGGTCCGGCCGGTCAC	GGAAGTGCTCGACAACGCCATGCC
H37Rv_3336_59 bp	ATCCCCGCGGTACCCATC	GCCAGCGGTGTCGACTATCC
H37Rv_3663_63 bp	GCCCAAAAAGCATGGGAACGTGCCCCT	GGTTGTCCCCGCAGTATCTC
H37Rv_3690_58 bp	AATCACGGTAACTTGGGTTGTTT	GATGCATGTTCGACCCGTAG
H37Rv_4052_111 bp	AACGCTCAGCTGTCGGAT	GGCCAGGTCCTTCCCGAT
H37Rv_4156_59 bp*	TGGTCGCTACGCATCGTGTCGGCCCGT*	TACCACCCGGGCAGTTTAC*
H37Rv_4348_53 bp	CGCATCGACAAACTGGAGCCAAAC	CGGAAACGTCTACGCCCCACACAT

* : the primers indicated are not the primers used in the princeps publication, but were designed for the present study, usually in order to reduce the size of the PCR product and consequently to improve allele size identification.

Results

Tandem repeats predicted to be of a different size in H37Rv and CDC1551

The size of tandem repeats in the two M. tuberculosis strains sequenced to date, H37Rv and CDC1551, was compared using the tandem repeat database [http://minisatellites.u-psud.fr]. Fifty-one of the tandem repeats identified in CDC1551 have repeat units longer than 9 base-pairs and a predicted overall size which differs from the H37Rv homolog estimate by at least 9 base-pairs. Seventeen have an expected product size above one kilobase. They include the DR locus and members of the family of PGRS sequences [13] and were not investigated further. Eighteen have been analyzed in previous investigations [7–9,11]. Three produced multiband patterns or inconsistent results. The results obtained for the remaining 13 loci together with the description of the 18 previously described loci are summarized in Table 1. In addition, Table 1 includes nine markers which are not polymorphic between H37Rv and CDC1551 but have already been quoted in the literature. Each locus is designated by its position (expressed in kilobases) on the H37Rv genome and by the repeat unit length as defined by the Tandem Repeat Finder software and indicated in the Tandem Repeat Database [http://minisatellites.u-psud.fr]. All thirteen newly evaluated loci are polymorphic as predicted. In two cases (Table 1) the expected product size is not the observed size. The expected size has not been observed in the collection of strains used here, which suggests that the incorrect prediction is due to an artifact along the sequencing process. Eight loci among the thirteen have polymorphism indexes above 0.50 (two are above 0.7). The vast majority of the repeats units are more than 50 bp long (Table 1) which makes them easy to assay by ordinary agarose gel electrophoresis when using the primer pairs indicated in Table 2. In one instance however (H37Rv_3663_63 bp) the PCR size products clearly do not differ by a perfect number of (63 bp) repeat units (Table 1).

Typing of strains and clustering analysis

The forty loci listed in Table 1 were used to genotype a collection of 90 strains from the M. tuberculosis complex, using the primers listed in Table 2. In our hands, some of the markers did not prove to be sufficiently robust for easy and reproducible typing in the conditions used here. On this basis, we have selected a collection of 21 markers (comprising thirteen previously described markers and eight among the new loci evaluated). The 21 markers used are italicised and underlined in Table 1 and 2. After analysis of the images using Bionumerics 3.0, and conversion of allele sizes in copy numbers of motifs in the tandem arrays, clustering analysis was done using the categorical and Ward parameters. The results of the clustering analysis are shown in Figure 1. The genotyping data from strains M. tuberculosis CDC1551 and M. bovis AF2122/97 was deduced (Table 1) from the sequence data and included in the analysis. Six major groups are defined (Figure 1). Group I contains the M. bovis strains and 5 of the M. africanum strains. Group II is composed of nine M. africanum strains. The third group includes three M. africanum strains and seven M. tuberculosis strains. Interestingly, five of these strains have been independently identified as representing the Beijing type [14] (the last two have not been tested). The last three groups comprise the vast majority of the M. tuberculosis strains. M. africanum strains which are negative for nitrate reduction (Africanum I type [15]) are among the first two groups, closer to the M. bovis strains as previously observed [16,17]. In contrast, the three M. africanum strains which are positive for nitrate reduction are in the third group, closer to M. tuberculosis strains. In order to facilitate the comparison with earlier investigations [16,17], Figure 1 displays the genotypes for the five ETR markers, extracted from the full data presented in Table 3. Group I in Figure 1 is reminiscent of group A in [17] and group A1 in [18]. Group II in Figure 1 is reminiscent of group B in [17] and group A2 in [18] which are both characterized by the 42432 ETR pattern.

The ETR panel alone discriminates 44 genotypes (instead of 84 with the panel of 21 loci; 86 genotypes when including the CDC1551 and AF2122/97 data, Figure 1) and is not sufficient to clearly separate the *M. africanum* strains from the *M. tuberculosis* strains (analysis not shown) as can be achieved using the 21 loci.

Internet-based identifications

The genotyping data presented in Table 3 can be queried directly via an internet service [http://bacterial-genotyping.igmors.u-psud.fr/bnserver/]. Figure 2 provides a brief description of the current *M. tuberculosis* query page (likely to evolve as updates are made). For each locus, allele sizes can be selected among a list of possibilities (observed sizes). Alternatively, more experienced users will go directly to a "copy-paste" page using the appropriate format. The results of the query indicate a similarity score and include links to the complete data for each strain listed. Help files are available, including a link to updated versions of Figure 1.

Testing the reproducibility of the approach

In order to test the reproducibility of the approach, ten blinded-coded control samples were typed. Figure 3 shows the typing of two markers, H37Rv_0802_54 bp (left, 54 bp unit; H37Rv allele : 1 unit, 199 bp PCR product) and H37Rv_1955_57 bp (right, 57 bp unit; H37Rv allele : 2 units, 206 bp PCR product). The number of units in each allele can be unambiguously deduced by comparison with the H37Rv control lanes and the 100 base-pairs

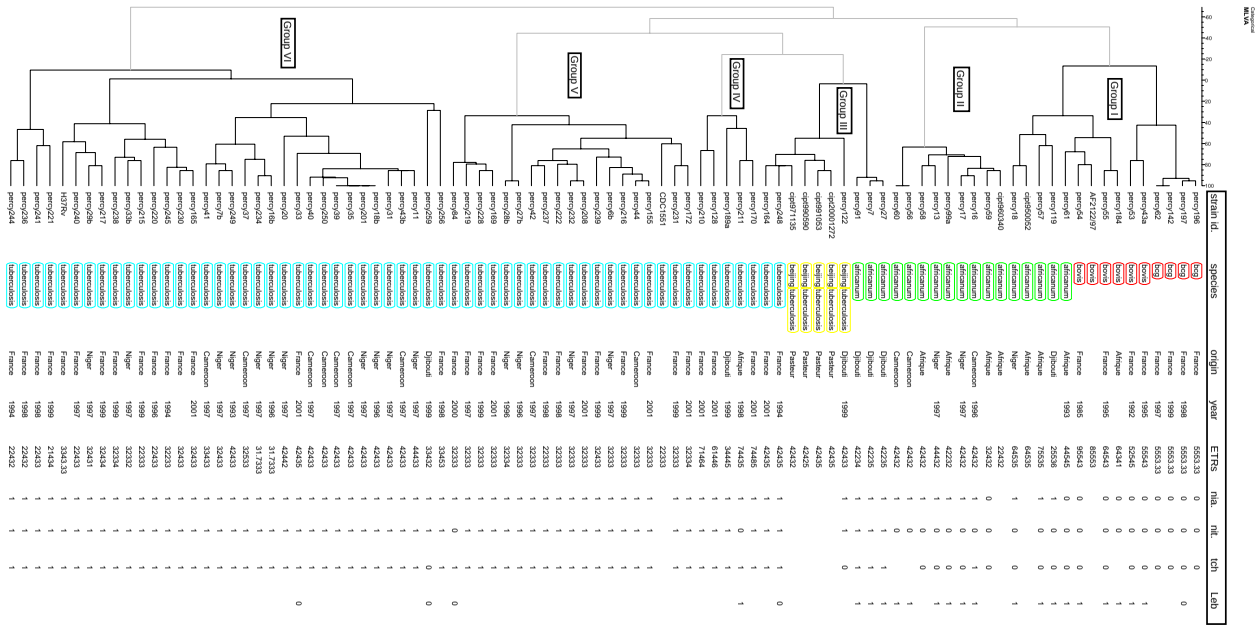


Figure 1

Dendrogram deduced from the clustering analysis of the 92 strains (including CDC1551 and AF2122/97) . The first column from the left identify the strains. The second column indicates the species (Red : *M. bovis* strains; green : *M. africanum* strains; yellow : *M. tuberculosis* strains known to be of the Beijing type and indicated "beijing tuberculosis"; blue : other *M. tuberculosis* strains). The third column indicates the geographic origin of the strain. The fourth column indicates the ETR pattern (ETR-A to ETR-E) extracted from the full data presented in Table 3. The last four columns indicate, from left to right, the result of the niacine production, nitrate reductase, TCH susceptibility and Lebek tests (0, negative ; 1, positive) when available.

ladder size marker. All ten unknown strains were correctly identified using the internet base service described above.

Discussion

The list of 40 markers given in Table 1 is close to representing the complete collection of tandem repeats of interest for MLVA typing in *M. tuberculosis*. It includes all loci with a different predicted size in H37Rv and CDC1551 and which are amenable to routine PCR typing. Nine additional loci which have been quoted in published reports are also included even if they do not fulfill this criteria. Clustering analysis (Figure 1) shows that the two strains CDC1551 and H37Rv (Figure 1) are relatively distant within the *M. tuberculosis* species. This would predict that tandem repeats of identical size in the two strains are likely to be poorly informative across the complex. However, this appears not to be absolutely true, since for instance, ETR-E (H37Rv_3192_53 bp) happens to have the same size in H37Rv, CDC1551 and even AF2122/97 (Table 1) in spite of its very high polymorphism index (0.69, Table 1). Consequently, the few additional loci, not explored here, which are of equal size in H37Rv and CDC1551, but differ with the predicted size for *M. bovis* strain AF2122/97 might also prove to be of interest.

As can be seen in Table 1, most repeat units are more than 50 bp long and allele sizes rarely exceed 1000 bp. As a result, the precision which can be achieved by ordinary agarose gel electrophoresis is sufficient to estimate the number of units in an allele. The selection of 21 markers proposed here was tested specifically in order to be easily assayed using this low-cost technological approach. Although a database system is necessary to efficiently manage a genotyping project with a high number of markers and strains, the identification of up to a few strains per day in a clinical setting for instance requires no sophisticated equipment nor costly consumables. Genotypes can be scored by visual analysis of the gel images, and a subset of the collection of available markers can be chosen for routine identification purposes. The data can then be analysed using the site described in Figure 2.

The role of tandem repeats in the *M. tuberculosis* genome is largely unknown. Twenty-one of the loci listed in Table 1 have repeat units which are a multiple of three basepairs. The majority (fifteen) falls within putative genes, often of unknown function, such as the PPE family of genes [19]. The most remarkable instance is probably PPE34 at position 2163–2165 of the genome (Rv1917c in [http://

Figure 2
Internet database interrogation page . The query page can be accessed via [http://bacterial-genotyping.igmors.u-psud.fr/bnserver]. The home page (not shown) includes a link to help files (and data updates information), and links to individual species query pages. Currently, identification pages are available for *Y. pestis*, *B. anthracis* (based on the data published in [1] and some additional unpublished data) and *M. tuberculosis*. Figure 2 shows the current *M. tuberculosis* query page. For each marker, allele sizes can be selected among the list of observed sizes. Allele sizes are indicated either as number of motifs, or as fragment sizes, assuming that the primers used are the primers listed in Table 2. The allele size listed in green corresponds to the H37RV control strain allele. More experienced users can go directly to a page on which data (expressed in base-pairs or in repeat unit number) can be directly pasted using the appropriate format.

genolist.pasteur.fr/TubercuList/]) which contains three minisatellites [20] (Table 1, Qub11a, Qub11b, ETR-A).

The present study includes 17 *M. africanum* strains. All strains have been identified as such independently, based on morphological features of the colonies grown on Lowenstein-Jensen medium, and biochemical analyses. *M. africanum* has long since been recognized as showing an extensive phenotypic heterogeneity [21], suggesting that *M. africanum* could display a phenotypic continuum between *M. tuberculosis* and *M. bovis*. This was recently supported by the study of deletion events distinguishing the H37Rv *M. tuberculosis* strain and the BCG *M. bovis* strain [22] and suggesting that *M. bovis* is the most recent member of the *M. tuberculosis* complex. The analysis of deletion events in the *M. africanum* strains investigated showed that West African strains fall into two groups, clearly distinguished from the *M. tuberculosis* strains. In contrast, no deletion event distinguished East African *M. africanum* strains from *M. tuberculosis* strains. The present study includes three Africanum type II strains (positive nitrate reductase test). All three originate from East Africa (Djibouti). Although the MLVA analysis presented here does confirm that they are very close to *M. tuberculosis* strains, they are clearly distinct, at least within the collection of strains evaluated. Interestingly, they appear to be closest to the Beijing type of *M. tuberculosis* strains (Figure 1, Group III, strains percy7, percy27 and percy91).

Conclusions

In its present form, the database should be considered as preliminary. More strains must be typed in order to provide a continuous and robust coverage of the *M. tuberculosis* complex, and the clustering analysis presented in Figure 1 should be considered as provisional. If the MLVA approach is considered to be of use by the community, and given that the associated data is highly portable, then it should be relatively easy, through collaborative efforts, to significantly expand the available data. It is hoped that this data will constitute an easy-to-use high-resolution classification resource which will then help address medical and epidemiological issues regarding the *M. tuberculosis* complex.

Methods

Strains and DNA preparation

Identification of mycobacteria used conventional morphological and biochemical tests as previously described [23]. In particular, *M. tuberculosis*, *M. africanum* and *M. bovis* were distinguished according to their morphology on Lowenstein-Jensen plates. *M. tuberculosis* strains are eugonic. The dysgonic *M. africanum* strains colonies are rough and flat. The dysgonic *M. bovis* colonies are smooth, hemispheric and white. Biochemical analyses included niacin production, nitrate reduction, TCH (thiophene-2-carboxylic acid hydrazide) sensitivity tests and growth characteristics on Lebek medium. DNA for PCR analysis was prepared using a simple thermolysis procedure. Briefly, a few colonies were resuspended in 1 ml water, and in-

Table 3: Genotype data for 21 loci and 92 strains (including CDC1551 and AF2122/97)

strain id.	species	24	79	154	577	580	802	959	1121	1644	1955	2163	2165	2347	2401	2461	2531	3006	3192	3663	3690	4348
percy196	bcg	9	11	2	5	3.3	2	2	4	3	1	9	5	2	2	5	4	3	3	1.7	2	2
percy197	bcg	9	11	2	5	3.3	2	2	4	3	1	6	5	2	2	5	4	3	3	1.7	2	2
percy142	bcg	9	11	2	5	3.3	2	2	4	3	1	11	5	2	2	5	4	3	3	1.7	2	2
percy62	bcg	9	11	2	5	3.3	2	2	4	3	1	11	5	2	2	5	4	3	3	1.7	2	2
percy43a	bovis	9	10	2	5	4	2	2	4	3	1	10	5	3	4	5	4	1	3	1.7	2	2
percy53	bovis	9	9	2	5	4	2	2	4	3	1	11	5	3	4	2	4	2	5	1.7	2	2
percy184	bovis	9	8	2	3	4	2	2	4	2	3	6	6	3	4	4	4	3	1	1.7	3	2
percy55	bovis	9	8	2	5	4	2	2	4	3	3	6	6	3	4	4	4	3	3	1.7	3	2
AF2122	bovis	9	7	2	5	5	2	2	4	2	3	10	8	3	4	5	4	3	3	1.7		2
percy54	bovis	9	8	2	5	4	1	2	4	2	3	10	9	3	4	5	4	3	3	1.7	2	2
percy61	africanum	9	7	2	5	4	2	5	4	2	3	10	4	3	4	4	4	3	5	1.7	4	2
percy119	africanum	9	8	2	5	3	1	7	4	4	3	10	2	3	4	5	4	3	6	1.7	1	2
percy57	africanum	9	6	2	5	3	1	10	4	2	4	10	7	3	4	5	4	3	5	1.1	5	2
cipt950052	africanum	9	7	2	5	3	2	4	4	4	4	10	6	2	4	4	4	3	5	1.7	3	2
percy18	africanum	9	8	2	5	3	2	5	4	4	4	10	6	3	4	4	4	3	5	1.7	6	2
cipt960340	africanum	9	5	2	4	3	1	4	4	4	2	9	2	3	4	2	4	4	2	1.7	4	2
percy59	africanum	9	6	2	4	3	1	4	4	4	2	9	3	3	4	2	4	4	2	1.7	3	2
percy16	africanum	9	5	2	4	3	1	4	4	3	2	9	4	3	4	2	4	4	2	1.7	3	2
percy17	africanum	9	5	2	4	3	1	4	4	4	2	8	4	3	4	2	4	4	2	1.7	3	2
percy99a	africanum	9	6	2	2	3	1	4	4	4	2	9	4	3	2	2	4	4	2	1.7	1	2
percy13	africanum	9	5	2	4	3	1	4	4	3	2	9	4	3	4	4	4	3	2	1.7	2	2
percy58	africanum	9	5	2	4	3	1	4	4	4	2	9	4	3	2	2	4	3	2	1.7	4	2
percy56	africanum	9	5	2	4	3	1	4	3	4	2	9	4	3	4	2	2	4	2	1.7	3	2
percy60	africanum	9	5	2	4	3	1	4	3	4	2	9	4	3	4	2	2	4	2	1.7	3	2
percy27	africanum	9	9	2	2	3	3	4	3	1	4	10	4	4	2	2	5	3	5	1.7	3	3
percy7	africanum	9	9	2	2	3	3	4	3	3	4	10	4	4	2	2	5	3	5	1.7	3	3
percy91	africanum	9	9	2	2	3	3	4	3	3	4	10	4	4	2	2	5	3	4	1.7	3	3
percy122	beijing tuberculosis	9	6	2	4	3	1	3	4	3	6	11	4	4	4	2	5	1	3	0.7	3	2
cipt20001272	beijing tuberculosis	9	11	2	4	3	3	3	4	4	5	6	4	4	4	2	5	3	5	0.7	3	2
cipt991053	beijing tuberculosis	9	11	2	4	3	3	2	4	3	5	9	4	4	4	2	5	3	5	0.7	3	2
cipt990590	beijing tuberculosis	9	11	2	4	2	3	2	4	3	5	9	4	4	4	2	3	3	5	0.7	3	3
cipt971135	beijing tuberculosis	9	11	2	4	3	3	3	4	3	4	9	4	4	4	2	5	3	2	0.7	3	3
percy248	tuberculosis	9	11	2	4	3	3	3	3	3	1	14	4	4	4	2	5	3	5	0.7	3	3
percy164	tuberculosis	9	6	2	4	3	3	3	4	3	5	10	4	4	4	2	5	3	5	0.7	3	3
percy170	tuberculosis	9	8	2	4	8	2	6	3	2	5	10	7	3	2	4	5	3	5	1.7	4	3
percy211	tuberculosis	9	8	2	4	3	3	3	4	2	5	11	7	3	2	4	5	3	5	1.7	4	3
percy189a	tuberculosis	9	9	2	4	4	3	4	4	3	6	5	3	3	2	4	4	3	5	1.7	7	3
percy128	tuberculosis	9	8	2	4	4	4	3	4	3	6	6	6	3	1	1	8	3	6	1.7	4	3
percy210	tuberculosis	9	8	2	4	6	4	4	4	3	7	24	7	3	1	1	6	3	4	1.7	4	3
percy172	tuberculosis	9	10	2	3	3	3	4	5	3	3	10	3	4	4	2	5	3	4	1.7	3	2
percy231	tuberculosis	9	10	2	3	3	2	4	4	3	3	11	3	4	4	2	5	3	3	1.7	3	2
CDC1551	tuberculosis	9	7	2	3	3	5	5	5	3	3	7	2	3	4	2	5	3	3	1.7		2
percy155	tuberculosis	9	8	2	3	3	3	5	4	3	2	10	3	4	4	2	5	3	3	1.7	3	2
percy44	tuberculosis	9	8	2	3	3	3	5	4	3	2	8	3	4	4	2	5	3	3	1.7	3	2
percy216	tuberculosis	8	8	2	3	3	3	5	4	3	2	7	3	4	4	2	5	3	3	1.7	3	2
percy6b	tuberculosis	9	8	2	3	3	4	5	4	3	4	7	3	4	4	2	5	3	3	1.7	3	2
percy239	tuberculosis	9	8	2	4	3	2	5	5	3	2	24	3	4	4	2	5	3	3	1.7	3	2

Table 3: Genotype data for 21 loci and 92 strains (including CDC1551 and AF2122/97) *(Continued)*

percy208	tuberculosis	9	8	2	3	3	3	5	4	3	3	11	3	4	4	2	5	3	3	1.7	5	2
percy232	tuberculosis	9	8	2	3	3	3	5	4	3	3	11	3	4	4	2	5	3	3	1.7	3	2
percy222	tuberculosis	9	8	2	3	3	3	5	4	2	3	10	3	4	4	2	5	4	3	1.7	3	2
percy237	tuberculosis	9	8	2	3	3	3	5	4	3	3	6	2	2	4	2	5	3	3	1.7	3	2
percy42	tuberculosis	9	8	2	3	3	2	5	4	3	3	4	3	3	4	2	5	3	3	1.7	3	2
percy27b	tuberculosis	9	8	2	3	3	6	3	4	3	4	7	3	4	4	2	5	3	3	1.7	3	3
percy28b	tuberculosis	9	8	2	3	3	6	3	4	3	4	7	3	4	4	2	5	3	4	1.7	3	3
percy169	tuberculosis	9	8	2	3	3	3	5	4	3	2	5	3	2	2	2	3	3	3	1.7	3	2
percy228	tuberculosis	9	8	2	3	3	3	5	4	3	3	24	3	2	4	2	3	3	3	1.7	3	2
percy219	tuberculosis	9	9	2	3	3	3	5	4	3	2	9	3	2	4	2	3	3	3	1.7	6	2
percy84	tuberculosis	9	8	2	3	3	3	4	4	2	2	11	3	2	4	2	3	3	3	1.7	3	2
percy256	tuberculosis	9	6	1	4	5	3	7	4	4	4	12	3	4	2	3	4	3	3	1.7	1	2
percy259	tuberculosis	10	6	2	4	3	2	1	5	3	1	26	3	2	2	3	5	1	2	1.7	3	1
percy11	tuberculosis	10	6	2	4	3	3	3	4	3	3	6	4	4	2	4	5	3	3	1.7	2	2
percy43b	tuberculosis	10	6	2	4	3	1	3	4	3	3	6	4	4	2	2	5	3	3	1.7	20	2
percy31	tuberculosis	10	6	2	4	3	3	3	4	2	3	6	4	4	2	2	5	3	3	1.7	17	2
percy18b	tuberculosis	10	6	2	4	3	3	3	4	3	3	6	4	4	2	2	5	3	3	1.7	3	2
percy201	tuberculosis	10	6	2	4	3	3	3	4	3	3	6	4	4	2	2	5	3	3	1.7	3	2
percy35	tuberculosis	10	6	2	4	3	3	3	4	3	3	6	4	4	2	2	5	3	3	1.7	3	2
percy39	tuberculosis	10	6	2	4	3	3	3	4	3	3	6	4	4	2	2	5	3	3	1.7	3	2
percy250	tuberculosis	10	6	2	4	3	3	3	4	3	3	6	4	2	2	2	5	3	3	1.7	3	2
percy40	tuberculosis	10	6	2	4	3	3	3	4	3	3	6	4	4	2	2	5	3	3	1.7		2
percy33	tuberculosis	10	6	2	4	3	3	3	4	3	2	25	4	4	2	2	5	3	5	1.7	4	2
percy20	tuberculosis	10	6	3	4	4	2	3	4	3	1	4	4	4	2	2	5	3	2	1.7	3	2
percy16b	tuberculosis	10	6	2	3	3	3	3	4	3	3	24	3	4	2	1.7	5	3	3	1.7	3	2
percy234	tuberculosis	10	6	2	3	3	3	3	4	3	3	25	3	4	2	1.7	5	3	3	1.7	2	2
percy37	tuberculosis	10	6	2	5	3	3	3	4	4	3	24	3	4	2	2	5	3	3	1.7	4	2
percy249	tuberculosis	10	6		4	3	3	3	4	2	3	25	4	4	2	2	5	3	3	1.7	2	2
percy7b	tuberculosis	10	6	2	4	3	3	3	4	2	3	25	3	4	2	2	5	3	3	1.7	9	2
percy41	tuberculosis	10	6	2	4	3	4	3	4	2	3	25	3	4	2	3	5	3	3	1.7	10	2
percy165	tuberculosis	10	6	2	4	3	3	3	4	2	2	10	3	4	2	2	5	3	3	1.7	3	2
percy230	tuberculosis	10	6	2	4	3	4	2	4	2	2	6	3	4	2	2	5	3	3	1.7	3	2
percy245	tuberculosis	10	6	2	2	3	2	2	4	2	2	11	3	4	2	2	5	3	3	1.7	3	2
percy220	tuberculosis	8	5	2	4	3	2	2	4	2	2	24	2	4	2	2	5	3	3	2.7	3	2
percy215	tuberculosis	10	6	2	3	3	2	3	4	2	2	5	2	4	2	2	5	3	3	2.7	3	2
percy33b	tuberculosis	10	6	2	3	3	2	3	4	2	3	6	3	4	2	2	5	3	2	1.7	3	2
percy238	tuberculosis	10	6	2	3	3	2	3	4	1	2	6	3	4	2	2	1	3	4	3.1	3	2
percy217	tuberculosis	10	6	2	4	3	4	3	4	3	2	24	3	4	2	2	6	3	4	2.7	3	2
percy29b	tuberculosis	10	6	2	4	3	4	3	4	3	1	25	3	4	2	2	5	3	1	2.7	3	2
percy240	tuberculosis	10	6	2	4	3	4	3	4	1	2	4	2	4	2	2	5	3	3	2.7	5	2
H37Rv	tuberculosis	10	6	2	4	3.3	1	3	4	2	2	3	3	4	2	3	6	3	3	2.7	5	2
percy221	tuberculosis	10	5	2	4	3	1	4	4	2	3	8	2	4	1	1	6	1	4	1.1	2	2
percy241	tuberculosis	10	5	2	4	3	3	4	4	2	3	6	2	4	2	2	6	3	3	1.7	1	2
percy236	tuberculosis	7	5	1	4	3	6	3	4	3	3	23	2	4	1	2	6	3	2	1.7	1	2
percy244	tuberculosis	10	5	1	4	3	4	2	4	3	3	8	2	4	1	2	6	3	2	1.7	3	2

Allele sizes were converted to number of repeats according to the correspondence indicated in Table 1. In some instances, decimal values are used, reflecting the existence of alleles with intermediate size. The markers are named and listed according to their position on the genome (Table 1). The strains are listed according to their position in the clustering analysis (Figure 1). *M. tuberculosis* CDC1551 and *M. bovis* AF2122/97 are included based on the predicted allele sizes (Table 1) with the exception of locus H37Rv_3690 (disagreement between observed and expected size for H37Rv at this locus).

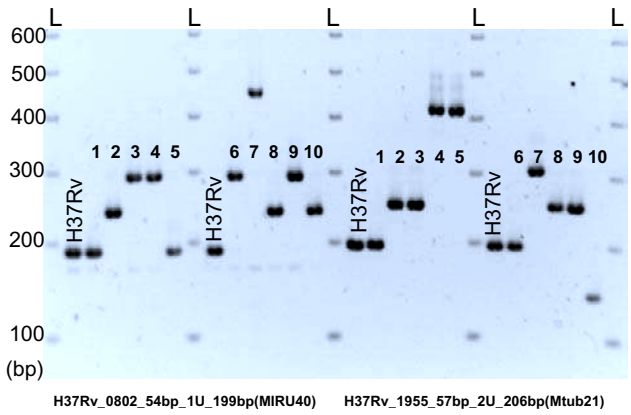


Figure 3
Set-up of the genotyping on agarose gels . The figure illustrates the usual setup for the running of pcr products on agarose gels. Twelve DNA samples (including two "H37Rv" control lanes) are typed at two loci. A 100 bp ladder size marker lane (L) flanks both sides of each group of 6 PCR products. The experiment shown is part of a reproducibility test. The ten blinded-coded samples are numbered from one to ten (percy59, percy55, percy40, percy189a, percy122, percy33, percy28b, percy33b, percy31, percy53). The number of units is easily deduced from the pattern observed, the largest alleles contain six copies of the repeat unit.

cubated at 95°C for 30 minutes. The tube was then centrifuged and the supernatant was recovered.

Identification of tandem repeats

The tandem repeats database described in [1] and accessible at [http://minisatellites.u-psud.fr] was used to identify tandem repeats with a predicted size which differs between the two strains H37Rv [24] and CDC1551 [19]. The database uses the Tandem Repeat Finder software [25] [http://tandem.biomath.mssm.edu/trf.html] to identify tandem repeats in bacterial genomes. Predicted PCR products size in *M. bovis* AF2122/97 was deduced using the *M. bovis* blast server at [http://www.sanger.ac.uk/Projects/M_bovis/blast_server.shtml].

Minisatellite PCR amplification and genotyping

PCR reactions were performed in 15 µl containing approximately 1 ng of DNA (2 µl of the thermolysate), 1× PCR buffer, 1 unit of Taq DNA polymerase, 200 µM of each dNTP, 0.3 µM of each flanking primer. The Taq DNA polymerase was obtained from Qbiogen and used as recommended by the manufacturer.

PCR reactions were run on a MJResearch PTC200 thermocycler. An initial denaturation at 94°C for five minutes was followed by 40 cycles of denaturation at 94°C for 1 minute, annealing at 62°C for one minute (except for H37Rv_0079 and H37Rv_2387 : annealing temperature 55°C), elongation at 72°C for 90 seconds, followed by a final extension step of 10 minutes at 72°C. Five microliters of the PCR products were run on standard 2% agarose gel (Qbiogen) in 0.5 × TBE buffer at a voltage of 10 V/cm (10× TBE is 890 mM Tris base, 890 mM boric acid, 20 mM EDTA, pH 8.3). Samples were manipulated and dispensed (including gel loading) with multi-channel electronic pipettes (Biohit) in order to reduce the risk of errors. Gel length of 20 cm were used. Gels were stained with ethidium bromide, visualized under UV light, and photographed.

Allele sizes were estimated using a 100 bp ladder (MBI Fermentas or Biorad) as size marker. Each 50 wells gel contained 8 regularly spaced size-marker lanes. In addition, strain H37Rv was included as a control for size assignments (one H37Rv control for each set of five DNA samples; see Figure 3). Gel images and resulting data were managed using the Bionumerics software package (version 3.0, Applied-Maths, Belgium).

Data analysis and on-line access

Band size estimates were exported from Bionumerics and converted to number of units. The resulting data was imported in Bionumerics as an opened character data set. Clustering analysis of genotyping data was performed using the Bionumerics package (categorical and Ward). The use of the categorical coefficient implies that the character states are considered as unordered. The same weight is given to a large vs. a small number of differences in the number of repeats at a locus. Among the many possibilities available for clustering analysis, the categorical and Ward combination were empirically selected for their ability to cluster the strains in almost perfect agreement with the microbiological analysis (Figure 1).

The web-page site running identifications was developed using the BNserver application (version 3.0, Applied-Maths, Belgium).

Authors' contributions

PLF has compiled and evaluated previously described markers, evaluated new markers, and genotyped the strains. FD has analyzed the H37Rv, CDC1551 and AF2122/97 sequence data to identify tandem repeats, and is the curator of the tandem repeat database [http://minisatellites.u-psud.fr] in which known data on individual markers is available. FD and GV have designed and set-up the internet strain identification service. GV conceived the study and participated in its design and coordination. MF and JLK have isolated and characterized the strains at the biochemical level, and also prepared PCR-quality DNA.

All authors contributed to the writing of the paper and approved the final manuscript.

Acknowledgements

We thank Drs V. Hervé (HIA Percy) and R. Teyssou (HIA Val de Grâce) for their support to this project. The setting up of a database for the identification of human pathogens is supported by grants from the Délégation Générale de l'Armement (DGA/DSA/SP-Num). The sequence data for *M. bovis* AF2122/97 was produced by the *M. bovis* Sequencing Group at the Sanger Institute and can be obtained from [ftp://ftp.sanger.ac.uk/pub/pathogens/mb]. We thank Dr V. Vincent, Institut Pasteur, Paris, for the provision of two *M. africanum* strains and four *M. tuberculosis* strains of the Beijing type.

References

1. Le Fleche P, Hauck Y, Onteniente L, Prieur A, Denoeud F, Ramisse V, Sylvestre P, Benson G, Ramisse F, Vergnaud G: **A tandem repeats database for bacterial genomes: application to the genotyping of *Yersinia pestis* and *Bacillus anthracis*.** *BMC Microbiol* 2001, **1:**2
2. Bayliss CD, Field D, Moxon ER: **The simple sequence contingency loci of *Haemophilus influenzae* and *Neisseria meningitidis*.** *J Clin Invest* 2001, **107:**657-666
3. Hermans PW, van Soolingen D, Bik EM, de Haas PE, Dale JW, van Embden JD: **Insertion element IS987 from *Mycobacterium bovis* BCG is located in a hot-spot integration region for insertion elements in *Mycobacterium tuberculosis* complex strains.** *Infect Immun* 1991, **59:**2695-2705
4. van Embden JD, van Gorkom T, Kremer K, Jansen R, van Der Zeijst BA, Schouls LM: **Genetic variation and evolutionary origin of the direct repeat locus of *Mycobacterium tuberculosis* complex bacteria.** *J Bacteriol* 2000, **182:**2393-2401
5. Sola C, Filliol I, Gutierrez MC, Mokrousov I, Vincent V, Rastogi N: **Spoligotype database of *Mycobacterium tuberculosis*: biogeographic distribution of shared types and epidemiologic and phylogenetic perspectives.** *Emerg Infect Dis* 2001, **7:**390-396
6. Kremer K, van Soolingen D, Frothingham R, Haas WH, Hermans PW, Martin C, Palittapongarnpim P, Plikaytis BB, Riley LW, Yakrus MA, *et al*: **Comparison of methods based on different molecular epidemiological markers for typing of *Mycobacterium tuberculosis* complex strains: interlaboratory study of discriminatory power and reproducibility.** *J Clin Microbiol* 1999, **37:**2607-2618
7. Frothingham R, Meeker-O'Connell WA: **Genetic diversity in the *Mycobacterium tuberculosis* complex based on variable numbers of tandem DNA repeats.** *Microbiology* 1998, **144:**1189-1196
8. Supply P, Mazars E, Lesjean S, Vincent V, Gicquel B, Locht C: **Variable human minisatellite-like regions in the *Mycobacterium tuberculosis* genome.** *Mol Microbiol* 2000, **36:**762-771
9. Roring S, Scott A, Brittain D, Walker I, Hewinson G, Neill S, Skuce R: **Development of variable-number tandem repeat typing of *Mycobacterium bovis*: comparison of results with those obtained by using existing exact tandem repeats and spoligotyping.** *J Clin Microbiol* 2002, **40:**2126-2133
10. Mazars E, Lesjean S, Banuls AL, Gilbert M, Vincent VV, Gicquel B, Tibayrenc M, Locht C, Supply P: **High-resolution minisatellite-based typing as a portable approach to global analysis of *Mycobacterium tuberculosis* molecular epidemiology.** *Proc Natl Acad Sci U S A* 2001, **98:**1901-1906
11. Skuce RA, McCorry TP, McCarroll JF, Roring SM, Scott AN, Brittain D, Hughes SL, Hewinson RG, Neill SD: **Discrimination of *Mycobacterium tuberculosis* complex bacteria using novel VNTR-PCR targets.** *Microbiology* 2002, **148:**519-528
12. Supply P, Lesjean S, Savine E, Kremer K, van Soolingen D, Locht C: **Automated high-throughput genotyping for study of global epidemiology of *Mycobacterium tuberculosis* based on mycobacterial interspersed repetitive units.** *J Clin Microbiol* 2001, **39:**3563-3571
13. Ross BC, Raios K, Jackson K, Dwyer B: **Molecular cloning of a highly repeated DNA element from *Mycobacterium tuberculosis* and its use as an epidemiological tool.** *J Clin Microbiol* 1992, **30:**942-946
14. van Soolingen D, Qian L, de Haas PE, Douglas JT, Traore H, Portaels F, Qing HZ, Enkhsaikan D, Nymadawa P, van Embden JD: **Predominance of a single genotype of *Mycobacterium tuberculosis* in countries of east Asia.** *J Clin Microbiol* 1995, **33:**3234-3238
15. Collins CH, Yates MD, Grange JM: **Subdivision of *Mycobacterium tuberculosis* into five variants for epidemiological purposes: methods and nomenclature.** *J Hyg (Lond)* 1982, **89:**235-242
16. Haas WH, Bretzel G, Amthor B, Schilke K, Krommes G, Rusch-Gerdes S, Sticht-Groh V, Bremer HJ: **Comparison of DNA fingerprint patterns of isolates of *Mycobacterium africanum* from east and west Africa.** *J Clin Microbiol* 1997, **35:**663-666
17. Frothingham R, Strickland PL, Bretzel G, Ramaswamy S, Musser JM, Williams DL: **Phenotypic and genotypic characterization of *Mycobacterium africanum* isolates from West Africa.** *J Clin Microbiol* 1999, **37:**1921-1926
18. Viana-Niero C, Gutierrez C, Sola C, Filliol I, Boulahbal F, Vincent V, Rastogi N: **Genetic diversity of *Mycobacterium africanum* clinical isolates based on IS6110-restriction fragment length polymorphism analysis, spoligotyping, and variable number of tandem DNA repeats.** *J Clin Microbiol* 2001, **39:**57-65
19. Fleischmann RD, Alland D, Eisen JA, Carpenter L, White O, Peterson J, DeBoy R, Dodson R, Gwinn M, Haft D, *et al*: **Whole-Genome Comparison of *Mycobacterium tuberculosis* Clinical and Laboratory Strains.** *J Bacteriol* 2002, **184:**5479-5490
20. Sampson SL, Lukey P, Warren RM, van Helden PD, Richardson M, Everett MJ: **Expression, characterization and subcellular localization of the *Mycobacterium tuberculosis* PPE gene Rv1917c.** *Tuberculosis (Edinb)* 2001, **81:**305-317
21. David HL, Jahan MT, Jumin A, Grandry J, Lehmann EH: **Numerical taxonomy of *Mycobacterium africanum*.** *Int J Syst Bacteriol* 1978, **28:**467-472
22. Brosch R, Gordon SV, Marmiesse M, Brodin P, Buchrieser C, Eiglmeier K, Garnier T, Gutierrez C, Hewinson G, Kremer K, *et al*: **A new evolutionary scenario for the *Mycobacterium tuberculosis* complex.** *Proc Natl Acad Sci U S A* 2002, **99:**3684-3689
23. Levy-Frebault VV, Portaels F: **Proposed minimal standards for the genus *Mycobacterium* and for description of new slowly growing *Mycobacterium* species.** *Int J Syst Bacteriol* 1992, **42:**315-323
24. Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, Gordon SV, Eiglmeier K, Gas S, Barry CE, *et al*: **Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence.** *Nature* 1998, **393:**537-544
25. Benson G: **Tandem repeats finder: a program to analyze DNA sequences.** *Nucleic Acids Res* 1999, **27:**573-580

2.2.3 Conclusions

La base de données des répétitions en tandem et les fonctionnalités associées (comparaison de souches, identification de souches…) sont un ensemble d'outils utiles à l'épidémiologie bactérienne. Dans le laboratoire, ces outils ont été (ou sont actuellement) mis à profit afin d'identifier des marqueurs polymorphes dans un nombre d'espèces bactériennes croissant. Outre *Yersinia pestis*, *Bacillus anthracis*, et *Mycobacterium tuberculosis*, le typage de répétitions en tandem a été achevé ou est encore en cours pour les organismes suivants:

- *Pseudomonas aeruginosa* (Onteniente 2003) : cette espèce a la particularité d'être extrêmement riche en répétitions en tandem mais très peu se sont révélées polymorphes.
- *Legionella pneumophila* (Pourcel 2003) : peu de marqueurs polymorphes ont pu être exploités dans cette espèce car les souches sont très divergées (difficulté à développer des amorces PCR « universelles » dans l'espèce).
- *Staphylococcus aureus* (article en préparation) : pour cette espère bactérienne 6 souches sont séquencées, ce qui fait de la page de comparaison un outil très efficace pour la sélection de répétitions en tandem polymorphes.
- *Burkholderia mallei/pseudomallei* (en cours)
- *Neisseria meningitidis* (en cours)
- *Brucella melitensis/suis* (en cours)
- *Coxiella burnetii* (en cours)

Par ailleurs, une vérification régulière des connexions à la base de données révèle de nombreuses utilisations de cette base par des laboratoires extérieurs, français ou étrangers, sur différentes espèces. En particulier, nous avons comptabilisé le nombre de fichiers d'alignements consultés entre janvier et août 2003 : ce chiffre s'élève à 5884. Le chargement d'un fichier d'alignement nécessite au préalable d'effectuer une requête dans la base de données puis de cliquer sur le lien vers l'alignement : nous ne prenons donc en compte que les démarches « actives », c'est-à-dire potentiellement intéressées, des internautes. L'histogramme présenté sur la Figure 20 montre la distribution du nombre de fichiers consultés par utilisateur distant. Les 5884 fichiers ont été consultés par 505 utilisateurs. Plus de 400 utilisateurs ont consulté moins de 5 fichiers (ce qui concerne 675 fichiers d'alignements) : on peut considérer ces requêtes comme anecdotiques. Une centaine d'utilisateurs ont effectué des requêtes plus soutenues (ce qui concerne 5209 fichiers d'alignements). Ils ont chargé plus de 5 fichiers d'alignements, en une ou plusieurs fois entre janvier et août 2003 (les utilisateurs ont été considérés comme identiques lorsque leurs adresses IP étaient identiques ou ne différaient que par le numéro de machine). Un tiers environ de ces utilisateurs ont effectué leurs requêtes dans des génomes bactériens. Ces requêtes ont concerné plus d'une vingtaine d'espèces : une vingtaine de projets MLVA (multilocus VNTR analysis) sont donc susceptibles de voir le jour

grâce (en partie) à la base de données. Il sera intéressant de surveiller la littérature dans les prochains mois pour vérifier si en effet, de tels projets ont été menés à bien.

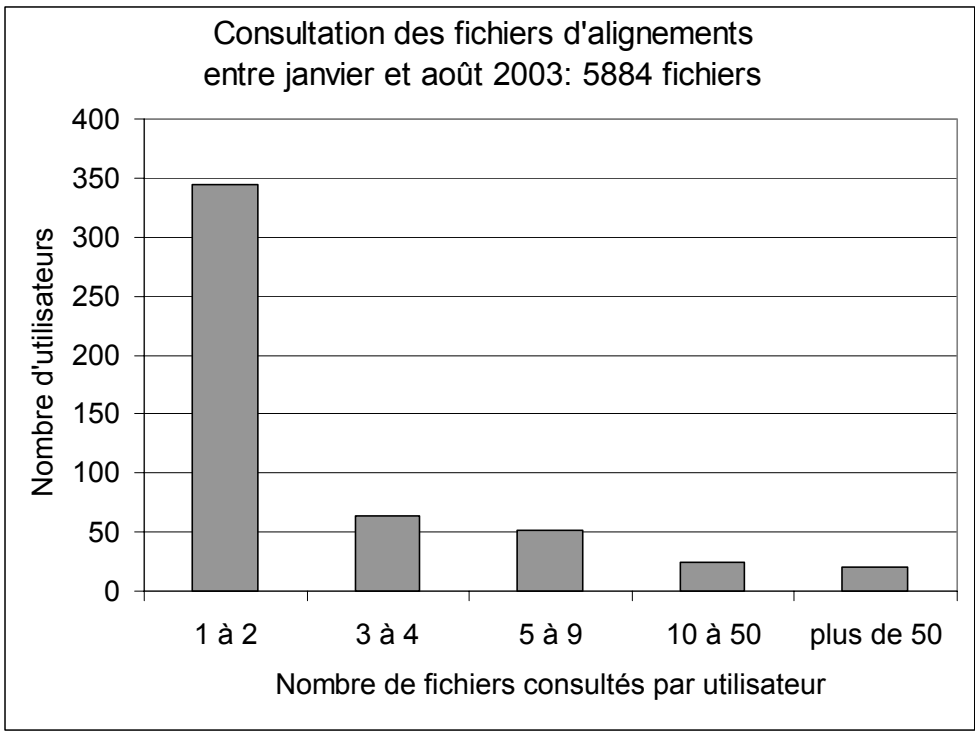


Figure 20 : Histogramme du nombre de fichiers d'alignements téléchargés dans la base de données par utilisateur, entre janvier et août 2003.

La base de données des répétitions en tandem ne contient pas uniquement des génomes de bactéries, mais également d'eucaryotes, ce qui rend possible l'étude des minisatellites à l'échelle de chromosomes eucaryotes entiers, comme nous le verrons au chapitre suivant. En particulier, deux tiers des requêtes soutenues effectuées entre janvier et août 2003 concernaient les chromosomes 20, 21 et 22 humains.

2.3 Utilisation de la base de données pour l'étude des minisatellites humains

2.3.1 Etude de la répartition des minisatellites dans des chromosomes eucaryotes entièrement séquencés

La revue suivante (Vergnaud & Denoeud 2000), intitulée « Minisatellites : mutability and genome architecture » (Minisatellites : mutabilité et architecture des génomes) présente une synthèse sur les minisatellites humains, en particulier hypermutables, ainsi qu'une étude de la répartition des minisatellites au sein de différents chromosomes eucaryotes, mettant à profit la base de données des répétitions en tandem. En effet, jusqu'alors, les minisatellites étaient identifiés par des techniques expérimentales telles que l'hybridation de sondes contenant des motifs répétés (Vergnaud 1989). Notre étude est la première à tirer parti du séquençage des génomes afin d'identifier de façon exhaustive les répétitions en tandem de chromosomes entiers.

En outre, cette étude met le doigt sur la difficulté de définir avec précision les minisatellites au sein de l'ensemble des répétitions en tandem. Nous y proposons une définition plus vaste que celle précédemment imposée par les contraintes expérimentales (Southern Blot) qui n'identifiaient que des minisatellites de longueur totale de l'ordre du kilobase, ou supérieure. Ces travaux antérieurs ont montré que cette catégorie de répétitions en tandem présente, chez l'homme, un biais de distribution télomérique très marqué (Amarger 1998). Il s'avère que des répétitions en tandem de quelques centaines de paires de bases tombent également dans la catégorie des répétitions en tandem plus souvent associées aux extrémités télomériques. Le nombre de minisatellites du génome humain peut alors être estimé à plusieurs dizaines de milliers, soit dix fois plus qu'initialement supposé.

Ainsi, il semble peu réaliste de mesurer le polymorphisme (par typages PCR, ou Southern Blot) de tous les minisatellites du génome humain, dans le but, par exemple, d'identifier des minisatellites hypermutables qui sont en premier lieu très polymorphes. Cependant, les données de séquences pourraient être mises à profit afin de trouver des critères prédictifs sur le polymorphisme, qui permettraient alors de se cantonner à un nombre raisonnable de candidats à typer. Un article (Denoeud 2003) présentant une étude sur la prédiction du polymorphisme des minisatellites, menée sur les chromosomes 21 et 22 humains, sera présenté au chapitre suivant.

Résumé :

Il a été montré que les minisatellites sont associés à des aspects importants de la biologie du génome humain tels que la régulation de gènes, les sites fragiles ou les phénomènes d'empreinte parentale. Notre connaissance de la biologie des minisatellites a fortement augmenté ces dix dernières années, grâce à l'identification et l'analyse détaillée des minisatellites hypermutables humains, aux modèles expérimentaux dans la levure, et aux études *in vitro* menées récemment sur les propriétés de recombinaison des minisatellites. En parallèle, les minisatellites ont été mis en avant comme des biomarqueurs potentiels pour mesurer l'effet d'agents génotoxiques comme les radiations ionisantes. Dans cette revue, nous synthétisons et discutons des observations récentes sur les minisatellites. De plus, nous mettons à profit la parution récente de séquences de chromosomes entiers pour proposer une vue unificatrice qui devrait faciliter l'annotation des répétitions en tandem.

Minisatellites: Mutability and Genome Architecture

Gilles Vergnaud [1,2,3] and France Denoeud[1]

[1]Institut de Génétique et Microbiologie, Université Paris Sud 91405 Orsay, France; [2]Centre d'Etudes du Bouchet, 91710 Vert le Petit, France

Minisatellites have been found in association with important features of human genome biology such as gene regulation, chromosomal fragile sites, and imprinting. Our knowledge of minisatellite biology has greatly increased in the past 10 years owing to the identification and careful analysis of human hypermutable minisatellites, experimental models in yeast, and recent in vitro studies of minisatellite recombination properties. In parallel, minisatellites have been put forward as potential biomarkers for the monitoring of genotoxic agents such as ionizing radiation. We summarize and discuss recent observations on minisatellites. In addition we take advantage of recent whole chromosome sequence data releases to provide a unifying view which may facilitate the annotation of tandem repeat sequences.

Classic Definition and Early Applications of Minisatellites

Minisatellites are usually defined as the repetition in tandem of a short (6- to 100-bp) motif spanning 0.5 kb to several kilobases. Although the first examples described 20 years ago were of human origin, (Wyman and White 1980), similar DNA structures have been found in many organisms including bacteria. Comparisons of the repeat units in classical minisatellites led early on to the notion of consensus or core sequences, which exhibit some similarities with the χ sequence of λ phage (GCTGTGG). In general, the majority of classical minisatellites are GC rich, with a strong strand asymmetry.

Because of their length polymorphism, which results from variations in the number of repeats, and the ability of some of these arrays to cross-hybridize with tens of other similar loci throughout the genome, minisatellites have opened the way to DNA fingerprinting for individual identification (Jeffreys et al. 1985). Minisatellites also provided the first highly polymorphic, multiallelic markers for linkage studies (Nakamura et al. 1987). The usefulness of polymorphic minisatellites (also called VNTRs for variable number of tandem repeats) in the early stages of human genome mapping is reflected in the Centre d'Etude du Polymorphisme Humain/National Institutes of Health consortium linkage maps (National Institutes of Health/Centre d'Etude du Polymorphisme Humain collaborative mapping group 1992).

In parallel, tandem repeats belonging to the minisatellite class were found to be associated with many interesting features of human genome biology and evolution, usually revealed by pathologies of genetic origin. In brief, minisatellites are thought to contribute to genome function in one of three ways: (1) Some are part of an open reading frame, which may or may not display polymorphism in the human population (for review, see Bois and Jeffreys 1999). (2) Some bind proteins with a variety of functional consequences, strongly suspected or still very hypothetical. Minisatellites located in the 5′ region of genes participate in the regulation of transcription (Kennedy et al. 1995). Others located within introns interfere with splicing (Turri et al. 1995). Minisatellites at imprinted loci are thought to play a role in the imprint control (Chaillet et al. 1995; Neumann et al. 1995). More speculatively, minisatellites have been proposed as intermediates in chromosome pairing initiation in some eukaryote genomes (Ashley 1994; Sybenga 1999), which might be related to their proposed recombinogenic properties (Boan et al. 1998; Wahls and Moore 1998). (3) Finally, minisatellites may constitute chromosome fragile sites (for review, see Sutherland et al. 1998) and have been found in the vicinity of a number of recurrent translocation breakpoints and in the switch recombination site in immunoglobulin heavy chain genes (Brusco et al. 1999). These aspects of minisatellite biology have been reviewed elsewhere and will not be further discussed in this article.

Novel Insights and Applications in Minisatellite Biology

Although the high degree of length polymorphism among minisatellites indicates that they are fast-evolving sequences, most of them are in fact quite stable, and neomutated alleles have been observed only at a few loci. Recent research has focused on identifying these rare hypermutable loci in human and

[3]Corresponding author.
E-MAIL Gilles.Vergnaud@igmors.u-psud.fr; FAX 33-1-69-15-66-78.

10:899–907 ©2000 by Cold Spring Harbor Laboratory Press ISSN 1088-9051/00 $5.00; www.genome.org

other genomes because they seem the most appropriate models to illustrate the mechanisms of minisatellite variability. Newly mutated alleles at human hypermutable minisatellites have been characterized in detail, leading to the current model of minisatellite mutation initiation by double-strand breaks (DSBs), and a number of attempts have been made to transfer human minisatellite instability into a more tractable system. We will present and discuss the work done on the subclass of minisatellites that are hypermutable in meiosis.

We will also show how investigations on the sensitivity of minisatellites to some genotoxic agents might provide new insight on minisatellite mutation processes. This work may lead to new applications for minisatellite sequences, such as the development of genotoxicity assays to specifically detect agents interfering with DNA recombination or replication.

Finally, the release of whole genome sequence data allows new approaches to minisatellite characterization. In spite of the fact that our understanding of minisatellite biology has improved very significantly in the last 10 years, minisatellites are usually not discussed or even annotated in releases of new sequence data. This is likely due to the lack of a clear and satisfying definition of these structures. We will briefly review the history of minisatellite characterization and chromosomal localization and compare the picture that these earlier investigations produced to the global view provided now by the sequencing of the human genome.

Insights from the Study of Mutant Alleles at Human Hypermutable Minisatellites

For practical reasons linked to the size of available pedigrees, a minisatellite will usually be classified as hypermutable if its average mutation rate in the germline is higher than 0.5% (the ratio of mutation events in the male and female germline is variable; it can be highly skewed toward paternal events as in CEB1, or equal as in MS1, see Table 1). As a rough estimate, approximately 300 human minisatellites have been typed across families (Armarger et al. 1998; Armour et al. 1990; Nakamura et al. 1987) and less than ten of these qualify as hypermutable (Table 1). The structural features of hypermutable minisatellites described in Table 1 are not specific for this subclass of tandem repeats, and the proportion of telomeric versus interstitial loci (MS32 and MS1 being interstitial) in this collection fits with the proportion of telomeric and interstitial loci among human minisatellites in general (see below).

All hypermutable minisatellites characterized so far possess internal variants, which have provided one way to undertake mutant allele analysis (Table 1). Jeffreys and colleagues developed a polymerase chain reaction-based assay (Jeffreys et al. 1991) which has proved very efficient at typing the position of variants along alleles. These internal maps can be used to identify the origin of additional repeats in mutant alleles as compared to their progenitors. An important part of our current knowledge of hypermutable minisatellite biology comes from this technology. Two reports in which neomutated alleles at the CEB1 and MS32 hypermutable minisatellites were typed (Buard and Vergnaud 1994; Jeffreys et al. 1994) pointed to DSBs as initiating events of the meiotic mutations. Both interallelic (gene conversion-like) and intra-allelic exchanges were observed, with a different proportion of the two classes of events at the two loci. The detailed typing achieved by the CEB1 study provided data showing that some of the interallelic insertions are flanked by duplicated motifs from the recipient allele. Figure 1 illustrates a model which fits with our current knowledge of meiotic DSBs within hotspots (i.e., in yeast they occur outside the tandem array (Debrauwère et al. 1999) and are almost blunt) while being compatible with observations on CEB1 in the human context.

Subsequent studies have investigated the role of the flanking sequence in the mutation process. This interest in flanking sequences was prompted by the observation that a point mutation very close to one end of the MS32 array was associated with a strongly reduced mutation rate of the corresponding allele (Monckton et al. 1994). In addition, meiotic mutation events in MS32 strongly clustered toward one end of the array (Jeffreys et al. 1994). In contrast, somatic mutations at MS32 (Jeffreys and Neumann 1997) do not show clustering toward one end; they occur at a much lower frequency and are simple intra-allelic events, predominantly deletions.

Experimental Models of Minisatellite Mutation

The development of experimental models to study minisatellite mutation processes has been necessary in order to analyze more precisely the timing of the mutation processes, the underlying genetics, and test the predictions made by the current models.

Attempts to develop animal models based on the identification of naturally occurring hypermutable minisatellites failed. Two hypermutable tandem repeats characterized in mice are the amplification of short (respectively, 4- and 5-bp) units, which do not fully qualify as minisatellites and are not amenable to variant typing (for review, see Bois and Jeffreys, 1999). For this reason, Jeffreys and colleagues developed a transgenic mouse model. The inserts injected were either an MS32 tandem array with only a few hundred base pairs of flanking sequence or a complete cosmid insert from the MS32 or CEB1 locus. Interestingly, although the mitotic instability was transferred, no mei-

Table 1. Known Human Hypermutable Minisatellites

Minisatellite name	Locus or D-number	Location	Accession number	Consensus length	Nb of variant positions	Percent match (%M)	GC% (bias)	Purine % (bias)	Instability	Proportion of paternal events
CEB1	D2S90	2q37.3 (T)	AF048727	39	8	90	72 (0.53)	68 (0.38)	6.7%	97%
CEB15	D1S172	1p36.33 (T)	AL096805	18	8	91	68 (0.33)	77 (0.56)	1.5%	100%
CEB25	D10S180	10q26.3 (T)	AL096806	52	>20	82	53 (0.73)	67 (0.36)	2.5%	65%
CEB36	D10S473	10q26.3 (T)	AL096810	42	18	93	64 (0.65)	66 (0.32)	1.8%	50%
CEB42	D8S358	8q24.3 (T)	AL096807	41	4	95	63 (0.11)	53 (0.08)	0.5%	100%
CEB72	D17S888	17q25 (T)	AL096808	21	0	100	71 (0.60)	80 (0.62)	1.8%	65%
MS1	D1S7	1p33–35 (I)	X14856	9	3	85	58 (0.93)	79 (0.60)	5%	50%
MS32	D1S8	1q42 (I)	AF048729	29	3	92	61 (0.37)	67 (0.35)	0.4%	50%
B6.7		20q13.3 (T)	AF081787	34	6	91	60 (0.5)	68 (0.37)	5.5%	64%
CSTB	EPM1	21q22.3 (T)	U46692	12	1	96	100 (0.5)	100 (0.5)	47%	76%

Minisatellites showing instability higher than 1% in the male or female germline are listed (with the exception of MS32, with a lower mutation rate, but which is a reference minisatellite in many investigations). The instability values indicated are most often average values, as measured usually in the large Centre d'Etudes du Polymorphisme Humain families (http://www.cephb.fr) with the exception of the cystatin B (CSTB) gene minisatellite. In this last case, the values given were measured in pathogenic, expanded alleles (Larson et al. 1999). Similarly, the mutation rate at CEB1 alleles has been shown to vary between <0.02% (at smaller alleles) and >20% (Buard et al. 1998). Percent match (%M) is the average similarity of motifs with the consensus motif. GC bias is the absolute value of (G% − C%)/(G% + C%). Purine bias is the absolute value of (pur% − pyr%).

otic instability was observed (Bois et al. 1997; Buard et al. 2000). However, in these investigations the integration site was random, and no attempts were made to target potentially more active loci of the mouse genome.

Alternative approaches have used yeast. The work in yeast was pioneered by Rannug, Cederberg, and colleagues, who showed meiotic induction of human minisatellite MS32 instability. The minisatellite was inserted in the vicinity of the LEU2 yeast hotspot for recombination initiation, where DSBs frequently form (Appelgren et al. 1997). Tetrad analysis demonstrated that interallelic mutants, which might look like bona fide crossover events (exchange of flanking markers, no complex secondary rearrangements), are in fact conversion events (Appelgren et al. 1999), which is of some importance when interpreting similar human data (Jeffreys et al. 1998). In a more recent investigation, the similar meiotic instability of a CEB1 allele introduced in yeast was shown to be dependent on the integration site (Debrauwère et al. 1999). Integration in a cold spot for recombination initiation resulted in a very low meiotic instability compared to integration adjacent to the ARG4 recombination hotspot. At this site, the tandem array did not modify the DSB hotspot: DSBs remained detectable on both sides of the array at a frequency comparable to the wild-type situation. Suppression of the DSBs, either by a failure in activating the site, as obtained in a rad50 deficient strain, or by the absence of the topoisomerase (Spo11) responsible for the DSBs (Bergerat et al. 1997), reduced the meiotic instability of the minisatellite to the mitotic level. Finally, taking advantage of mismatch repair-deficient strains, the predicted heteroduplex intermediates (Fig. 1) have been observed in some (but not all) mutant alleles.

These observations, combined with the attempts to develop a mouse model and the data in humans, very strongly suggest that the production of an experimental model in which the minisatellite shows meiotic

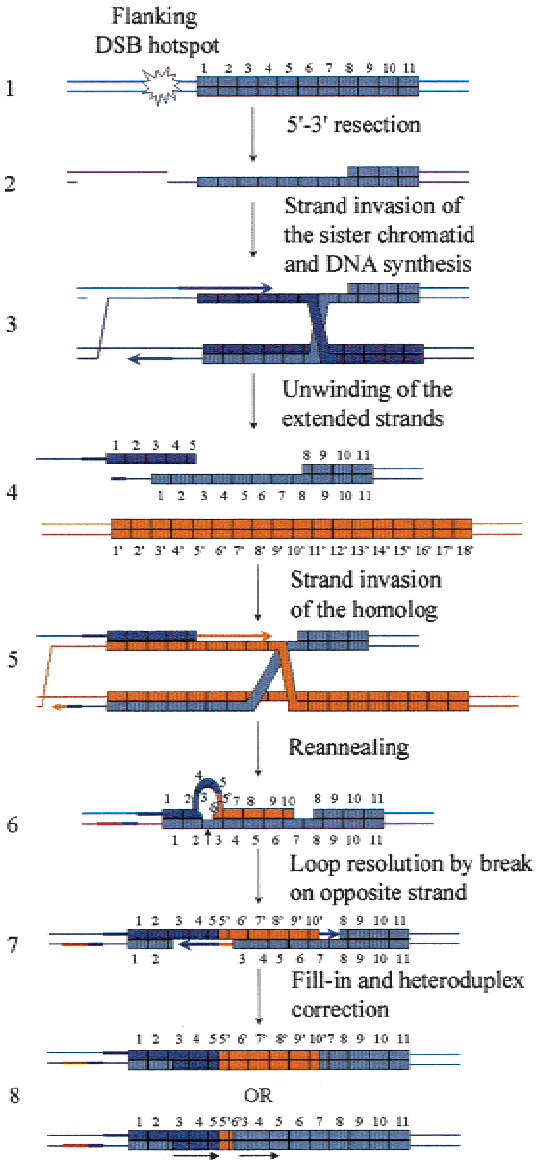


Figure 1 Revised model for meiotic mutation events demonstrating the formation of interallelic events with duplications flanking the converted motifs. In order to explain the observed duplication flanking meiotic interallelic events, the model initially proposed by Buard and Vergnaud (1994) and subsequently adopted by others (Bois and Jeffreys 1999) invoked DSBs initiated within the array by staggered single-strand breaks separated by 80 nucleotides or more. This would require a strong associated helicase activity and does not fit with the view now provided by the yeast work (Debrauwère et al. 1999). Alternatively, the presented model adapted from Debrauwère et al. (1999) shows how an almost blunt DSB, produced in a flanking DSB hotspot outside the minisatellite (step 1), can produce interallelic exchanges with a duplication flanking the converted motifs, as well as most, if not all, minisatellite rearrangements observed in man or yeast. After 5′-3′ resection (step 2), the repair is initiated by invading the sister chromatid and priming DNA synthesis on one or both (as suggested here, step 3) strands. After DNA synthesis, the newly synthesized strands independently unwind (step 4) and are free to engage in other DNA–DNA interactions (here, both strands are shown invading the homolog, step 5). Eventually, the newly synthesized strands reanneal together with properly aligned flanking sequences. A loop may form on one (as shown here, step 6) or both (Debrauwère et al. 1999) strands. This loop can be converted into the corrected allele (Debrauwère 2000) via a single strand cut on the opposing DNA strand (step 7), or removed. Depending on which strand is used to correct the heteroduplex, a direct duplication of repeats flanking the converted patch is produced (step 8, bottom). All models proposed so far predict the existence of patches of heteroduplex intermediates produced by the reannealing of similar, but different, minisatellite motifs (here, step 6). This last prediction was successfully tested in Debrauwère et al. (1999). An interesting aspect of this model is that the lower strand may extend in the flanking sequence at steps 2 and 4. This will produce a heteroduplex region in the flanking sequence (steps 6 and 7; left unrepaired here in step 8) which, once repaired, may introduce a conversion patch in the final product flanking sequence.

instability depends on the coincidence of a tandem repeat with a DSB hotspot. Conversely, minisatellites can be made unstable in mitosis in yeast strains deficient for some aspects of DNA replication (Kokoska et al. 1998), so the yeast model has already provided experimental support for the view that at least two mechanisms promote minisatellite instability (Jeffreys and Neumann 1997). Accordingly, it can be anticipated that agents interfering with one of these mechanisms will induce minisatellite instability.

Genotoxicity

A number of studies indicate that hypermutable minisatellites might provide biomarkers for exposure to some genotoxic agents. One such class of genotoxic agents is ionizing radiation. The first hint of the sensitivity of minisatellites to ionizing radiation was obtained in mouse (Dubrova et al. 1993; Dubrova et al. 1998). This was followed by studies in humans exposed to chronic low doses of radiation based on populations living in regions contaminated by the release of radioactive material after the explosion at the Chernobyl power station in 1986. The investigation used Southern blotting to genotype father-mother-child trios at hypermutable minisatellite loci and to count the frequency of mutant alleles in a control and an exposed population (Dubrova et al. 1997). The data obtained indicated that the frequency of mutant alleles in the exposed population was twice the frequency observed in the control population (from the United Kingdom). Importantly, the exposed population was split into two parts according to the degree of soil contamination in regions from which families were collected, suggesting a dose-effect relationship. The results are in contrast with the Hiroshima-Nagasaki survivors investigations (Satoh and Kodaira 1996), but the situation in the Hiroshima-Nagasaki study is very different because children were conceived years after parental exposure. At this time, minisatellite mutation rate in the germline should be back to normal, if data obtained with the mouse model can be extrapolated to human (Dubrova et al. 1998).

Several chemicals released in the environment are also suspected of inducing meiotic minisatellite mutations. Germline mutation rate monitored by DNA fingerprinting was twice as high in herring gulls inhabiting a heavily industrialized area as compared to birds living in rural sites (Yauk and Quinn 1996). Similarly, instability of the human minisatellite MS32 introduced in yeast also appears to be modulated by some chemicals (Appelgren et al. 1999).

Taking Advantage of the Global View Provided by Large-Scale Sequencing

During the 1980s and early 1990s, a number of approaches were developed to detect and/or clone minisatellite loci. Because DNA fingerprinting, using so-called multilocus minisatellite probes, previously demonstrated the property of some tandem arrays to cross-hybridize with a number of others, the majority of these approaches was based on cross-hybridization detection (Vergnaud 1989). Given the technology which was used, i.e., Southern blotting, a minisatellite would be defined as a tandem repeat with allele length usually in the range that can be assayed by Southern blots, i.e., above approximately 800 bp.

The overall frequency of such minisatellites in five mammalian genomes investigated at a significant scale is similar (Amarger et al. 1998; Bois et al. 1998; Georges et al. 1991). The distribution is however very different, with a high bias toward chromosome ends in human and a much lower bias in mouse and rat. The situation in the pig is intermediate, and a closer look at the synteny relationships suggests that, in a common ancestor, the interstitial minisatellite clusters were telomeric (Amarger et al. 1998). One conclusion of these investigations is that the tandem repeats which can be analyzed on a Southern blot are predominantly associated with chromosome ends, and internal clusters of such tandem repeats are very likely to be the result of secondary rearrangements such as chromosome ends fusion.

However, analyses limited to the usual definition of minisatellites (>800 bp) are not altogether satisfying because this definition represents only a fraction of tandem repeats, many of which are smaller than the 500-bp arbitrary limit, but do not fit in the microsatellite class of tandem repeats. Furthermore, this definition has a limited value when dealing with sequence data for at least two reasons: 1) tandem repeats which clearly qualify as minisatellites often have some alleles in the human population which are much shorter than the 500-bp limit, and 2) during the assembly of raw sequence data, the true allele length of minisatellites is not always correctly inferred, especially when the internal array is very homogeneous. CEB1, the most hypermutable minisatellite characterized so far (Vergnaud et al. 1991), illustrates both of these drawbacks of the current definition: (1) Small alleles with 5 repeat units (total array length: 200 bp) have been described, and their meiotic mutation rate is still high at 0.4% (Buard et al. 1998). (2) The cosmid from which CEB1 was originally isolated has been sequenced (Murray et al. 1999). Although the CEB1 allele present in this cosmid is 3.6 kb long, as estimated by restriction enzyme analysis, the deposited cosmid sequence contains only six CEB1 motifs spanning 240 bp, presumably because of difficulties encountered in sequencing the array.

The release of whole chromosome sequence data for a number of eukaryotes including human, the nematode *Caenorhabditis elegans*, and the plant *Arabidopsis thaliana* now opens the way to more systematic,

sequence-based investigations. For this purpose, we have constructed a prototype tandem repeat database (http://minisatellites.u-psud.fr) using the Tandem Repeats Finder software (Benson 1999) to identify the repeats. The database contains more than 14,000 tandem repeats for the acrocentric chr22 (34.6 Mb) (Dunham et al. 1999) and can be queried according to a number of features (see legend, Fig. 2). One-third of human chr22 tandem repeats (Fig. 2A) satisfies an enlarged definition of minisatellites, as used in this review (at least three units, repeat unit longer than 6 bp). Among them, minisatellites with repeat units longer than 16 bp and total length greater than 100 bp display a distribution strongly biased toward the chr22 long-arm telomere (Fig. 2B).

Figure 2C shows the result of a query mimicking characteristics of classical minisatellites, i.e., query "B" plus a high GC content, strong strand bias, and strong

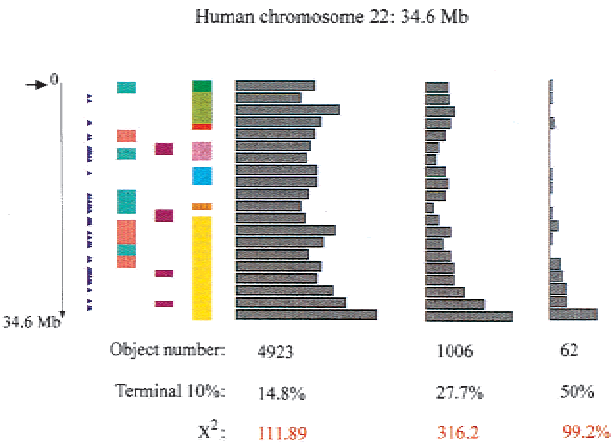


Figure 2 Distribution of tandem repeats corresponding to different queries along human chromosome 22. Tandem repeats have been identified within the human chromosome 22 sequence using the Tandem Repeats Finder (TRF) software with the following options: alignment parameters = (2,3,5), minimum alignment score to report repeat = 50, maximum period size = 500. Redundancy was then eliminated, and Alu and satellite sequences (152 were identified) were filtered. The arrow (top left) shows the centromere position. The position of the 51 chr22 Genethon microsatellites present in the database is shown with arrow heads. GC-rich (pink) or -poor (green) areas, regions of increased recombination, and known mouse synteny correspondence are as indicated in Dunham et al. (1999) Distributions obtained with different queries: (*A, Left*) U > = 6, N > = 3 (*B, Middle*) U > 16, N > = 3, L > 100 (*C, Right*) U > 16, N > = 3, %GC > = 65%, BGC > = 0.3, %M > = 85% (U = unit length, N = copy number, L = total length, %GC = GC percent, BGC = G/C bias = |%G-%C|/(%G + %C), %M (percent matches) is the average similarity of each motif with the consensus motif). Percentages reported correspond to the proportion of objects in the last 10% of total length. χ^2 values were calculated by comparing the last 10% of the chromosome with the mean number of objects along the whole chromosome. χ^2 threshold of significance (homogeneity hypothesis is rejected if χ^2 is greater than threshold) is 3.841 with P = 5%, and 10.827 with P = 0.1% (1 degree of freedom).

internal homogeneity (see legend, Fig. 2, for details). Half of the 62 minisatellites fitting this query are located within the terminal 10% of chr22. Such simple queries demonstrate that a fraction of minisatellites, comprising hundreds of loci on chromosome 22 alone, do behave as initially suggested by the subset of classical minisatellites, i.e., are present at a much higher frequency in the terminal R band of human chromosomes (Amarger et al. 1998).

Chromosome ends appear to be relatively poor in recombination nodules during human male meiosis, which is surprising given the very high male recombination rates observed toward chromosome ends. Subtelomeric minisatellites are one class of sequences that have been put forward as candidates to help explain this paradox (Ashley 1994; Sybenga 1999). Specific mechanisms would be activated in male meiosis, and minisatellites would be involved in chromosome pairing, either directly or via interactions with pairing proteins. This predicts that minisatellites should not display subtelomeric clustering in plants, where no such discrepancy between recombination nodules and rates is observed. Figure 3 presents comparisons of the three species using *C.elegans* chromosome 1 (12.75 Mb) and *A. thaliana* chromosome 4 (17.8 Mb) (The *C. elegans* Sequencing Consortium 1998; Mayer et al. 1999). The total number of tandem repeats found with Tandem Repeats Finder in the three species is not proportional to chromosome length (Fig. 3). It is significantly higher in the nematode (637 Mb) when compared to man and *A. thaliana* (415 Mb and 445 Mb, respectively).

The result of a representative query is shown in Figure 3, bottom row. The number of positive minisatellites is similar in the three species, taking into account chromosome size difference. A strong telomeric bias is observed for *C. elegans* chr1, (right panel) reminiscent of the situation in human chr22. In contrast, the distribution of minisatellites in *A. thaliana* (middle) is strikingly different from that of the two other genomes: tandem repeats are mainly located around the centromere. Figure 4A plots, for each species, the ratio of telomeric versus nontelomeric tandem repeats according to repeat unit length. *C. elegans* chr1 demonstrates telomeric bias for both short units (in particular, 6- and 12-bp units, due to the presence of many (TTAGGC)n telomere-like tandem arrays (The *C. elegans* Sequencing Consortium 1998) and longer units (above approximately 18 bp). Human chr22 demonstrates telomeric bias for repeat units above 17 bp. It may be worth noting that in yeast, 16 bp is the threshold above which mismatch repair mechanisms are unable to correct DNA loops (Sia et al. 1997). Figure 4B plots the same measure of telomeric bias according to the overall array length. In contrast with *C. elegans*, the telomeric bias for human chr22 appears only for arrays longer than 120–140 bp. This threshold is reminiscent

Figure 3 Comparison of tandem repeats distribution in three species. See legend, Fig. 2, for information on database construction and other details. Arrows show centromere position (unknown for nematode). The Tandem Repeats Finder software identifies tandem repeats at a frequency of 415 per Mb for human chromosome 22, 445 per Mb for *Arabidopsis thaliana* chromosome 4, and 637 per Mb for *Caenorhabditis elegans* chromosome 1 ("No query" panel). Bottom panel: the query applied was U > = 10, N >= 3, L > 100, 0.3 > = BGC > = 0.55, %M > = 70. Chromosomes were fragmented in areas of comparable length: 1.73 Mb (20 areas), 1.78 Mb (10 areas), and 1.82 Mb (7 areas), for human chromosome 22, plant chromosome 4 and nematode chromosome 1, respectively. For human chromosome 22 and nematode chromosome 1, significant telomeric biases are observed. In contrast, the plant chromosome shows a bias toward the centromeric region.

of triplet repeat instability observed above 40–50 repeats. No telomeric bias is observed in *A. thaliana* chr4.

Concluding Remarks and Perspectives

Previously, the number of classical minisatellites has been estimated to be a few thousand in the human genome, which translates to a few tens on chromosome 22. Such rare objects would not likely play a significant role in genome metabolism. The view now provided by the availability of whole human chromosome sequence reveals a much larger number of small minisatellites with repeat units similar to the classical structures and a similarly biased distribution toward chromosome ends, which is not observed in *A. thaliana*. These observations give much more credibility to these structures (Boan et al. 1998; Wahls and Moore 1998). Obviously, comparisons with additional, larger human chromosomes will be of some interest.

It is tempting to speculate that the meiotic hypermutability of some minisatellite structures is the by-product of the coincidence of an ordinary minisatellite with a DSB hotspot (Debrauwère et al. 1999). The disappearance of a hotspot, as proposed by Boulton et al. (1997) will then remove the hypermutability of the neighboring tandem repeat. In this model, the study of hypermutable minisatellites is demonstrating more about human DSB hotspots, the majority of which would exist independently of neighboring tandem repeats in human (Badge et al. 2000) as in yeast, than about minisatellites in general. The model presented in Figure 1 shows how a double strand break occurring outside of the array (as suggested by Debrauwère et al. 1999) can indeed produce the complex interallelic events observed in man, including duplications flanking the converted patch. The model also accommodates conversion patches in the flanking sequence, which may include mosaics of intra- and interallelic origin. In contrast, the making of minisatellites in general would result from replication mechanisms, favored by deficiencies in enzymes involved in replication such as *Saccharomyces cerevisiae* Rad27 as proposed in Tishkoff et al. (1997). In the process, sequence features of the motif, likely to produce secondary structures or slow down the polymerase on the lagging strand during replication (G-rich DNA strands, palindromic motifs in AT rich minisatellites, GC richness), may be important.

In this regard, no information regarding minisatellite instability or even polymorphism is obtained using the tandem repeat database presented here. This will be an important further step of the database development, which might benefit from the current knowledge of variant motif interspersion patterns along hypermutable minisatellite alleles. In addition, tandem repeat polymorphism predictions will be facilitated by the expected availability, in the near future, of sequence data from more than one allele.

Genotoxicity is a promising domain for minisatellite-related investigation. It may combine short-term applications toward the development of genotoxicity assays specifically identifying recombinogenic activities with more basic investigations into the purpose of minisatellites and what triggers them. One question raised by these investigations is whether the tandem array itself is the target of the genotoxic agent, whether it is the flanking DSB hotspot which is further activated by the agent, or whether it is the replication machinery which is affected. In the second hypothesis, hypermutable minisatellites would act as markers for the activity of their flanking recombination hotspot, whereas in the first (and perhaps also third) hypothesis, any minisatellite could act as a biomarker for the genotoxic agent. Recently developed yeast models may help address such issues.

ACKNOWLEDGMENTS

We thank Christine Pourcel for comments and critical reading

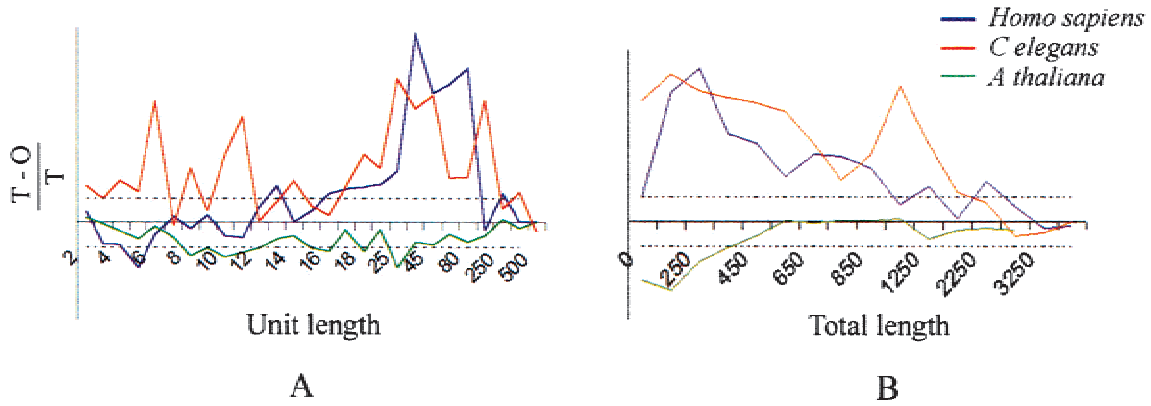


Figure 4 Comparison between terminal and other regions according to unit length (*A*) or total length (*B*). On Y-axis: {number of objects in the terminal 10% of the sequence [T] − number of object in other regions [mean for 10%] [O])/O (corresponds to a Z-score). If >0, terminal 10% are richer than the rest of the genome; else they are poorer, with a significance threshold of 1.96 (dotted lines). On X-axis: unit length (*A*) or total array length (*B*).

of this work. Current minisatellite work in the laboratory is supported by a grant from Délégation Générale de l'Armement (DGA/DSP/STTC).

REFERENCES

Amarger, V., Gauguier, D., Yerle, M., Apiou, F., Pinton, P., Giraudeau, F., Monfouilloux, S., Lathrop, M., Dutrillaux, B., Buard, J. et al. 1998. Analysis of the human, pig, and rat genomes supports a universal telomeric origin of minisatellite sequences. *Genomics* **52:** 62–71.

Appelgren, H., Cederberg, H., and Rannug, U. 1997. Mutations at the human minisatellite MS32 integrated in yeast occur with high frequency in meiosis and involve complex recombination events. *Mol. Gen. Genet.* **256:** 7–17.

Appelgren, H., Cederberg, H., and Rannug, U. 1999. Meiotic interallelic conversion at the human minisatellite MS32 in yeast triggers recombination in several chromatids. *Gene* **239:** 29–38.

Appelgren, H., Hedenskog, M., Sandstrom, C., Cederberg, H., and Rannug, U. 1999. Polychlorinated biphenyls induce meiotic length mutations at the human minisatellite MS32 in yeast. *Environ. Mol. Mutagen* **34:** 285–290.

Armour, J.A.L., Povey, S., Jeremiah, S., and Jeffreys, A.J. 1990. Systematic cloning of human minisatellites from ordered array charomid libraries. *Genomics* **8:** 501–512.

Ashley, T. 1994. Mammalian meiotic recombination: a reexamination. *Hum. Genet.* **94:** 587–593.

Badge, R.M., Yardley, J., Jeffreys, A.J., and Armour, J.A. 2000. Crossover breakpoint mapping identifies a subtelomeric hotspot for male meiotic recombination. *Hum. Mol. Genet.* **9:** 1239–1244.

Benson, G. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27:** 573–580.

Bergerat, A., de Massy, B., Gadelle, D., Varoutas, P.-C., Nicolas, A., and P. Forterre. 1997. An atypical topoisomerase II from archaea with implication for meiotic recombination. *Nature* **386:** 414–417.

Boan, F., Rodriguez, J.M., and Gomez-Marquez, J. 1998. A non-hypervariable human minisatellite strongly stimulates in vitro intramolecular homologous recombination. *J. Mol. Biol.* **278:** 499–505.

Bois, P., Collick, A., Brown, J., and Jeffreys, A.J. 1997. Human minisatellite MS32 (D1S8) displays somatic but not germline instability in transgenic mice. *Hum. Mol. Genet.* **6:** 1565–1571.

Bois, P. and Jeffreys, A.J. 1999. Minisatellite instability and germline mutation. *Cell Mol. Life Sci.* **55:** 1636–1648.

Bois, P., Stead, J.D., Bakshi, S., Williamson, J., Neumann, R., Moghadaszadeh, B., and Jeffreys, A.J. 1998. Isolation and characterization of mouse minisatellites. *Genomics* **50:** 317–330.

Boulton, A., Myers, R.S., and Redfield, R.J. 1997. The hotspot conversion paradox and the evolution of meiotic recombination. *Proc. Natl. Acad. Sci.* **94:** 8058–8063.

Brusco, A., Saviozzi, S., Cinque, F., Bottaro, A., and DeMarchi, M. 1999. A recurrent breakpoint in the most common deletion of the Ig heavy chain locus (del A1-GP-G2-G4-E). *J. Immunol.* **163:** 4392–4398.

Buard, J., Bourdet, A., Yardley, J., Dubrova, Y., and Jeffreys, A.J. 1998. Influences of array size and homogeneity on minisatellite mutation. *EMBO J.* **17:** 3495–3502.

Buard, J., Collick, A., Brown, J., and Jeffreys, A.J. 2000. Somatic versus germline mutation processes at minisatellite CEB1 (D2S90) in humans and transgenic mice. *Genomics* **65:** 95–103.

Buard, J. and Vergnaud, G. 1994. Complex recombination events at the hypermutable minisatellite CEB1 (D2S90). *EMBO J.* **13:** 3203–3210.

Chaillet, J.R., Bader, D.S., and Leder, P. 1995. Regulation of genomic imprinting by gametic and embryonic processes. *Genes Dev.* **9:** 1177–1187.

Debrauwère, H. 2000. Analyse des mécanismes d'instabilité des séquences répétées humaines de type minisatellite dans la levure *S. cerevisiae*. *Biologie-Sciences de la vie*. PhD thesis, University Paris **VI:** 56–61.

Debrauwère, H., Buard, J., Tessier, J., Aubert, D., Vergnaud, G., and Nicolas, A. 1999. Meiotic instability of human minisatellite CEB1 in yeast requires DNA double-strand breaks. *Nat. Genet.* **23:** 367–371.

Dubrova, Y.E., Jeffreys, A.J., and Malashenko, A.M. 1993. Mouse minisatellite mutations induced by ionizing radiation. *Nat. Genet.* **5:** 92–94.

Dubrova, Y.E., Nesterov, V.N., Krouchinsky, N.G., Ostapenko, V.A., Vergnaud, G., Giraudeau, F., Buard, J., and Jeffreys, A.J. 1997. Further evidence for elevated human minisatellite mutation rate in Belarus eight years after the Chernobyl accident. *Mut. Res.* **381:** 267–278.

Dubrova, Y.E., Plumb, M., Brown, J., Fennelly, J., Bois, P., Goodhead, D., and Jeffreys, A.J. 1998. Stage specificity, dose response, and doubling dose for mouse minisatellite germ-line mutation induced by acute radiation. *Proc. Natl. Acad. Sci.* **95:** 6251–6255.

Dunham, I., Shimizu, N., Roe, B.A., Chissoe, S., Hunt, A.R., Collins, J.E., Bruskiewich, R., Beare, D.M., Clamp, M., Smink, L.J. et al. 1999. The DNA sequence of human chromosome 22. *Nature* **402:** 489–495.

Georges, M., Gunawardana, A., Threadgill, D.W., Lathrop, M., Olsaker, I., Mishra, A., Sargeant, L.L., Schoeberlein, A., Steele, M.R., Terry, C. et al. 1991. Characterization of a set of variable number of tandem repeat markers conserved in bovidae. *Genomics* **11:** 24–32.

Jeffreys, A.J., MacLeod, A., Tamaki, K., Neil, D.L., and Monckton, D.G. 1991. Minisatellite repeat coding as a digital approach to DNA typing. *Nature* **354:** 204–209.

Jeffreys, A.J., Neil, D., and Neumann, R. 1998. Repeat instability at human minisatellites arising from meiotic recombination. *EMBO J.* **17:** 4147–4157.

Jeffreys, A.J. and Neumann, R. 1997. Somatic mutation processes at a human minisatellite. *Hum. Mol. Genet.* **6:** 129–136.

Jeffreys, A.J., Tamaki, K., MacLeod, A., Monckton, D.G., Neil, D.L., and Armour, J.A.L. 1994. Complex gene conversion events in germline mutation at human minisatellites. *Nat. Genet.* **6:** 136–145.

Jeffreys, A.J., Wilson, V., and Thein, S.L. 1985. Individual-specific 'fingerprints' of human DNA. *Nature* **316:** 76–79.

Kennedy, G.C., German, M.S., and Rutter, W.J. 1995. The minisatellite in the diabetes susceptibility locus IDDM2 regulates insulin transcription. *Nat. Genet.* **9:** 293–298.

Kokoska, R.J., Stefanovic, L., Tran, H.T., Resnick, M.A., Gordenin, D.A., and Petes, T.D. 1998. Destabilization of yeast micro- and minisatellite DNA sequences by mutations affecting a nuclease involved in Okazaki fragment processing (rad27) and DNA polymerase delta (pol3-t). *Mol. Cell Biol.* **18:** 2779–2788.

Larson, G.P., Ding, S., Lafreniere, R.G., Rouleau, G.A., and Krontiris, T.G. 1999. Instability of the EPM1 minisatellite. *Hum. Mol. Genet.* **8:** 1985–1988.

Mayer, K., Schuller, C., Wambutt, R., Murphy, G., Volckaert, G., Pohl, T., Dusterhoft, A., Stiekema, W., Entian, K.D., Terryn, N. et al. 1999. Sequence and analysis of chromosome 4 of the plant *Arabidopsis thaliana*. *Nature* **402:** 769–777.

Monckton, D.G., Neumann, R., Guram, T., Fretwell, N., Tamaki, K., MacLeod, A., and Jeffreys, A.J. 1994. Minisatellite mutation rate variation associated with a flanking DNA sequence polymorphism. *Nat. Genet.* **8:** 162–170.

Murray, J., Buard, J., Neil, D.L., Yeramian, E., Tamaki, K., Hollies, C.R., and Jeffreys, A.J. 1999. Comparative sequence analysis of human minisatellites showing meiotic repeat instability. *Genome Res.* **9:** 130–136.

Nakamura, Y., Leppert, M., O'Connell, P., Wolff, R., Holm, T., Culver, M., Martin, C., Fujimoto, E., Hoff, M., Kumlin, E. et al. 1987. Variable number of tandem repeat (VNTR) markers for human gene mapping. *Science* **235:** 1616–1622.

Neumann, B., Kubicka, P., and Barlow, D.P. 1995. Characteristics of imprinted genes. *Nat. Genet.* **9:** 12–13.

NIH/CEPH Collaborative Mapping Group. 1992. A comprehensive genetic linkage map of the human genome. *Science* **258:** 67–83.

Satoh, C. and Kodaira, M. 1996. Effects of radiation on children. *Nature* **383:** 226.

Sia, E.A., Kokoska, R.J., Dominska, M., Greenwell, P., and Petes, T.D. 1997. Microsatellite instability in yeast: dependence on repeat unit size and DNA mismatch repair genes. *Mol. Cell Biol.* **17:** 2851–2858.

Sutherland, G.R., Baker, E., and Richards, R.I. 1998. Fragile sites still breaking. *Trends Genet.* **14:** 501–506.

Sybenga, J. 1999. What makes homologous chromosomes find each other in meiosis? A review and an hypothesis. *Chromosoma* **108:** 209–219.

The *C. elegans* Sequencing Consortium. 1998. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* **282:** 2012–2018.

Tishkoff, D.X., Filosi, N., Gaida, G.M., and Kolodner, R.D. 1997. A novel mutation avoidance mechanism dependent on S. cerevisiae *RAD27* is distinct from DNA mismatch repair. *Cell* **88:** 253–263.

Turri, M.G., Cuin, K.A., and Porter, A.C. 1995. Characterisation of a novel minisatellite that provides multiple splice donor sites in an interferon-induced transcript. *Nucleic Acids Res.* **23:** 1854–1861.

Vergnaud, G. 1989. Polymers of random short oligonucleotides detect polymorphic loci in the human genome. *Nucleic Acids Res.* **17:** 7623–7630.

Vergnaud, G., Mariat, D., Apiou, F., Aurias, A., Lathrop, M., and Lauthier, V. 1991. The use of synthetic tandem repeats to isolate new VNTR loci: cloning of a human hypermutable sequence. *Genomics* **11:** 135–144.

Wahls, W.P. and Moore, P.D. 1998. Recombination hotspot activity of hypervariable minisatellite DNA requires minisatellite DNA binding proteins. *Somat. Cell Mol. Genet.* **24:** 41–51.

Wyman, A. R. and White, R. 1980. A highly polymorphic locus in human DNA. *Proc. Natl. Acad. Sci.* **77:** 6754–6758.

Yauk, C.L. and Quinn, J.S. 1996. Multilocus DNA fingerprinting reveals high rate of heritable genetic mutation in herring gulls nesting in an industrialized urban site. *Proc. Natl. Acad. Sci.* **93:** 12137–12141.

2.3.2 Prédiction du polymorphisme de minisatellites humains

L'article suivant (Denoeud 2003), intitulé « Predicting human minisatellite polymorphism » (prédiction du polymorphisme de minisatellites humains), présente une étude menée sur les chromosomes 21 et 22, visant à trouver un moyen de prédire le polymorphisme des minisatellites à partir de leur séquence : en effet, comme nous l'avons vu avec l'article précédent, le génome humain est riche en minisatellites et il serait donc d'une grande utilité de pouvoir identifier les minisatellites d'intérêt c'est-à-dire polymorphes, et potentiellement hypermutables, sans avoir recours à des typages PCR trop nombreux.

Cet article a mis en évidence deux critères corrélés au polymorphisme : le pourcentage en GC et un critère, nommé HistoryR, reflétant la « facilité » avec laquelle on peut reconstruire l'histoire des duplications successives ayant généré le minisatellite, ce qui se révèle par la présence de co-mutations dans différents motifs.

D'autre part, nous avons montré que, contrairement à la comparaison de souches bactériennes, qui est une technique efficace pour identifier des répétitions en tandem polymorphes, la comparaison entre les séquences du génome humain produites par le consortium public « Human Genome Project » et la société CELERA manque d'efficacité. Ceci peut résulter du fait que les deux séquences ne sont pas indépendantes (la société CELERA ayant utilisé les séquences publiques pour générer son assemblage) et que le séquençage des répétitions en tandem est souvent de mauvaise qualité dans la version « CELERA » : le nombre de répétitions est inférieur à celui de la plage d'allèles observés parmi les individus typés, ce qui doit correspondre à des erreurs d'assemblage. Nous recommandons donc, pour l'étude des répétitions en tandem, l'utilisation préférentielle des séquences du consortium public.

Enfin, cette étude a permis d'identifier un minisatellite hypermutable appartenant à une séquence codante prédite : il s'agit d'une protéine hypothétique similaire à la protéine « erythrocyte membrane-associated giant protein antigen 332 –Plasmodium falciparum-» (Locuslink : LOC129238 [http://www.ncbi.nlm.nih.gov/LocusLink/]). Cette similitude n'est basée que sur la présence d'une répétition en tandem de 11 acides aminés contenant la succession des acides aminés PVEE dans les deux protéines. Les parties de la protéine hypothétique situées hors du minisatellite n'ont aucune homologie avec des protéines connues. Cette prédiction nécessite d'être confirmée, et, si cette protéine existe bel et bien, il serait intéressant de mener une étude de l'influence du minisatellite hypermutable sur sa fonction.

Résumé :

Nous cherchons à définir des critères prédictifs basés sur la séquence, qui permettraient d'identifier des minisatellites polymorphes et hypermutables dans le génome humain. Le

polymorphisme d'un ensemble représentatif de minisatellites, issus des chromosomes 21 et 22 a été mesuré expérimentalement par typages PCR dans une population d'individus non-apparentés. Deux approches prédictives ont été testées. La première utilise des caractéristiques simples des séquences des répétitions en tandem (taille du motif, nombre de répétitions, biais nucléotidique…) et une mesure plus complexe, appelée HistoryR, basée sur la présence de mutations associées dans les répétitions en tandem. Nous montrons que la mesure HistoryR et le pourcentage en GC sont fortement corrélés au polymorphisme et qu'en tant que critères prédictifs, ils réduisent de moitié le nombre de répétitions à typer en augmentant la proportion de minisatellites ayant une hétérozygotie supérieure ou égale à 0.5 de 43% à 59%. La deuxième approche utilise les différences de taille entre les minisatellites de deux versions de la séquence du génome humain (provenant du consortium public et de la société CELERA). Ce prédicteur augmente de façon similaire la proportion de minisatellites polymorphes mais d'une façon moins efficace qu'attendu (un nombre trop élevé de minisatellites polymorphes est manqué). Enfin, le typage des minisatellites fortement polymorphes dans des grandes familles a permis d'identifier un nouveau minisatellite hypermutable, situé dans une séquence codante prédite. Il pourrait s'agir du premier minisatellite hypermutable humain codant.

***Remarque**:* les données supplémentaires associées à cet article figurent en Annexe 3.

Predicting Human Minisatellite Polymorphism

France Denoeud,[1,4] Gilles Vergnaud,[1,2] and Gary Benson[3]

[1]*Laboratoire GPMS, Institut de Génétique et Microbiologie, Université Paris-Sud, 91405 Orsay cedex, France,* [2]*Centre d'Etudes du Bouchet, 91710 Vert le Petit, France, and* [3]*Department of Biomathematical Sciences, Mount Sinai School of Medicine, New York, New York 10029, USA*

We seek to define sequence-based predictive criteria to identify polymorphic and hypermutable minisatellites in the human genome. Polymorphism of a representative pool of minisatellites, selected from human chromosomes 21 and 22, was experimentally measured by PCR typing in a population of unrelated individuals. Two predictive approaches were tested. One uses simple repeat characteristics (e.g., unit length, copy number, nucleotide bias) and a more complex measure, termed HistoryR, based on the presence of variant motifs in the tandem array. We find that HistoryR and percentage of GC are strongly correlated with polymorphism and, as predictive criteria, reduce by half the number of repeats to type while enriching the proportion with heterozygosity ≥0.5, from a background level of 43% to 59%. The second approach uses length differences between minisatellites in the two releases of the human genome sequence (from the public consortium and Celera). As a predictor, this similarly enriches the number of polymorphic minisatellites, but fails to identify an unexpectedly large number of these. Finally, typing of the highly polymorphic minisatellites in large families identified one new hypermutable minisatellite, located in a predicted coding sequence. This may represent the first coding human hypermutable minisatellite.

[Supplemental material is available online at www.genome.org.]

Tandem repeats represent a significant fraction of vertebrate genomes and have been classified as satellites, minisatellites, and microsatellites according to the length of the repeated unit and the overall length of the array. Minisatellites are usually defined as the tandem repeats of a short (10- to 100-bp) motif spanning several hundred to several thousand base pairs and are associated with interesting features of genome biology (for review, see Vergnaud and Denoeud 2000).

Minisatellites frequently exhibit length polymorphism, which results from variation in the number of internal copies, making them valuable genomic markers. They provided the first highly polymorphic, multiallelic markers for linkage studies (Bell et al. 1982; Nakamura et al. 1987) and were used in the early stages of human genome mapping (NIH/CEPH Collaborative Mapping Group, 1992). Chromosomal distribution of minisatellites in the human genome is highly skewed toward telomeres and ancestrally telomeric regions (Amarger et al. 1998). Highly polymorphic minisatellites are thus a good tool for detection of microdeletions in the ends of chromosomes, associated with human pathologies such as mental retardation (Giraudeau et al. 2001). Polymorphic minisatellites are also found in bacterial genomes (Le Fleche et al. 2001), in which they have proven to be a powerful tool for bacterial strain identification.

Although the abundance of polymorphic minisatellites suggests that they are fast-evolving sequences, most of them are, in fact, quite stable. New alleles that display changes in the number of tandem copies have been observed at only a few loci, called hypermutable minisatellites. Changes at these loci in the germline can be observed in the next generation, and in humans, one locus, D2S90 (CEB1), has been found to change in as many as 13% of the gametes (Vergnaud et al. 1991; Vergnaud and Denoeud, 2000). Hypermutable minisatellites may provide a potent source of information on the mechanism of minisatellite instability. In humans, this instability apparently arises at least in part through gene conversion events, during or shortly after meiosis, many of which involve interallelic transfers of information (Buard and Vergnaud 1994; Jeffreys et al. 1994; May et al. 1996; Buard et al. 1998). Similar intraallelic and interallelic recombination events are found in MS32 and CEB1 minisatellite sequences, when they are placed close to a meiotic hotspot in *Saccharomyces cerevisiae* (Appelgren et al. 1997, 1999; Debrauwère et al. 1999). Most likely, these events result from the gene conversion repair of double-strand breaks, as recent evidence indicates that meiotic recombination in mammals and yeast is initiated by the Spo11p endonuclease (Bergerat et al. 1997; Keeney et al. 1997; Baudat et al. 2000; Romanienko and Camerini-Otero 2000), which is also essential to the meiotic instability of the minisatellites introduced in yeast (Debrauwère et al. 1999). In agreement with these observations, it has been proposed that the meiotic hypermutability of some minisatellite structures is the byproduct of the coincidence of an ordinary minisatellite with a double-strand break hotspot (Vergnaud and Denoeud 2000).

Interestingly, hypermutable minisatellites might additionally provide biomarkers for low-dose exposure of the human germline to ionizing radiation (Dubrova et al. 1993, 1997; Dubrova and Plumb 2002). Unfortunately, <10 human hypermutable loci have been characterized so far, using approaches developed >10 years ago, whereas the population studies conducted to evaluate the effect of low-dose irradiation would greatly benefit from the availability of a larger panel of probes.

Given the multifaceted utility of minisatellites, determining which are polymorphic/hypermutable would seem a valuable task. Efficient tandem repeat detection software en-

[4]Corresponding author.
E-MAIL France.Denoeud@igmors.u-psud.fr; FAX 33-1-69-15-66-78.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.574403. Article published online before print in April 2003.

ables the identification of tandem repeats across entire genomes (Benson 1999; Vergnaud and Denoeud 2000), so that testing for polymorphism is all that is required. But although the polymorphism of the few dozen minisatellites usually present in a small genome can be systematically assayed at a reasonable cost (Le Fleche et al. 2001), this is not a realistic option for the human genome. There, the number of minisatellite loci is estimated in the thousands (based on the sequence of chromosome 22; Vergnaud and Denoeud 2000), the proportion of highly polymorphic minisatellites among these is not known, and previous efforts to identify hypermutable loci among minisatellites have produced only very low yields (~1% to 3% of those examined). Furthermore, sequence analysis of a few hypermutable loci has not yet revealed specific features that might facilitate their identification (Murray et al. 1999). Of need are predictive criteria that can be applied before the expensive and labor-intensive step of polymorphism typing.

Earlier attempts at polymorphism prediction for tandem repeats focused on microsatellites. Fondon III et al. (1998) identified polymorphic loci by selecting microsatellites in which the individual copies were at least 90% identical to a core pattern, but that study did not include a control group to test whether selection yielded higher polymorphism values than the background rate. Wren et al. (2000) improved polymorphic microsatellite identification by requiring perfect homogeneity of the repetitive unit. Such results are in accordance with the mutation process of microsatellites (replication slippage): They are stabilized by variant repeats (Weber 1990), the presence of which facilitates detection of slipped-strand DNA by the mismatch repair system (Strand et al. 1993; Heale and Petes 1995). In the case of minisatellites, in which internal conservation is not the rule at currently known hypermutable loci (Murray et al. 1999; Vergnaud and Denoeud 2000), such a high conservation requirement imposes too great a restriction on the set of potentially useful repeats and, as we report below, would preclude finding both highly polymorphic and hypermutable repeats.

The purpose of this report is to define inexpensive strategies to accelerate the search for highly polymorphic minisatellites. The goal has been the development of sequence-based predictive criteria for polymorphism. Results are based on the study of a representative pool of minisatellites selected from human chromosomes 21 and 22. Polymorphism for these loci was experimentally measured by typing in a population of unrelated individuals. This was followed by typing the most polymorphic loci across a number of large families to test for hypermutability. Two predictive approaches were tested. The most straightforward takes advantage of the availability of two different releases of the human genome sequence: one from the public genome sequencing project and the other from the private Celera project. The second approach uses sequence-based characteristics of the repeats—including such simple measures as unit length, copy number, degree of conservation, percentage of GC (%GC)— and a more complex measure based on the internal organization of variant motifs in the tandem array. A repeat that contains several distinct sets of nearly identical mutations exhibits prima facie evidence of multiple rounds of expansion and may be more likely to exist as multiple alleles than a repeat that contains mostly unique mutations (Fig. 1). This later measure is analyzed by using history reconstruction (Benson and Dong 1999), a type of parsimony analysis that infers how the present day sequence could have evolved from a single ancestral copy while undergoing a minimum number of point mutations interspersed with duplications.

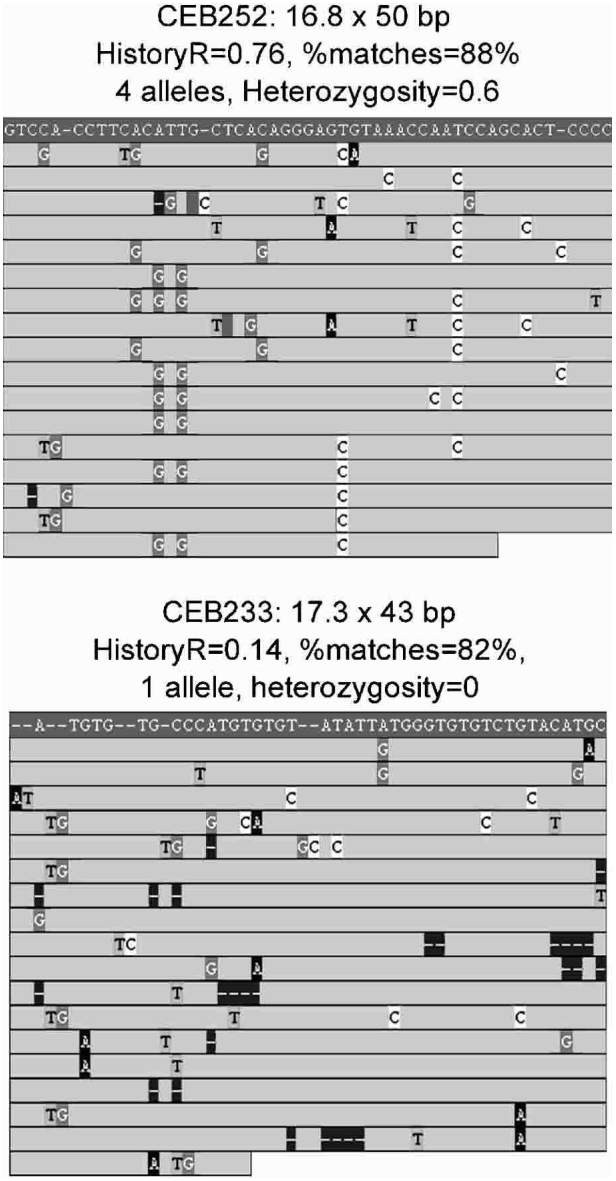


Figure 1 Multiple alignments of tandem repeats CEB252 and CEB233. In each alignment, the upper darker line is a consensus pattern for the basic unit, shown for reference, and the lighter lines are the individual copies, ordered from *top* to *bottom* as they occur in the repeat. Only differences with the consensus are shown. Heterozygosity for CEB252 is 0.6. Note several redundant patterns of mutation resulting in a high HistoryR score. Heterozygosity for CEB233 is zero. No clear organization of mutations resulting in a low HistoryR score.

RESULTS

Characterization of Chromosome 21 and 22 Minisatellites

Human chromosomes 21 and 22 contain ~15,000 tandem repeats each (as detected by tandem repeats finder [TRF] in the

publicly available sequences that exclude heterochromatin; Benson 1999). For this study, the empirical definition of minisatellites follows the suggestion made in Vergnaud and Denoeud (2000), which is more stringent than the usual definition of minisatellites mentioned in the introduction: (1) unit length ≥17 bp, (2) copy number ≥10, (3) total length ≥350 bp, (4) percent matches ≥70%, and (5) GC bias (i.e., strand asymmetry for G and C; see Methods section) ≥0.35. This definition includes repeats clearly classified as minisatellites, not microsatellites, allows minisatellites shorter than the ≥800 bp usually identified by Southern blotting (Vergnaud 1989; Amarger et al. 1998) and removes repeats with highly diverged copies. On chromosomes 21 and 22, 127 tandem repeats fulfill these criteria. Table 1 indicates their position on the chromosomes. As described before for minisatellites derived by classical approaches (Amarger et al. 1998), they are mainly located toward chromosome ends (both chromosomes are acrocentric). Analysis shows no statistically significant differences between the minisatellites from chromosome 21 and 22 for any of the characteristics listed in Supplementary Table 1. The two chromosomes will subsequently be considered together.

PCR Typing Results

Polymorphism results, that is, number of alleles observed and heterozygosity, are given in Table 1, as well as dbSNP accession numbers for the polymorphic minisatellites that were submitted to the SNP database (http://www.ncbi.nlm.nih.gov/SNP/index.html). Supplementary data about polymorphism is also available at http://minisatellites.u-psud.fr. For the minisatellites that were typed first ("training set"), the study was made on a population of 76 unrelated individuals. Results were comparable to those obtained with a subset of 28 unrelated individuals from the set of 76. Subsequent PCR typings (minisatellites from the "test set") were performed only on the 28 individuals, except for the most polymorphic loci that were typed in all 76 individuals in order to evaluate their polymorphism more accurately.

Among the 127 minisatellites, 118 were successfully amplified (55 on chromosome 21 and 63 on chromosome 22) by using the selected primer pair (Table 1). Not surprisingly, long minisatellites (>2 kb) are the most difficult to amplify: Only five among eight were successfully amplified under the conditions used. Figure 2A shows the image of the gel obtained for minisatellite CEB285 on 32 individuals (including 28 unrelated individuals): Six different alleles can be assigned. About 75% of the minisatellites successfully amplified are polymorphic (i.e., two alleles or more), and 42% have a heterozygosity value ≥0.5.

Polymorphism Prediction: Sequence Characteristics and History Reconstruction

Training Set

Twenty-five out of 60 and 32 out of 67 minisatellites were picked randomly, from chromosomes 21 and 22 respectively, to be typed first: They form the training set. PCR amplification was successful on 51 out of 57. A comparison of the sequence and polymorphism characteristics between the training set and the remaining minisatellites showed that the two sets have comparable distributions except for percentage of matches, purine/pyrimidine bias, and GC bias. To determine if some sequence characteristics are associated with high polymorphism, correlations between sequence characteristics and allele number or heterozygosity were calculated for the training set. The greatest correlations were obtained for HistoryR (a measure derived from the tandem repeats history reconstruction algorithm [Benson and Dong 1999]; see Methods section) and %GC (Fig. 3). Weaker correlations were also found for average entropy (strongly correlated with HistoryR), and unit length (data not shown). Based on these observations, we chose to test three predictive criteria: criterion 1, minisatellites with HistoryR ≥ 0.54; criterion 2, minisatellites with %GC ≥ 48%; and criterion 3, minisatellites with HistoryR ≥ 0.54 and %GC ≥ 48%.

Test Set

Of the remaining 70 minisatellites, 67 were successfully amplified and used as a test set in order to confirm the predictive criteria deduced from the training set. For each of the three criteria, the test set was partitioned into two groups: a positive group fitting the predictive criterion and a negative group. Figure 4A illustrates the results: All three criteria are predictive, that is, heterozygosity and allele number are significantly higher in the positive group compared with the negative group. The best polymorphism prediction was obtained with criterion 3 (HistoryR and %GC combined). It produces an enrichment of repeats having heterozygosity ≥0.5 from 43% (29 of 67) in the test set to 59% (19 of 32) in the positive group and a diminishment of monomorphic repeats from 25% (17 of 67) in the test set to 6% (two of 32) in the positive group. Criterion 3 thus reduces by half (67 to 32) the number of minisatellites to type while eliminating most monomorphic minisatellites and keeping most polymorphic ones (Fig. 4A). One among five highly polymorphic minisatellites (heterozygosity ≥0.85%) would have been missed using criterion 3.

Polymorphism Prediction: Direct Sequence Comparison

The experimental polymorphism values measured here indicate that greatly enhanced efficiency of polymorphic loci identification is possible if the sequences of two independent alleles for each locus are available. The reasoning is that two random samples of a moderately or highly polymorphic locus will, with high probability, yield different alleles, whereas for a monomorphic or only slightly polymorphic locus, the alleles will likely be identical. Thus, selection based on observed allele difference in the two samples should enhance the proportion of loci obtained that are polymorphic. The applicability of this approach was directly tested by comparing sequences from the Human International Genome Sequencing Consortium (HGP) and Celera genomics. We establish selection criterion 4 to be different reported lengths in these two sequences. For the 127 minisatellites previously identified in the HGP sequence, repeat sizes in the sequence provided by Celera (Venter et al. 2001) were obtained by BLAST with the PCR primers. Three tandem repeats were not found in the Celera sequence, including two that were typed (CEB230, CEB256) and one long repeat (CEB215; length expected from HGP = 2834 bp) that could not be typed. Of the remainder, 51% (29 of 57) have a different length in the two sequences for chromosome 21 and 22% (15 of 67) for chromosome 22. From the measured heterozygosity values, we would expect 37% (43 of 116) to have different lengths between the two sequences, essentially the same as found. None of these

Table 1. List of the 118 Minisatellites That Were Typed: PCR Conditions, Polymorphism Results, and Allele Size Information

IDENTIFIER INFORMATION				PCR CONDITIONS			POLYMORPHISM		ALLELE SIZE INFORMATION		
Chr	Name	dbSNP ss#	Physical position (kb)	Left primer	Right primer	Annealing temperature	Number of alleles (28ind)	Heterozygosity (28ind)	Length predicted by Human Genome Project sequence (bp)	Length predicted by Celera sequence (bp)	Observed size range (bp)
21	CEB256	6313628	220	CAACCTCCACCTCCCAGAAAAAGAAAG	CGTGCTGTGGACGTATTAACTTACTGGAAA	68°C	3	0.57	1454	?	900-1450
21	CEB255		3948	CTGGAACCCTGACAATTTTCAAGTGAGG	TTTTTGTTGCTGGAACCCTGACAATTTA	55°C	1	0	1045	682	1050
21	CEB258		4445	ACATAACAATCAAAGCAGAGCCTCACTGAC	TACATTTTCTGACTTCTGTGGTCTTCATGG	59°C	1	0	720	720	720
21	CEB260		7748	GATGTAGTTGCATTGCTTGAGTGCATTAAC	ATCGCCAAGTCATAAAGGTGTACTGTGGT	64°C	1	0	891	891	890
21	CEB261	6313629	10501	TGCAAAATCTCTCCCTCTCTGTTGATAAAA	GCTTATGTATGAGGCAGCAAATAGGATCAG	63°C	4	0.7	543	491	490-550
21	CEB263		19963	TTTTAAATTCTGATTTTCTTGCGAAGGTGA	CTTGCAGTATGAGGTCCTCTGTATCTGGTC	68°C	1	0	917	917	920
21	CEB264		21556	GCACTTTTGTCCCCATGTGTCATTCTAAC	AACACACAGAGAGCCGCAGCAGAGAC	68°C	1	0	1086	1026	1090
21	CEB265	6313630	23006	CATACAGATACGGATGATTCTTGCTCTTGG	TCTATCTTATTTGACCTCTGCCCGTAGTCC	68°C	2	0.14	830	831	830-880
21	CEB234	6313631	23639	CGAGGTGCCCCAAGGAGGGGAGGAG	GAGAGGCTGCCGTCCCCCGATTGCT	68°C	2	0.36	636	636	610-640
21	CEB266	6313632	24912	GCACCAACCAGATAGGCCACTGAGAT	ATTTCTCTGGGTATTTTTCATCTGGAAGCA	68°C	2	0.36	821	819	820-870
21	CEB268	6313633	26682	GGAAAGTGACAGCTCCCTGTTTGAATTAT	AAGACATAAGTGCTCCAGGTGTTAACAGGA	68°C	6	0.64	498	357	500-690
21	CEB269		28940	CTGGAGGCCAGAAGTTCAAAATCAAGTT	TATCCAAGTGGGCTCTAGATCCAGTGATAC	68°C	14*	0.88*	1192	1044	1600-4500
21	CEB235		29314	GCTTTCAGTGGCTCTGGAGGTTTTAGTAAG	AATAGCGAAGAGGATGTTCCCAAACAAAAT	68°C	2	0.04	813	813	1500-1850
21	CEB270	6313634	30145	CTTGAGGAGGACTGAGCCTTCAGAAGTTAG	TAAACTAGATGAAAGAGTTGCTGCGCCTGA	68°C	2	0.39	642	642	640-690
21	CEB271		30440	ACTCCTTGAGTCTTGGAGGGACTGACAC	ACTCATCTCCTGGTGGAGAAGACACTCAC	57°C	10*	0.83*	2227	2038	1100-2900
21	CEB236		30516	TGCATTTTCTTCTAGGGGAGTATGACAAGT	CTACTTGGGATGCTGAGGCAGGATTATG	68°C	1	0	899	904	900
21	CEB272	6313635	30670	AATTGTTGAGGAGACTTCATATTGCTTTCC	GACATCCAAAAGGCCAATGAGTATATGAAA	59°C	2	0.47	600	498	600-640
21	CEB273		30771	CAGCTTGGGCAATGGAGTGAGACTGTCT	ACTGCACTCCAGCCTCTCCCATCCCTA	68°C	1	0	580	580	580
21	CEB274	6313636	30829	AAAGAGCCAAGTGAAGTCCCTTCTCTGAA	CTTCTTGGGAACATCCATGGCTCAG	68°C	8	0.6	1241	1017	660-1420
21	CEB275		31147	AAACAAAGAGTCCAGGAGACGCCTGAGAGA	AATCTGCTCTTTGCCTCCATTCCTCAT	68°C	1	0	1508	1511	1510
21	CEB276	6313637	31252	GGGTTCTGGTCTGCAGTTTTTATCTGAGTT	AGGTTCAGTATTTGAAACAGCCAGATGGTA	68°C	2	0.19	838	541	840-870
21	CEB237	6313638	31420	ATGGAATCCAGAGAAGCAAGTTACACCAAT	CAGTATTCTCACCACTGCACTCAGCAAGT	68°C	2	0.29	912	690	910-960
21	CEB238		31572	GAGTAGCCACAGGACAGAACTGAGAAGC	CCTGAAGAGAAGACAAGGAGAAAGGATGAC	68°C	7*	0.79*	3109	7305	830-3100
21	CEB277		31619	TGTAATAATTCATCCACCCAGATTTGTATGC	ATCTTAAATGGGCAGGTTAATGGATTGATG	57°C	1	0	1654	1657	1660
21	CEB239	6313639	31675	GAGACAGTATCAGACACACCAGACAAAAGC	CTGTACCCGGTTAGATCCACACCCTATG	68°C	3	0.56	650	539	650-850
21	CEB278	6313640	31833	CTGAACGAATTAAGTGATGTACCCAGAAGC	GGAATGTTTAGCAGCAACCGAATATACCAG	64°C	4	0.69	1033	1000	810-1200
21	CEB240	6313641	32177	AGGTGTACCAGTACAGCAGCTTTGGACCTA	GTTTGTCTCCTCTTGCCCTCTGGAAAGTAA	68°C	3	0.51	1040	1040	900-1060
21	CEB279		32246	CTTCAGGAGGCTCTGTCTCTGGGTAGAAG	AGAGTGTGTGTGCACAACTGTCAGTGAA	68°C	1	0	761	763	760
21	CEB241	6313642	32529	AGAGCACACACCCACCACTAAAATCACTG	AGTGCATGGACTGCAGATATTGGGACT	54°C	4	0.57	2553	1241	1980-2550
21	CEB242	6313643	32690	TTCATTCCTGTGAAGCACAGCGTTT	AGAAAACAGGAGACTCACCACGATCAACT	68°C	7	0.66	737	483	650-1200
21	CEB280	6313644	33109	CTGTGAACATTTTGTAGCCATGTTGTGTTA	AAAGAAGAAAAGAAAAGCAGGGCTCATACC	68°C	2	0.07	589	590	590-670
21	CEB281	6313645	33286	GCTCAGTTCTCTCTCCTATTCGACTTGGTC	ATGAAGCTGACGCGGAAGATGGTTCT	68°C	3	0.37	635	635	600-670
21	CEB243	6313646	33317	CCCTAGGGAGGGGAGCCTAAGACCA	ATACCAGAGTCCAGCCAAGTTAGCCGTTT	68°C	3	0.27	765	765	410-770
21	CEB244	6313647	33318	CTGCTGTAACCCAGGCTCACAAACCT	ACCCTAGATGACCCTAGTGGGACCTACAC	68°C	2	0.17	643	644	640-860
21	CEB245	6313648	33383	TCTGTGGATAAACGTGAATATGCCCGAAAT	CCCAGAATCCCATTCCTGCCCAATG	68°C	2	0.04	901	901	850-900
21	CEB282	6313649	33481	AGAGGTGATGAGCACAGGTGGTGAGAG	AGCATAACACATTCTGTTCTTTGGGCATTA	57°C	3	0.23	645	647	610-710

(continued)

Table 1. (*Continued*)

21	CEB283	6313650	33711	AGAACTCTTCTGGTTTCCCGCTGCT	GAAGAGAACTTTCAGATCAGAGCACGAGTT	68°C	4	0.71	1114	1045	1020-1110
21	CEB284	6313651	33832	ACTAAAGCAGTACTGGCTCCCTCCCCTCT	AATCCTAGGTGCATTTCCGTAAACGGTGGT	68°C	3	0.55	1322	545	1300-1340
21	CEB246	6313652	33920	GATGCTGACTCACGTGGCTTCTCTGT	ATCATTTTAGATCCATGACTCCCCTTGGA	68°C	5	0.66	758	758	600-1350
21	CEB247		34282	GTCACCTTCGTGTTTTCTGCCATCAG	CCCAGGGTTATAGGACAATTTTGAACCTG	68°C	1	0	649	539	650
21	CEB285	6313653	34327	AACAGAAGCTCCTGCTGATGAATAATTTCC	GTTTAGAGAGAGGAATGGACGCAGAGTGTC	59°C	6	0.3	1092	1092	880-1090
21	CEB248	6313654	34384	CCTGTACATAGTGAAGGTGGGTTCTATTGC	GTAAACCCAACATCGAGAAAACAAGGATGG	68°C	5	0.64	1510	1514	830-1800
21	CEB286	6313655	34390	ACTTCCTCCACTCCTGGACATCGTAGTCTC	CAAGACCCAGCTGTCCTCCAGGAAT	64°C	5	0.66	713	713	560-850
21	CEB287	6313656	34454	ATCCTAACTTTCAGGGGCTTTGGTCCT	AGCTGGAAGAAAGCAGCAGGGTCAC	68°C	3	0.51	570	570	430-570
21	CEB288	6313657	34474	AATTAGAAGAGCAGTGAAACCACAGAGTCG	AAGACTAGTTTCTTTGCAGAGCACCAGAAG	68°C	3	0.55	822	394	690-820
21	CEB249	6313658	34683	TTTTTGCTTTCCGTAAGATAACAATTTTCC	AAGCGAAGAGAGTGTGTGTGACAGTGACTA	68°C	2	0.5	1305	430	1300-1350
21	CEB289		34741	ATTGCACTGTGGTTATCTGGATGTTGTTTT	AATTACTTATATTCCGGCCCATCTGTGTGT	68°C	1	0	2144	3451	5000
21	CEB290	6313659	34830	GATACTTCTCCAGCAGGGGAAACAAGAAGT	CTGAGCAGGGACAGAGGCTCTCATCT	68°C	4	0.64	833	374	670-920
21	CEB291		34854	ACAGCTCAAAGTGCGAGACAGGAACAC	GGAGGCCCCTCACAGGGAGTAGATA	68°C	10	0.87	772	775	770-1300
21	CEB250	6313660	34932	CTTTGAGGTGAGTGGTGACTCCTGCTC	CAGCAGCTAATTCTAACGTTTAAGGTCATCC	68°C	19	0.93	1045	487	540-1850
21	CEB292	6313661	34950	GGGACCCTGCATTTCCGTTTCAGGT	GAATCCCATGAGGCAGCTGAGAGAG	60°C	5	0.55	647	648	580-1100
21	CEB251	6313662	34992	GGTGACAAAGTGCCAGTCAGTGATTGTAT	AATCATCTCTGGGGAGGTGCCGTTTACATA	68°C	3	0.39	1737	?	1740-1960
21	CEB252	6313663	35084	TTTGGTGCAAGGTACAGATATCTCCTTATG	GAAAATGTAATCAAGGGACAGGAAGGAAAC	68°C	4	0.61	983	1033	980-1330
21	CEB253		35145	ACTCAGGCAGTTAGGGGTACACATCCTAT	CAGACTTAAATTCTCCCTTAAATTCACAAA	62°C	6	0.49	798	663	870-1100
21	CEB254	6313664	35165	ATAAACTAGGTTTTCTTGGAGCAGCAGGAG	AAACTTAAACGAACCGTGTTAATATGTCCA	68°C	3	0.2	656	539	630-680
22	CEB224	6313665	1528	CTAGCCTTACCCATCCCAACTAGCCTTACC	CAAGGAATCCTGACTGGTAGTGGTGCT	68°C	6	0.75	1066	1019	940-1500
22	CEB213		1571	CTACTTCCCCTGCTTAGGAGCTTAGCCATC	GACTTATGAATCAATCAAGGGCTTGTGCTG	68 °C	5	0.53	1401	1092	900-1700
22	CEB293	6313666	3522	CCCATGTATGTGTGCATTTCTCTATCCATT	CTGACACTCCATCAGTAGGATGGACACTG	64°C	3	0.35	858	858	820-910
22	CEB225	6313667	4095	TCTTCTCATTACAAAAGAAGCATGTTCAAA	AACTCCAGGAAACTGGCAAGTCAGTCAG	59°C	2	0.23	555	555	550-600
22	CEB216	6313668	10395	GTTTCCCAATGCAAGTGTTTTGGTTTATTT	GGAGACTAACAGTGGCTACGGGATGTGTTA	60°C	2	0.07	1563	1290	1300-1600
22	CEB294	6313669	11741	ACACTTACCTCCATAGTGTGTGGCTGTGT	GAGGATGACCATGGGTATAATGCACAA	68°C	3	0.5	624	628	630-800
22	CEB226		12766	AGCCCAGAGGTTCAAGGCTACGATTAG	CACCCCGGCTGTGTGTGTAACCTCT	68°C	1	0	650	652	650
22	CEB295		13055	ACCAATGTAATCACAGGGTCCTTAGAGAGG	GAGAATCCATTCCTTGCCTTTTCAACCT	68°C	9*	0.8*	805	809	1020-4700
22	CEB222		15802	AGCTTTTTCTACCACAGATACCCTCACCTG	AAGGGCCCCCAAGTCACTGGAATACAT	60°C	1	0	822	822	1200
22	CEB296	6313670	16228	GGACATGCTTGAGGGAATTTTTACTTTGTC	AGAGGCTCTCCCAGAGCACCTACAAT	68°C	5	0.52	1866	1868	1860-2200
22	CEB227		16333	CCCAAGGTCACACAGGATGTTATATTTCTT	AAAGTGAGGACAGTAACACAAGGCTTATCG	68°C	1	0	670	671	670
22	CEB297	6313671	16658	ACCTGCCTGATTCCTACATTCCTACCAC	TCAGTAATTGTTTGCTTCTCCTCTCCTTCA	60°C	2	0.24	738	738	740-880
22	CEB298		18295	ATTAAGATACAGACAAAAGCAGGGATGCTG	TAATCTTAGTTCACCCAGACATGCCCTAAG	68°C	8	0.78	569	567	1080-1300
22	CEB299	6313672	19972	CAACTCAGTCTCATTCCCACTCTGAGATTT	TCCCATTTCCTACTTTAGAAACTTTGCGATT	68°C	8	0.79	1016	1993	1020-2500
22	CEB228		20986	GGGAACAACAATTATCACCAGAGCCTAATA	CTGAAGAGTTGTGTGCAAGGATGCTCT	68°C	1	0	723	723	700-720
22	CEB300		21044	AGATGGACAGGAGAGCCAAGGCCTAAGT	GACACAGCCTCCAGGTGACCCCACT	68°C	3	0.56	968	758	1420-1670
22	CEB201	6313673	21327	ATCCCTGGTTCTGAAATCCTCAGCTTC	AAGGAGAAGGAGCCAGACAATGTGGAC	68°C	3	0.58	820	2301	320-900
22	CEB301	6313674	21541	CTCAGGCTGCCTCTACACGTGAAATC	GTTGTCTCGTTGAAAGGAAAAGGACTGTGT	60°C	3	0.33	1790	1796	1650-2090
22	CEB214		22157	GAGAGGTCAGCTATCAGGCCCATCC	GCTCCTGCCACCATGTCCATCTAAT	58°C	1	0	668	418	660
22	CEB302	6313675	22447	AAGTAAGGACTGAAAGGTCAGCATTCTTGC	CTCCTTACGAGTGGATGAGGCTCGTTTAT	68°C	2	0.04	737	739	680-740
22	CEB303	6313676	23529	GAAGCAAGAAACAGAGGGATTTAGGATTCA	CTTCTGCATCCTCTGCACCCAGCAT	60°C	2	0.13	713	713	710-770
22	CEB304		25144	GCCCTGCCCTAGATGAAGTAAGTTAGAATC	ACAGGATCTCATGAACCTGAGTCAGTGG	68°C	1	0	533	533	530
22	CEB229	6313677	26130	AGAGACGAATAAACCAGTGAGGGTAAAAGG	TGTAAAAGGACATTAGCAAACCACACGATT	68°C	2	0.5	500	500	500-520

(continued)

Table 1. (*Continued*)

22	CEB305	6313678	27086	CCACCGAAACTTAAATATTCCCCACCTATG	GTCAGCCATGAGGAGAAGGAGAATGAG	68°C	14	0.89	1557	1109	1060-2150
22	CEB306		27209	CAGGTAGATGTTCCCAAAGGTAGAACAGGT	CTAAAACGGAAGCCATTATCCAATGGTGAG	68°C	1	0	629	629	630
22	CEB307	6313679	27382	CAGCTTCAGAATCTAAACCCCTGGTTCTAA	CTGAGCAGAGCCAAGGACAATAATAGAGAC	68°C	4	0.58	528	528	430-640
22	CEB217		27398	CTCTGCAGAAAGCATCTTCCCTTCTTGA	TACAGCTTCCATCTGCGTGGTGCTTTAGAC	68°C	1	0	1527	1527	1530
22	CEB308	6313680	27654	AGTAGCCTCAGTAAATCGAGGAACTCTCCA	CTTAACTGACGCCTTGCTCTCTGTCGTAT	63°C	7*	0.78*	1191	1191	600-1200
22	CEB309		27773	CTCCAAAACCAAAAATCTCTATGACCCAAT	CAAAGTCACATGCTTTACCCTCAAACAAAG	58°C	1	0	513	513	513
22	CEB212	6313681	28490	CCCCAGCTGACCTACCTTGATCACTAT	TATAGTTGGGTTTAGGCCCACCATCTGTTA	68°C	4	0.49	1201	1203	1100-1400
22	CEB202	6313682	29067	AGAAGCGGCTCAGCAAATTACAGTGTGAAC	ACTTTTCATCCCTCGCACCAGCTCAG	60°C	19*	0.92*	907	909	630-1900
22	CEB230		30257	GTGTGTACGAGGCTGAGATCTAGGGATG	CGTGCCTCCACTGGTACTTGACACC	68°C	5	0.77	683	684	720-1200
22	CEB231	6313683	30336	GAGTGTGCACTGAACCCATTCTTTTATCAG	GTTCTGGCTTCTGAGGGTAACTGGTTATG	68°C	3	0.45	1862	1844	1620-1850
22	CEB310	6313684	30541	CCTTTTTATGGCTAAGTAGTATTCCATCGT	CGTTCAGGGAAAGAAAAACAGAGATGACTT	68°C	12	0.85	998	998	900-1550
22	CEB311	6313685	30860	GTCTGGGTCGGTTGCAGTTGACAGTAG	CCAGCTGGATAAACTCTAAAGGTCTCAGGA	68°C	2	0.46	1115	1111	1110-1180
22	CEB232		31534	CTAAACCATTTGTCACCCTCTGGATTTGTT	CATAAGTGTGGAATTGTTGGTGGTTCTGAT	68°C	1	0	572	572	500-605
22	CEB312	6313686	31865	AGACTCTGCCAGGTGGAAATTTAAGATTGG	GCCTGATATGCCAGAGGAGATGCTTAG	68°C	2	0.04	615	615	610-640
22	CEB313		32217	ATGGTCAGCAAATAAACACCCCATGTATTT	GGCTGTTATCAGATTGTAGAGCAGAGCATC	64°C	1	0	811	811	810
22	CEB314	6313687	32267	AGAAGCTTGAAAGACAAGACTGGAGTGTCC	TCTGAGCTCTTCCCAGGTATTCCACATATT	63°C	4	0.56	1037	1036	630-1040
22	CEB203		32298	AACCACCTTCACATTTGAGAACTCGCTCTG	ACACACCCAACCCATCATCCTGTCTCTAT	68°C	1	0	695	695	700
22	CEB315		32458	AAAGACCTCAGGGTGTAAGGACAGAAGAAA	TAGGCCATCCTAAAGGAAGGGACAGAG	68°C	5	0.72	1228	1245	1400-1600
22	CEB218		32693	ACCCACACAGCCTCCGTGAATTATTTAGTA	AACCTACAAGAGCACTTGGAAACCGAAG	68°C	1	0	590	590	600
22	CEB316	6313688	32741	ACAAAGAGTGGACCCTGAATCACAATGAAT	AAAAATTTCCGCTGTTTAAGCTGCCTGGAC	68°C	5	0.62	757	759	700-810
22	CEB219	6313689	32911	ACACTGAACATTGGGAAGGGCTTCTTCTAC	CCTGTGGCTTCTTCCTGTCTCAGGTAAC	68°C	3	0.51	559	559	500-700
22	CEB317		32915	ACCTTCAATATCCGTTCACCCTCTGTATCT	AAAAAGTGCTGGGAAAAGCCCTCATCT	68°C	1	0	594	594	600
22	CEB204	6313690	32948	CAGTTCCTCAAACCCCAAGTGAAGATGA	CAGGATACTAGGGCCTAAGCAGTTTGGACA	68°C	5	0.66	875	867	850-1500
22	CEB318	6313691	32980	CACTATCACAAGAACAGCACGGGAAAACCT	CACATGTGAGGCTGCATGGGAAGAAC	68°C	3	0.43	777	777	730-860
22	CEB205	6313692	33057	GCGTTTAGAAAGACAGGTGCAAGAATTAGG	AGTGGTTGAGCGTCTCCTTCTGTCTCAAT	68°C	19*	0.93*	1318	1317	550-2500
22	CEB206	6313693	33318	AGAACACAGCAGCTGAAATGCCATACC	CCTAACAAAGAAAGGTCAGAGTGAGCTGTG	68°C	2	0.44	1227	1226	820-1270
22	CEB233		33400	TGCTAGGAAATCCTTGTCTGCTGTGTATAA	CAGCACATGGCATTTGTTAGGATACACATA	68°C	1	0	856	856	850
22	CEB319		33414	ACTCTCCCTCCCATCCTCCACTCTC	CTGCTGTTGCTTCTGCTCTCAGTTCATA	68°C	1	0	572	572	570
22	CEB320	6313694	33419	CTCCTCTCAGACCCTGTCACAGAACAAAC	GCTCATGAGAAAACAGTAGATCCATCTGAG	68°C	2	0.04	691	691	690-760
22	CEB321	6313695	33434	GAAGAGACCTCATCATGGGCAGTCC	ATCCTCTACTAGCCCTGATAGCACCACATC	63°C	5	0.62	1272	1272	1240-1500
22	CEB322	6313696	33545	CCTCACTGCTCATTGGCATCAGACAT	GGTACTGCTTCTTGAGGACAGGGCTAACT	68°C	2	0.04	578	578	540-580
22	CEB220	6313697	33592	AAGACCCAAGATCTGAACCCTCAACTTCCT	GTGACTTGGCTTTCCATCTCTCCTCTGT	68°C	3	0.28	785	785	700-930
22	CEB323		33618	ATGGAGAGTGGGGGCCCTTGAGTTAG	CTTTTCTTGCAACCTTAAAGGGCATCTG	68°C	1	0	872	872	870
22	CEB207	6313698	33817	ACAACACAATAAACAAATCTGGCCCAACTC	AGGAGTTTCTTAACTGTGACAGGGGATGCT	68°C	2	0.13	2617	2663	2600-3500
22	CEB324	6313699	33825	GTGGGGCAAGAGGCATCTCCGTGAGT	CGCCCGGCAATAGGGGGGGTCTTTA	68°C	19*	0.93*	985	244	400-3500
22	CEB325	6313700	33854	GCCCCTTCCTCTCTGGTTCCATACTG	CTCTGCCTTCAGAACCCCATACCTCT	58°C	4	0.53	926	928	930-1400
22	CEB221	6313701	33864	CAAAATAATTGGAGTAGGAGTGGGTGAAGC	AAGTGGTTTTGCACCCAAATCATTAGAAGA	68°C	3	0.51	802	686	750-830
22	CEB326	6313702	33965	AAAGCAGAAAGTGCATCTGGAAAACAAAG	TAAAGATCCTTGCATGTTTTCTGAGGGATG	68°C	3	0.25	1408	802	1250-1680
22	CEB327		33983	CAGGAGGCGCGTGGGACTACACTT	GGCCGCCTTCCTCCACTCTCACCT	68°C	1	0	793	793	800
22	CEB223	6313703	34031	AATACCAGCACGATCGATTCTATCAGGACA	CCCTGTGGAGACAGTGTTTTGTGATG	68°C	2	0.5	1836	1864	1850-1950

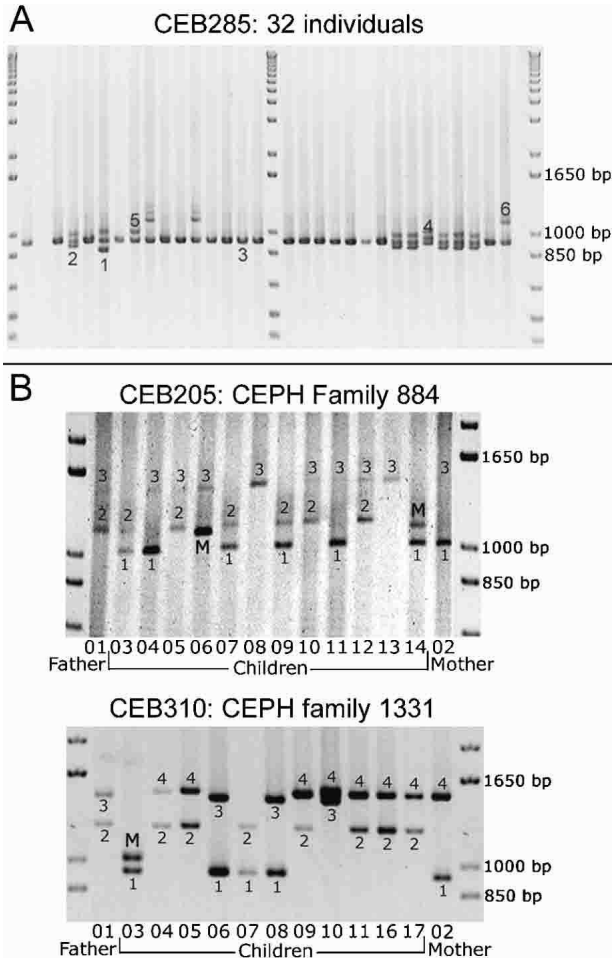


Figure 2 (*A*) Ethidium bromide–stained agarose gel showing PCR products for minisatellite CEB285. Six different alleles are scored among 32 individuals (in some cases, three bands are seen for one individual [the upper one is a PCR artifact as shown by segregation patterns in families]; this artifact occurs only in heterozygotes [data not shown], indicating a mechanism involving an interaction between the two alleles). (*B*) Image of the gels obtained for minisatellites CEB205 and CEB310 on CEPH families 884 and 1331, respectively. Two children inherit mutant alleles for CEB205, and one child inherits a mutant allele for CEB310. For CEB205, larger alleles are missed in the procedure used: The results were confirmed by Southern blot.

should be monomorphic, and ~75% (32 of 43) should have a heterozygosity value ≥0.5.

Heterozygosity and allele number are significantly higher in the positive group for criterion 4 (over the entire set of typed repeats) compared with the negative group (Fig. 4B). Criterion 4 produces an enrichment of repeats having heterozygosity ≥0.5 from 42% (49 of 118) in the whole set to 61% (25 of 41) in the positive group and a diminishment of monomorphic repeats from 25% (30 of 118) in the whole set to 12% (5 of 41) in the positive group. Criterion 4 thus reduces to nearly one third (116 to 41) the number of minisatellites to type while eliminating most monomorphic minisatellites and retaining 50% of the most polymorphic ones. By comparison, criterion 3, if applied to the entire set of typed repeats, (Fig. 4B) would reduce their number by roughly half (118 to 61), eliminating just two fewer monomorphs while retaining 69% (34 of 49) of the most polymorphic repeats. Additionally, criterion 4 eliminates half (four of eight) of the highly polymorphic (heterozygosity ≥0.85) minisatellites, whereas criterion 3 retains 75% (six of eight) of these.

We note that for some highly polymorphic minisatellites, (CEB202, CEB205, CEB310, CEB291), predicted lengths are identical in the two sequences. In addition, the results for criterion 4 are not uniform for the two chromosomes, owing to the much greater agreement on predicted loci length in chromosome 22. We presume that this reflects the fact that the Celera sequence was assembled by using both public and Celera sequence reads (Venter et al. 2001). More surprisingly, for five minisatellites, which we found to be monomorphic, predicted lengths differ (CEB214, CEB255, CEB264, CEB247, CEB289). These findings raise unresolved questions about the accuracy of the HGP and Celera sequences with regard to minisatellites. Tandem arrays can present significant sequence assembly problems, in particular when the internal array contains regions of high homology and, potentially more seriously, when the repeat exhibits length polymorphism and data are drawn from more than one individual, as was done for the Celera sequence (Venter et al. 2001).

To examine this further, we compared the HGP and Celera predictions to the alleles we detected (in Table 1, predicted lengths are underlined and not shaded when they correspond to an observed allele). In 65% (75 of 116) of the repeats, HGP and Celera predict an identical allele length, which corresponds to an observed allele length with five exceptions (Table 2) and is the most common allele in 81% of these cases. In 35% of the repeats (41 of 116), HGP and Celera predict different length alleles (Table 2). The length predicted by the HGP sequence fits with an observed allele size in 36 cases (most common allele length in 20 of these), whereas the Celera prediction fits with an observed size in 10 cases (and was once the most common allele).

Among the tandem repeats that provide PCR products unmatched by the HGP sequence, six sufficiently informative ones (CEB230, CEB253, CEB295, CEB298, CEB315, CEB269), with at least three different alleles among the four parental chromosomes, were typed in large CEPH families to check their chromosomal origin. All map to the expected area of chromosome 21 or 22, indicating that the discrepancy between sequence data and PCR product size probably results from a sequencing error (or the sequencing of a very rare allele) and not from a PCR specificity problem.

χ^2 tests were used to examine whether the similarities in prediction of the HGP and Celera findings could be explained by chance (see Methods). Differences identified by the tests had, in all cases, less than one one-thousandth probability of occurring by chance. Specifically, cases in which predictions disagreed and both allele sizes were detected were underrepresented (compared to expected frequency) in all tests, and cases in which only one or neither predicted size was detected were overrepresented in all but one test.

Identifying Hypermutable Loci

Hypermutable minisatellites are expected to belong to the class of highly polymorphic loci because they are, by definition, subject to frequent rearrangements that generate new alleles. For practical reasons linked to the size of available pedigrees, a minisatellite will usually be classified as hypermutable if its average mutation rate in the germline is >0.5%, that is, if an average of at least one or two mutant alleles is observed among 100 children.

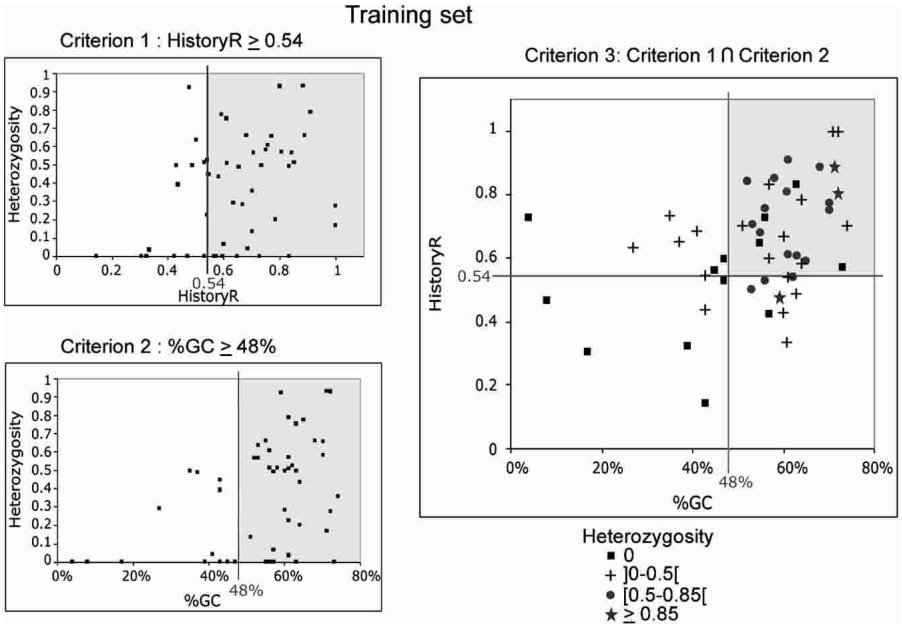


Figure 3 Criteria 1, 2, and 3 applied to the training set. For criteria 1 and 2, heterozygosity (28 individuals) versus HistoryR (criterion 1) or percentage of GC (%GC; criterion 2) are plotted. Correlations are significant at the 0.01 level. For criterion 3, HistoryR versus %GC is plotted, with different symbols representing the polymorphism. Lines represent the selected thresholds, and shaded areas contain the minisatellites selected by the criteria (criterion 1, HistoryR ≥0.54; criterion 2, %GC ≥48%; criterion 3, criteria 1 and 2 combined). Plots show that criteria select most of the polymorphic minisatellites and eliminate a majority of monomorphs or slightly polymorphic ones.

We typed the eight most polymorphic minisatellites (i.e., with heterozygosity ≥0.85) in the eight largest CEPH families (102 children) to search for mutant alleles. Comparing the results obtained by PCR and Southern blotting shows that even when some larger alleles are missing in the PCR products, the estimated heterozygosity rate (see Methods) is close to the heterozygosity rate obtained with Southern blots. This helps validate the simplified PCR-based polymorphism measurement. Among the eight minisatellites (CEB202, CEB205, CEB250, CEB310, CEB269, CEB291, CEB305, CEB324), two showed mutant alleles (CEB205 and CEB310; Fig. 2B). Both yielded two mutant alleles among 204 meioses, that is, 102 children (mutation rate, 0.12% to 3.5%; 95% confidence interval). For minisatellite CEB205, one mutation event occurred in the mother and the other in the father, whereas for CEB310, both mutations occurred in the father. The remaining six minisatellites yielded no mutant allele among 102 children (mutation rate, 0 to 1.79%; 95% confidence interval). They were not investigated further but can not be strictly excluded from being hypermutable. The two minisatellites that appeared hypermutable among 102 children were then typed in more families (32 other reference CEPH families). For CEB205, one new mutant allele was found among 352 meioses (mutation rate, 0.54%; 95% confidence interval, 0.11% to 1.57%), but no other mutant allele was detected for CEB310 among 476 additional meioses (mutation rate, 0.29%; 95% confidence interval, 0.04% to 1.06%). Based on these results, CEB205 appears to be hypermutable. It is a GC-rich minisatellite with a unit length of 33 bp repeated 10 to 70 times, located at 1.5 Mb from the end of the chromosome 22 sequence. It seems to be part of a predicted coding region (gene *LOC129238*; see http://www.ncbi.nlm.nih.gov/LocusLink/LocRpt.cgi?l=129238, 31 July 2002 update).

DISCUSSION

This study, performed on the scale of entire human chromosomes, provides a first global evaluation of minisatellite polymorphism based on genome sequence data. The repeats studied here, chosen by using a detailed definition that is more stringent than the broad definition mentioned in the Introduction, are, in majority (75%), polymorphic in the population investigated, and 42% have a heterozygosity value ≥0.5. Minisatellites from chromosomes 21 and 22 are similar in physical distribution (higher frequency toward chromosome ends), sequence features, and polymorphism. Assuming that chromosomes 21 and 22 are representative of all human chromosomes and given that the two chromosomes represent ~2% of the genome, we speculate that the entire human genome contains ~6,000 minisatellites that match our definition, including 4,800 polymorphic and 2,500 very polymorphic ones. A few 10s of these might be expected to qualify as hypermutable loci. Because our definition precluded many other potentially polymorphic minisatellites, future research should seek to expand the category of minisatellites that are tested against our polymorphism prediction criteria.

Predicting Polymorphism

We showed that using the sequence properties %GC and HistoryR effectively improves polymorphic minisatellite selection. With them, we reduce the number of minisatellites for typing by about half while increasing the frequency of repeats with heterozygosity ≥0.5% from the background rate of 43% to 59%. Internal conservation, used as a polymorphism predictor for microsatellites, is not applicable to minisatellites, presumably owing to the greater complexity of their mutation processes.

That %GC correlates with polymorphism is in agreement with earlier observations. Some of the first minisatellites to be characterized were detected via a shared 10- to 15-bp "core" sequence similar to the generalized recombination signal (χ) of *Escherichia coli* (GCTGTGG; Jeffreys et al. 1985). The majority of classical minisatellites (mostly polymorphic and/or hypermutable ones) are GC-rich, with a strong purine/pyrimidine strand asymmetry (Vergnaud and Denoeud 2000). In other genomes, though, (for instance bacterial genomes), %GC does not seem to be associated with minisatellites polymorphism (Le Fleche et al. 2001). Such a criterion may therefore not be universal, especially because GC content varies significantly across genomes.

The HistoryR criterion is based on the hypothesis that tandem repeats expand through multiple rounds of duplication, with the new copies sharing the mutations that occur before duplication, whereas unique mutations accumulate once the repeat is no longer evolving. For example (Fig. 1),

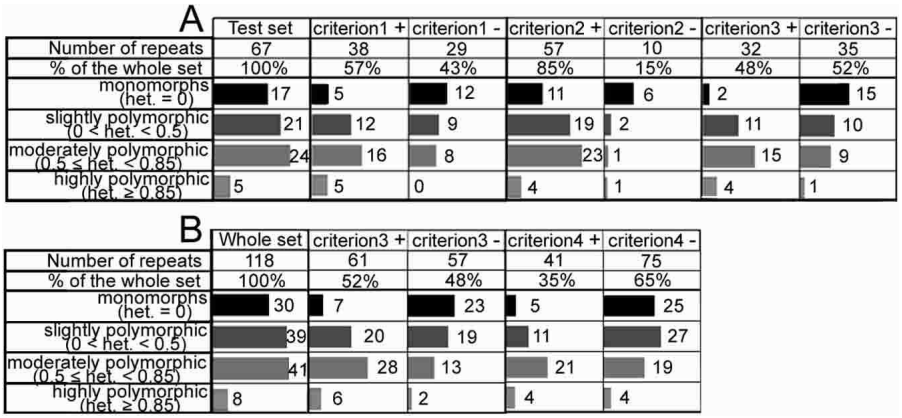


Figure 4 (*A*) Application of criteria 1 (HistoryR ≥0.54), 2 (%GC ≥48%), and 3 (HistoryR ≥0.54 and %GC ≥48%) to the test set. For each criterion, the distributions of minisatellites (from monomorphs to highly polymorphic) between positive (retained by the criterion) and negative (excluded by the criterion) sets are compared. All differences between sets + and − are statistically significant at the 0.01 level. (*B*) On the whole set, comparison of the results obtained with criterion 4 and criterion 3.

minisatellite CEB252 shows several redundant patterns of mutation, resulting in a high HistoryR score, whereas CEB233 shows no clear organization of mutations, resulting in a low HistoryR score.

This polymorphism criterion is likely to be applicable to any genome, even though the history reconstruction algorithm makes simplifying assumptions about the possible biological mechanisms involved in array expansion. These mechanisms, which include mutational events during mitotic replication and meiotic recombination, comprising both intraallelic and interallelic events, might occur independently or jointly. At present, there are no rules to predict which mechanism will occur preferentially at which locus (Maleki et al. 2002). Moreover, the individual mechanisms themselves are still poorly understood and, thus, impossible to model. Meiotic events, for instance, have been shown to result from the activity of nearby meiosis-specific double-strand break hot-spots. The nature of these sites, better known in yeast, is still unknown in the human genome (Debrauwère et al. 1999; Tamaki et al. 1999; Vergnaud and Denoeud 2000). In view of the current state of knowledge, it may be premature to hope for a perfect polymorphism predictor based on apparent array expansion.

Use of Two Human Sequences to Select for Polymorphic Loci Is Problematic

The availability of two versions of the human genome sequence provides an additional avenue to improve polymorphic minisatellite identification. However, in the repeats studied here, selection based on reported length differences discarded half the highly polymorphic minisatellites and, in particular, the hypermutable one from chromosome 22. In both chromosomes, the number of loci with different predicted lengths in the HGP and Celera sequences that were nonetheless both found was significantly underrepresented. This is apparently owing to the lack of independence resulting from sharing of data during assembly of the Celera sequence. In addition, in both chromosomes, the number of loci in which only one or no predicted allele was found is overrepresented, apparently owing to assembly errors. Because the Celera sequence—which when not in agreement with the HGP data—usually provides copy numbers unobserved in any allele, it appears that the Celera sequence, at least with respect to minisatellites, is more prone to assembly error. As a result of the lack of independence/assembly errors in the Celera/HGP data, polymorphism prediction based on sequence comparison did not perform as well as anticipated.

One New Hypermutable Locus in a Coding Region

This study revealed one hypermutable minisatellite, CEB205, showing three mutant alleles among 278 children (mutation rate, 0.54%; 95% confidence interval, 0.11% to 1.57%). Interestingly, CEB205, with a 33-bp pattern, may be part of a coding region. The corresponding putative protein is 614 amino acids long, half of which are derived from the tandem repeat (11 codon repetition) at the N terminus. Of the minisatellites studied here, 26 among 60 (43%) on chromosome 21, and 22 among 67 (33%) on chromosome 22 belong to genes (i.e., exons, introns, or UTRs), as determined by sequence similarity analyses in the human genome sequence (using BLAST and http://www.ncbi.nlm.nih.gov/genome/seq/, release of November 2001). None except CEB205 appear to contribute to the coding sequence itself. Although the proportion of tandem repeats that contribute to coding regions is important in bacterial genomes, it is relatively low in the human genome, and CEB205 might represent the first known, coding hypermutable minisatellite.

CEB310—which exhibited meiotic mutation events, but which we do not here classify as hypermutable—is unusual in that its sequence is 80% AT. It is reminiscent of the tandem

Table 2. Success of the Public Human Genome Project (HGP) and Celera Sequences in Predicting Alleles That Were Actually Found to Occur

116 tandem repeats (predicted by both sequences)					
Predicted lengths match: n = 75		Predicted lengths differ: n = 41			
Allele found with predicted length:		Allele found with predicted length:			
Yes	No	Yes, both predictions	HGP prediction only	Celera prediction only	Neither prediction
70	5	10	26	0	5

repeats studied in Giraudeau et al. (1999), that is, minisatellites made up of degenerated microsatellite-like repeated units (in this case, [AC]m[AT]n). Although most hypermutable minisatellites known to date are GC-rich, some have been described as having a very high AT content, for instance, the one constituting the chromosomal fragile site FRA16B (Yu et al. 1997; Yamauchi et al. 2000). The highly polymorphic minisatellite MSY1, from human chromosome Y, is also very AT-rich (75% to 80%; Jobling et al. 1998).

Future research will expand the systematic exploration of human tandem repeat polymorphism by testing the %GC, HistoryR, and HGP/Celera criteria on other human chromosomes as the sequences are progressively finished and released (Deloukas et al. 2001).

METHODS

Constructing the Tandem Repeats Database

Tandem repeats were identified from chromosome 21 (Hattori et al. 2000) and chromosome 22 (Dunham et al. 1999) sequences by using the TRF software (Benson 1999) with the following options: alignment parameters of (2,3,5), minimum alignment score to report repeat of 50, maximum period size of 500. When the program reported redundant (overlapping) repeats, the redundancy was eliminated in the following way. For each group of overlapping repeats, two values were determined: Lmax, the maximal total length among the redundant alignments, and Mmax, the maximal percent matches among the redundant alignments with total length ≥80% of Lmax. Then, of all the alignments in the group with total length ≥80% of Lmax and percentage of matches ≥Mmax − 0.1, the one with smallest unit length was stored in the database. The nominal length of the stored repeat is the total length of the overlapping region, that is, from the first position of the first overlapping repeat to the last position of the last overlapping repeat. Twenty-two tandem repeats showed differences of >5% between the nominal length and the length of the stored repeat, (the difference exceeded 10% in 14 cases, and 30% in three cases: CEB311, 33%; CEB320, 46%; and CEB327, 50%). For these latter three, TRF cut the repeats into two parts, which were combined for further analysis. Variation between nominal and stored size of repeats does not affect allele size prediction, which is based on length of sequence between primers. The database, publicly available at http://minisatellites.u-psud.fr, can be queried according to a number of simple features (e.g., total length, unit length, copy number, %GC) and provides links to repeat alignments and flanking sequence data as described previously (Le Fleche et al. 2001).

PCR Typing of Minisatellites

DNA was provided by Centre d'Etudes du Polymorphisme Humain (CEPH; http://www.cephb.fr/). PCRs were performed in 15 µL reactions, using 50 ng of genomic DNA, Roche long template PCR buffer (1.75 mM $MgCl_2$, 50 mM Tris-HCl at pH 9.2 and 25°C, 16 mM $[NH_4]_2[SO_4]$), 0.033 U/µL Taq polymerase (Roche), 0.003 U/µL Pwo (Roche), 200 µM of each dNTP (Amersham-Pharmacia biotech), and 0.6 µM of each flanking primer (Table 1; primers were selected within the flanking sequences provided by TRF using Primer3 software: http://www-genome.wi.mit.edu/cgi-bin/primer/primer3_www.cgi). PCRs were cycled for 5 min at 96°C, then for 15 sec at 96°C: for 20 sec at annealing temperature (Table 1; this temperature was optimized for each primer pair by using the temperature gradient provided by MJResearch PTC200), for 5 min at 68°C for 30 cycles, and for 10 min at 68°C, on Perkin Elmer 9600 thermocycler or MJResearch PTC200. Samples were run through a 13-cm-long 1% standard agarose (Qbiogen) gel in 0.5× TBE buffer at 10 V/cm for 1.5 h and visualized by ethidium bromide staining using UV (1× TBE buffer is 89 mM Tris, 89 mM boric acid, 2 mM EDTA at pH 8).

Polymorphism Measures

A population of 96 CEPH individuals (from the 40 reference families) were typed for minisatellite polymorphism. This population includes 13 mother/father/child trios and altogether comprises 76 unrelated individuals. The 76 unrelated individuals form subpopulation 1. A subset of 28 unrelated individuals forms subpopulation 2. The exact list of the 96 individuals typed is provided in Supplementary Table 2.

In this study, we examined only length polymorphism, not internal sequence variation. Two values, calculated on unrelated individuals, were used to quantify polymorphism: the number of alleles observed and the heterozygosity, calculated as $1 - \Sigma f^2$, where f are the allelic frequencies observed in the population of unrelated individuals. Heterozygosity represents the probability of having two different alleles. The simple PCR and ethidium bromide staining assay used here will usually detect only the smallest allele in individuals showing large length differences between alleles (as is often the case for highly polymorphic loci). The shorter allele often masks the longer one because it is easier to amplify. Such PCR artifacts are indicated with an asterisk in Table 1. They were detected because of the mother/father/child segregation controls and also because they do not satisfy the Hardy-Weinberg equilibrium, as tested with the HWE program, from the publicly available Linkage Utilities package (Ott 1999). For these loci, the heterozygosity value calculated from allelic frequencies was obtained by counting only one allele for individuals showing a single band (i.e., by assuming that the individual is heterozygous with one allele masked) instead of counting the same allele twice, as was done for loci in which homozygosity was not in question. The resulting heterozygosity value could be underestimated (if too many alleles are not seen), but it is sufficient to roughly evaluate the polymorphism.

Mutation Rate Estimation

Mutation rate of the most polymorphic (i.e., potentially hypermutable) minisatellites was evaluated by a combination of Southern blot hybridization and PCR typings, in recognition of the "masking" phenomenon described above. Typings were performed by using DNA from the eight largest CEPH families (F102, F884, F1331, F1332, F1347, F1362, F1413, F1416). Five µg of DNA were digested with *Alu*I (CEB202, CEB250, CEB269, CEB291) or *Hin*fI (CEB205, CEB324, CEB305; Boehringer Mannheim), electrophoresed through a 1% agarose gel and transferred to nylon membranes (Nytran+, Schleicher and Schuell) under vacuum (Pharmacia Biotech). Probes were obtained from PCR products and recovered from agarose using QIAquick gel extraction kit (Qiagen). Probes were labeled with a-[^{32}P]dCTP (Amersham Pharmacia Biotech) by the random priming procedure (Feinberg and Vogelstein 1984). Hybridization was conducted as described in Vergnaud (1989) in an hybridization oven at 65°C. After hybridization, the filters were washed in 1× SSC/0.1% SDS or 0.1× SSC/0.1% SDS at 65°C. Membranes were revealed by using a phosphoimager (Storm 860 Molecular Dynamics).

Sequence Characteristics of Repeats

The following sequence characteristics (calculated from the HGP sequence) were tested for correlation with either allele number or heterozygosity. Characteristics did not differ markedly when evaluated in the Celera sequence (in which differences with HGP typically involved deletion of adjacent copies reported in the HGP sequence):

1. Unit length: the length of the repetitive unit (consensus pattern).
2. Copy number: the number of copies of the repetitive unit.

3. Total length: the length of the entire tandem array.
4. Percent matches: the frequency at which a nucleotide at a position in one unit matches the corresponding nucleotide in the next unit (reading from left to right).
5. %GC: the percentage of nucleotides that are either G or C.
6. GC bias: strand asymmetry for G and C, |%G − %C|/(%G + %C).
7. Purine/Pyrimidine bias: strand asymmetry for purines and pyrimidines, |%Pur − %Pyr|/(%Pur + %Pyr).
8. Average entropy: from the columns of a multiple alignment of the repeat copies, the average, over all columns, of the entropy calculated from nucleotide frequencies.
9. HistoryR: described below.

HistoryR is derived from the tandem repeats history reconstruction algorithm (Benson and Dong 1999), a greedy algorithm that chooses a series of least-cost contractions to convert a multicopy tandem array into a single putative ancestral copy. Greedy algorithms are not guaranteed to find the overall least-cost solution, but testing has shown this approach to work very well on simulated sequences. Input is a multiple alignment, $\boldsymbol{M}$, of the individual copies in the repeat, with $\boldsymbol{n}$ rows (number of copies) and $\boldsymbol{k}$ columns (length of alignment). $\boldsymbol{M_{i,j}}$ represents the $\boldsymbol{i}$th row and $\boldsymbol{j}$th column of $\boldsymbol{M}$, and each $\boldsymbol{M_{i,j}}$ contains one of the alphabet symbols (A,C,G,T,–). In a contraction, two or more consecutive, equal-length subsequences (the contraction copies) are replaced by a single subsequence (the merged copy) of the same length (all subsequences selected have length equal to a multiple of $\boldsymbol{k}$). Each contraction reduces the number of rows in $\boldsymbol{M}$. If the contraction copies are identical, then one becomes the merged copy. Otherwise at every position at which the contraction copies differ, the merged copy contains the character that occurs most often, with ties being represented by an ambiguous character, that is, a set of all the most frequently occurring characters at that position. An ambiguous character created in one contraction may be converted to a single character in a subsequent contraction. This method is analogous to that used by Sankoff (1975; Sankoff and Rousseau 1975). The cost of a contraction is a ratio. The numerator is the cost of obtaining the contraction copies from the merged copy; that is, at each position of the merged copy, subtract the number of times the most frequent character occurs in the contraction copies from the total number of contraction copies, then sum all these differences. The denominator is the combined length of rows by which $\boldsymbol{M}$ is reduced, that is, the length of all contraction copies minus the length of the merged copy.

History reconstruction yields four numerical values: (1) Max, the maximum possible history cost; (2) Min, the minimum possible history cost; (3) BinaryActual, the calculated history cost when the number of contraction copies in every contraction is restricted to exactly two; and (4) ManyActual, the calculated cost when the number of contraction copies is unrestricted. Max and Min are sums of column values from the original alignment $\boldsymbol{M}$. In the case of Max, the value of a single column is the number of characters that are not the most frequent character. Max is therefore the cost if the most frequent character is ancestral and if every character different from the ancestral character was produced by its own mutation. For Min, the value is one less than the number of distinct characters in a column, that is, at most four. Min is the history cost if every distinct character different from the ancestral character arose by a single mutation (with identical characters produced by duplication).

Combinations of the four numerical values were tested for polymorphism prediction in the training set and HistoryR, which produced the highest correlation with heterozygosity, was used for the remainder of the study. It is defined as

$$History\ R = \begin{cases} (Max - BestActual)/(Max - Min) & \text{when } Max \neq Min \\ 1 & \text{otherwise} \end{cases}$$

where BestActual is the minimum of BinaryActual and ManyActual. Usually, this was BinaryActual. The HistoryR value can be thought of as the proportion of mutations that could be accounted for by duplication that actually are. When Max ≠ Min, HistoryR ≤1, with a higher ratio indicating more mutations accounted for by duplications (Fig. 1). When Max = Min, each mutation is unique, and we arbitrarily set the ratio to one. This occurred in only one repeat with a total of four mutations. The history reconstruction program is freely available for interactive use at http://tandem.biomath.mssm.edu/cgi-bin/history/history.exe.

Statistical Analysis

All statistical analysis was done with the SPSS program except for χ^2 tests which were done with StatXact 4. Correlations were determined by three methods: Pearson correlation, and nonparametric Kendall's τ_b and Spearman's ρ. Correlations are considered significant at the 0.01 level (two-tailed) of the test statistics. Group comparisons were determined by first conducting two tests of normality, Kolmogorov-Smirnov and Shapiro-Wilkinson, on the values within each group. Values were assumed to be normally distributed unless the test statistic fell within the 0.05 level of significance. If the values were normally distributed in the two groups, then a *t* test was used to compare the means, which were judged significantly different at the 0.01 level of the statistic (two-tailed). If the values were not normally distributed in either of the two groups, then a nonparametric Mann-Whitney test was used to compare the distributions, which were judged significantly different at the 0.01 level of the statistic (two-tailed).

χ^2 tests were used to analyze HGP/Celera prediction data for chromosomes 21 and 22 separately. The data were divided into three categories: (1) identical predictions/allele size detected, (2) different predictions/both alleles sizes detected, and (3) one or neither predicted allele size detected. Two estimates for frequency of unobserved alleles were used (in order to calculate the probability of alleles being detected): 10% which corresponds to the largest frequency in the population for which the chance of not appearing in our sample of 28 individuals is ≥0.05, and an arbitrary low estimate of 1%. The probability of identical predictions in the HGP and Celera sequences was obtained by summing the estimated heterozygosity values calculated separately for each locus based on observed frequencies in our sample (equivalent to using the average observed heterozygosity over all loci).

ACKNOWLEDGMENTS

We would like to thank Carol Bodian for extensive discussions and help with design of the χ^2 analysis. G.B. is supported in part by NSF grants CCR-0073081 and DBI-0090789. F.D. and G.V. are supported by grants from Délégation Générale de l'Armement.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Amarger, V., Gauguier, D., Yerle, M., Apiou, F., Pinton, P., Giraudeau, F., Monfouilloux, S., Lathrop, M., Dutrillaux, B., Buard, J., et al. 1998. Analysis of the human, pig, and rat genomes supports a universal telomeric origin of minisatellite sequences. *Genomics* **52:** 62–71.

Appelgren, H., Cederberg, H., and Rannug, U. 1997. Mutations at the human minisatellite MS32 integrated in yeast occur with high frequency in meiosis and involve complex recombination events. *Mol. Gen. Genet.* **256:** 7–17.

———. 1999. Meiotic interallelic conversion at the human minisatellite MS32 in yeast triggers recombination in several chromatids. *Gene* **239:** 29–38.

Baudat, F., Manova, K., Yuen, J.P., Jasin M., and Keeney, S. 2000. Chromosome synapsis defects and sexually dimorphic meiotic

progression in mice lacking Spo11. *Mol. Cell* **6:** 989–998.
Bell, G.I., Serby M.J., and Rutter, W.J. 1982. The highly polymorphic region near the human insulin gene is composed of simple tandemly repeating sequences. *Nature* **295:** 31–35.
Benson, G. 1999. Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Res.* **27:** 573–580.
Benson, G. and Dong, L. 1999. Reconstructing the duplication history of a tandem repeat. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 44–53.
Bergerat, A., de Massy, B., Gadelle, D., Varoutas, P.C., Nicolas, A., and Forterre, P. 1997. An atypical topoisomerase II from archaea with implication for meiotic recombination. *Nature* **386:** 414–417.
Buard, J. and Vergnaud, G. 1994. Complex recombination events at the hypermutable minisatellite CEB1 (D2S90). *EMBO J.* **13:** 3203–3210.
Buard, J., Bourdet, A., Yardley, J., Dubrova Y., and Jeffreys, A.J. 1998. Influences of array size and homogeneity on minisatellite mutation. *EMBO J.* **17:** 3495–3502.
Debrauwère, H., Buard, J., Tessier, J., Aubert, D., Vergnaud, G., and Nicolas, A. 1999. Meiotic instability of human minisatellite CEB1 in yeast requires DNA double-strand breaks. *Nat. Genet.* **23:** 367–371.
Deloukas, P., Matthews, L.H., Ashurst, J., Burton, J., Gilbert, J.G., Jones, M., Stavrides, G., Almeida, J.P., Babbage, A.K., Bagguley, C.L., et al. 2001. The DNA sequence and comparative analysis of human chromosome 20. *Nature* **414:** 865–871.
Dubrova, Y.E. and Plumb, M.A. 2002. Ionising radiation and mutation induction at mouse minisatellite loci: The story of the two generations. *Mutat. Res.* **499:** 143–150.
Dubrova, Y.E., Jeffreys, A.J., and Malashenko, A.M. 1993. Mouse minisatellite mutations induced by ionizing radiation. *Nat. Genet.* **5:** 92–94.
Dubrova, Y.E., Nesterov, V.N., Krouchinsky, N.G., Ostapenko, V.A., Vergnaud, G., Giraudeau, F., Buard J., and Jeffreys, A.J. 1997. Further evidence for elevated human minisatellite mutation rate in Belarus eight years after the Chernobyl accident. *Mut. Res.* **381:** 267–278.
Dunham, I., Shimizu, N., Roe, B.A., Chissoe, S., Hunt, A.R., Collins, J.E., Bruskiewich, R., Beare, D.M., Clamp, M., Smink, L.J., et al. 1999. The DNA sequence of human chromosome 22. *Nature* **402:** 489–495.
Feinberg, A.P. and Vogelstein, B. 1984. Addendum: a technique for radiolabeling DNA restriction endonuclease fragments to high specific activity. *Anal. Biochem.* **137:** 266–267.
Fondon III, J.W., Mele, G.M., Brezinschek, R.I., Cummings, D., Pande, A., Wren, J., O'Brien, K.M., Kupfer, K.C., Wei, M.H., Lerman, M., et al. 1998. Computerized polymorphic marker identification: Experimental validation and a predicted human polymorphism catalog. *Proc. Natl. Acad. Sci.* **95:** 7514–7519.
Giraudeau, F., Petit, E., Avet-Loiseau, H., Hauck, Y., Vergnaud, G., and Amarger, V. 1999. Finding new human minisatellite sequences in the vicinity of long CA-rich sequences. *Genome Res.* **9:** 647–653.
Giraudeau, F., Taine, L., Biancalana, V., Delobel, B., Journel, H., Moncla, A., Bonneau, D., Lacombe, D., Moraine, C., Croquette, M.F., et al. 2001. Use of a set of highly polymorphic minisatellite probes for the identification of cryptic 1p36.3 deletions in a large collection of patients with idiopathic mental retardation: Three new cases. *J. Med. Genet.* **38:** 121–125.
Hattori, M., Fujiyama, A., Taylor, T.D., Watanabe, H., Yada, T., Park, H.S., Toyoda, A., Ishii, K., Totoki, Y., Choi, D.K., et al. 2000. The DNA sequence of human chromosome 21: The chromosome 21 mapping and sequencing consortium. *Nature* **405:** 311–319.
Heale, S.M. and Petes, T.D. 1995. The stabilization of repetitive tracts of DNA by variant repeats requires a functional mismatch repair system. *Cell* **83:** 539–545.
Jeffreys, A.J., Wilson, V., and Thein, S.L. 1985. Hypervariable "minisatellite" regions in human DNA. *Nature* **314:** 67–73.
Jeffreys, A.J., Tamaki, K., MacLeod, A., Monckton, D.G., Neil, D.L., and Armour, J.A.L. 1994. Complex gene conversion events in germline mutation at human minisatellites. *Nat. Genet.* **6:** 136–145.
Jobling, M.A., Bouzekri, N., and Taylor, P.G. 1998. Hypervariable digital DNA codes for human paternal lineages: MVR-PCR at the Y-specific minisatellite, MSY1 (DYF155S1). *Hum. Mol. Genet.* **7:** 643–653.
Keeney, S., Giroux, C.N., and Kleckner, N. 1997. Meiosis-specific DNA double-strand breaks are catalyzed by Spo11, a member of a widely conserved protein family. *Cell* **88:** 375–384.
Le Fleche, P., Hauck, Y., Onteniente, L., Prieur, A., Denoeud, F., Ramisse, V., Sylvestre, P., Benson, G., Ramisse, F., and Vergnaud, G. 2001. A tandem repeats database for bacterial genomes: application to the genotyping of *Yersinia pestis* and *Bacillus anthracis*. *BMC Microbiol.* **1:** 2.
Maleki, S., Cederberg, H., and Rannug, U. 2002. The human minisatellites MS1, MS32, MS205 and CEB1 integrated into the yeast genome exhibit different degrees of mitotic instability but are all stabilised by RAD27. *Curr. Genet.* **41:** 333–341.
May, C.A., Jeffreys, A.J., and Armour, J.A.L. 1996. Mutation rate heterogeneity and the generation of allele diversity at the human minisatellite MS205 (D16S309). *Hum. Mol. Genet.* **5:** 1823–1833.
Murray, J., Buard, J., Neil, D.L., Yeramian, E., Tamaki, K., Hollies, C.R., and Jeffreys, A.J. 1999. Comparative sequence analysis of human minisatellites showing meiotic repeat instability. *Genome Res.* **9:** 130–136.
Nakamura, Y., Leppert, M., O'Connell, P., Wolff, T., Holm, T., Culver, M., Martin, C., Fujimoto, E., Hoff, M., Kumlin, E., et al. 1987. Variable number of tandem repeat (VNTR) markers for human gene mapping. *Science* **235:** 1616–1622.
NIH/CEPH collaborative mapping group. 1992. A comprehensive genetic linkage map of the human genome. *Science* **258:** 67–83.
Ott, J. 1999. *Analysis of human genetic linkage*, 3d ed. Johns Hopkins University Press, Baltimore, MD.
Romanienko, P.J. and Camerini-Otero, R.D. 2000. The mouse Spo11 gene is required for meiotic chromosome synapsis. *Mol. Cell* **6:** 975–987.
Sankoff, D. 1975. Minimal mutation trees of sequences. *J. Appl. Math.* **28:** 35–42.
Sankoff, D. and Rousseau, P. 1975. Locating the vertices of a Steiner tree in an arbitrary metric space. *Math. Programming* **9:** 240–246.
Strand, M., Prolla, T.A., Liskay, R.M., and Petes, T.D. 1993. Destabilization of tracts of simple repetitive DNA in yeast by mutations affecting DNA mismatch repair. *Nature* **365:** 274–276.
Tamaki, K., May, C.A., Dubrova, Y.E., and Jeffreys, A.J. 1999. Extremely complex repeat shuffling during germline mutation at human minisatellite B6.7. *Hum. Mol. Genet.* **8:** 879–888.
Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., et al. 2001. The sequence of the human genome. *Science* **291:** 1304–1351.
Vergnaud, G. 1989. Polymers of random short oligonucleotides detect polymorphic loci in the human genome. *Nucleic Acids Res.* **17:** 7623–7630.
Vergnaud, G. and Denoeud, F. 2000. Minisatellites: Mutability and genome architecture. *Genome Res.* **10:** 899–907.
Vergnaud, G., Mariat, D., Apiou, F., Aurias, A., Lathrop, M., and Lauthier, V. 1991. The use of synthetic tandem repeats to isolate new VNTR loci: Cloning of a human hypermutable sequence. *Genomics* **11:** 135–144.
Weber, J.L. 1990. Informativeness of human (dC-dA)n (dG-dT)n polymorphisms. *Genomics* **7:** 524–530.
Wren, J.D., Forgacs, E., Fondon III, J.W., Pertsemlidis, A., Cheng, S.Y., Gallardo, T., Williams, R.S., Shohet, R.V., Minna, J.D., and Garner, H.R. 2000. Repeat polymorphisms within gene regions: phenotypic and evolutionary implications. *Am. J. Hum. Genet.* **67:** 345–356.
Yamauchi, M., Tsuji, S., Mita, K., Saito, T., and Morimyo, M. 2000. A novel minisatellite repeat expansion identified at FRA16B in a Japanese carrier. *Genes Genet. Syst.* **75:** 149–154.
Yu, S., Mangelsdorf, M., Hewett, D., Hobson, L., Baker, E., Eyre, H.J., Lapsys, N., Le Paslier, D., Doggett, N.A., Sutherland, G.R., et al. 1997. Human chromosomal fragile site FRA16B is an amplified AT-rich minisatellite repeat. *Cell* **88:** 367–374.

WEB SITE REFERENCES

http://minisatellites.u-psud.fr; the tandem repeats database.
http://tandem.biomath.mssm.edu/cgi-bin/history/history.exe; history reconstruction program
http://www.ncbi.nlm.nih.gov/genome/seq/; Human Genome Sequencing at NCBI.
http://www-genome.wi.mit.edu/cgi-bin/primer/primer3_www.cgi; Primer3 primer picking software.
http://www.cephb.fr; Centre d'Etudes du Polymorphisme Humain.
http://www.ncbi.nlm.nih.gov/LocusLink/LocRpt.cgi?l=129238; locuslink at NCBI, predicted gene *LOC129238*.
http://www.ncbi.nlm.nih.gov/SNP/index.html; dbSNP home page.

Received July 1, 2002; accepted in revised form January 28, 2003.

2.3.3 Recherche de minisatellites potentiellement polymorphes dans les séquences codantes

2.3.3.1 Introduction

Comme nous l'avons vu précédemment (Vergnaud & Denoeud 2000), le génome humain est riche en répétitions en tandem. Le chromosome 22, qui représente environ 1% du génome humain, contient de l'ordre de 15000 répétitions en tandem (détectées par le Tandem Repeats Finder (Benson 1999)) : on peut donc estimer le nombre de répétitions en tandem sur l'ensemble du génome à 1500000, dont plus de 5000 appartiendraient à la classe des minisatellites.

Cependant, seule une quinzaine de minisatellites polymorphes appartenant à des séquences codantes, c'est-à-dire générant des répétitions en tandem au niveau de la séquence en acides aminés de la protéine correspondante, ont été étudiés, la plupart appartenant à la famille des mucines (Vinall 1998) : voir Tableau 9, paragraphe 1.2.3.3.3.2. D'autre part, l'étude menée sur les chromosomes 21 et 22 (Denoeud 2003) a mis en évidence un minisatellite hypermutable appartenant à une séquence codante prédite, mais l'existence de cette protéine hypothétique nécessitera d'être vérifiée. Les répétitions survenant au niveau de séquences codantes sont susceptibles de générer une variabilité au niveau des protéines, éventuellement nécessaire à leur rôle physiologique (système immunitaire…) ou générant des pathologies. Il semble donc intéressant d'étudier plus largement cette classe de répétitions en tandem encore peu explorée.

Certains articles traitent plus particulièrement des microsatellites survenant dans les séquences codantes : une étude menée sur des protéines de primates a par exemple montré que les répétitions trinucléotidiques étaient sur-représentées dans les exons (Borstnik 2002). Wren et collègues ont étudié les microsatellites polymorphes présents dans les ARNs messagers humains : ils sont plus fréquents dans les séquences non-codantes (5' ou 3') que dans les séquences codantes. Les auteurs prédisent toutefois que 3,7% des gènes humains contiendraient des répétitions en tandem polymorphes, dont 92% correspondraient à une unité répétée multiple de 3 (Wren 2000). Il apparaît en outre que certaines répétitions d'un acide aminé (trinucléotides) sont plus fréquentes que d'autres : poly-alanine, proline, glycine, glutamine, lysine, leucine et sérine (Borstnik 2002 ; Wren 2000).

Marcotte *et al.* ont mis en évidence que les protéines d'eucaryotes sont plus riches en répétitions (principalement duplications) que les protéines de bactéries ou d'archées : ils ont émis l'hypothèse que des mécanismes similaires seraient à l'origine des duplications dans les trois règnes, mais que la machinerie de synthèse protéique eucaryote, prenant mieux en charge

les protéines multi-domaines, non-globulaires, permettrait le maintien des structures répétées chez ces organismes préférentiellement (Marcotte 1999).

De Fonzo *et al.* ont mené une étude sur les répétitions en tandem (microsatellites comme minisatellites) présentes dans des gènes associés à des pathologies (en particulier celles qui sont associées à un phénomène d'anticipation) : certains candidats ont été trouvés à proximité de gènes ou dans des gènes (leurs unités répétées étant alors systématiquement multiples de 3) mais leur polymorphisme nécessite encore d'être testé (De Fonzo 1998).

2.3.3.2 Séquences traitées

Afin d'identifier les minisatellites appartenant à des séquences codantes, j'ai rapatrié toutes les séquences d'ARNs messagers du génome humain présentes dans Refseq en juin 2003. Refseq, « The Reference Sequence » collection (http://www.ncbi.nih.gov/RefSeq/) contient l'ensemble des séquences non-redondantes (ADN génomique, transcrits (ARNs messagers) et produits protéiques) pour différents organismes. Les ARNs messagers provenant de Refseq appartiennent à deux catégories principales : les gènes, dont le produit protéique est connu, et les modèles de gènes, prédits par des logiciels de détection de gènes et étayés par différents types de preuves : alignement avec des ARNs messagers (ARNm) et/ou des EST (Expressed Sequence Tags : séquences provenant de banques d'ADN complémentaire). Parmi les 37547 ARNm rapatriés, 18474 correspondent à des gènes et 19016 correspondent à des modèles. Les autres ARNs messagers correspondent à des gènes d'ARNs ou des pseudogènes : ils n'ont pas été considérés dans cette étude. Les ARNm considérés totalisent 76921544 paires de bases, soit environ 2,5% de l'ensemble du génome humain.

2.3.3.3 Caractéristiques des répétitions en tandem des ARNs messagers

Tableau 10 : Caractéristiques des séquences des chromosomes 20, 21 et 22 et de l'ensemble des ARNs messagers du génome humain.

	Chromosome 20	Chromosome 21	Chromosome 22	ARNs messagers
Longueur de la séquence traitée (pb)	59422997	33824148	33563000	76921544
Densité en gènes (gènes/Mb) (Venter 2001)	16	13	23	
% GC (Venter 2001)	44%	41%	48%	
% de répétitions (dispersées…) (Venter 2001)	41%	38%	44%	
Nombre de répétitions en tandem	24484	14996	14369	26153
Densité en répétitions en tandem (/Mb)	412	443	428	340
Nombre de minisatellites (définition au § 2.3.3.4)	51	/	/	50
% de minisatellites parmi les répétitions en tandem	0,21 %	/	/	0,19 %
Densité en minisatellites (/Mb)	0,86	/	/	0,65

Les séquences d'ARNm ont été soumises au Tandem Repeats Finder (Benson 1999) puis les répétitions en tandem identifiées ont été importées dans la base de données comme décrit précédemment (chapitre 2.1.1) : 26153 répétitions en tandem ont été identifiées dans les ARNs messagers, ce qui correspond à une densité en répétitions en tandem de 340/Mb, inférieure aux densités observées pour les chromosomes 20, 21 et 22 entiers : voir Tableau 10. Cette observation est cohérente avec une pression de sélection plus forte dans les régions codantes par rapport aux régions intergéniques.

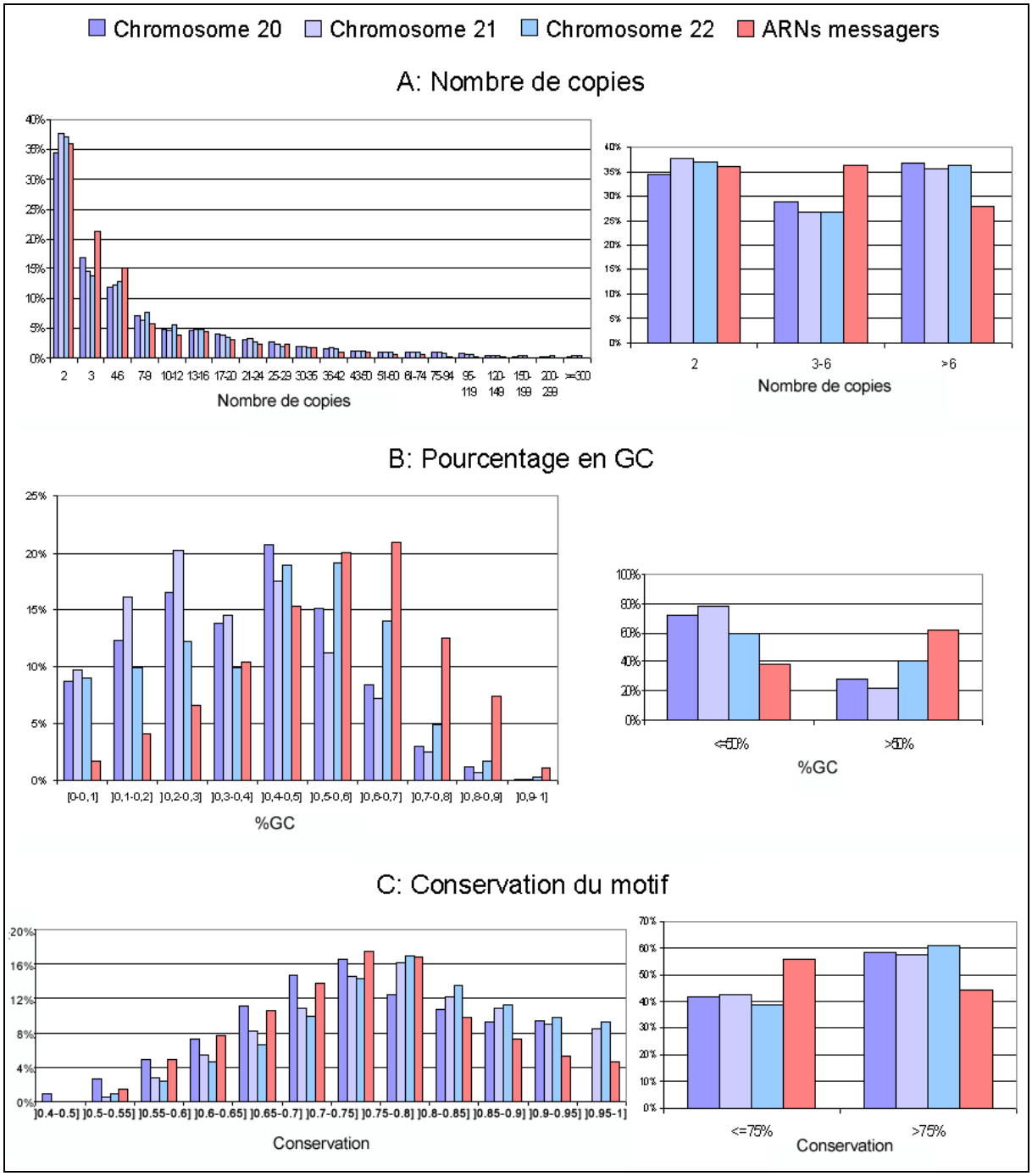


Figure 21 : Comparaison des distributions du nombre de copies, du %GC et de la conservation pour les répétitions en tandem des chromosomes 20, 21 et 22, et des ARNs messagers.

Nous avons comparé les caractéristiques des répétitions en tandem des ARNs messagers à celles des chromosomes 20, 21 et 22. Les histogrammes présentés sur la Figure 21 font apparaître que les répétitions en tandem des ARNs messagers ont, par rapport à l'ensemble des répétitions en tandem des trois chromosomes considérés :

- un plus faible nombre de copies (Figure 21A) : pour les ARNm, il y a un excès de répétitions en tandem avec un nombre de copies entre 3 et 6 et un défaut de répétitions en tandem avec un nombre de copies supérieur à 6 par rapport aux trois chromosomes entiers. Cette observation reflète probablement le fait que les répétitions en tandem trop longues peuvent nuire à la fonction des gènes et sont donc contre-sélectionnées.
- un plus fort pourcentage en GC (Figure 21B) : pour les ARNm, il y a un excès de répétitions en tandem avec un pourcentage en GC supérieur à 50% par rapport aux trois chromosomes entiers, même si leur situation est hétérogène : en effet, les chromosomes 20, 21, et 22 ont des %GC variables : 44%, 41%, et 48% respectivement (Venter 2001). Le fait que les minisatellites d'ARNs messagers soient plus riches en GC reflète probablement le fait que les régions riches en GC du génome humain sont plus riches en gènes (Lander 2001 ; Venter 2001).
- un plus faible degré de conservation entre unités répétées (Figure 21C) : pour les ARNm, il y a un excès de répétitions en tandem avec une conservation inférieure à 75% par rapport aux chromosomes 20, 21 et 22.

Chez les bactéries, la plupart des répétitions en tandem a un motif répété d'une taille multiple de 3 (Le Flèche 2001), ce qui reflète probablement le fait qu'une majorité des répétitions se trouve dans des séquences codantes (en tout cas dans certaines espèces), les génomes bactériens étant plus denses en gènes. Pour les ARNs messagers humains, nous avons donc comparé la proportion de répétitions en tandem avec des unités répétées multiples de 3 entre les ARNm et les chromosomes 20, 21, et 22 : voir Figure 22. Il y a un très net excès de répétitions en tandem avec des unités répétées multiples de 3 dans les ARNs messagers par rapport aux trois chromosomes considérés. On peut donc supposer qu'une proportion conséquente de ces répétitions en tandem se situe dans les régions codantes et non pas dans les régions non traduites en 5' (5'UTR) et en 3' (3'UTR).

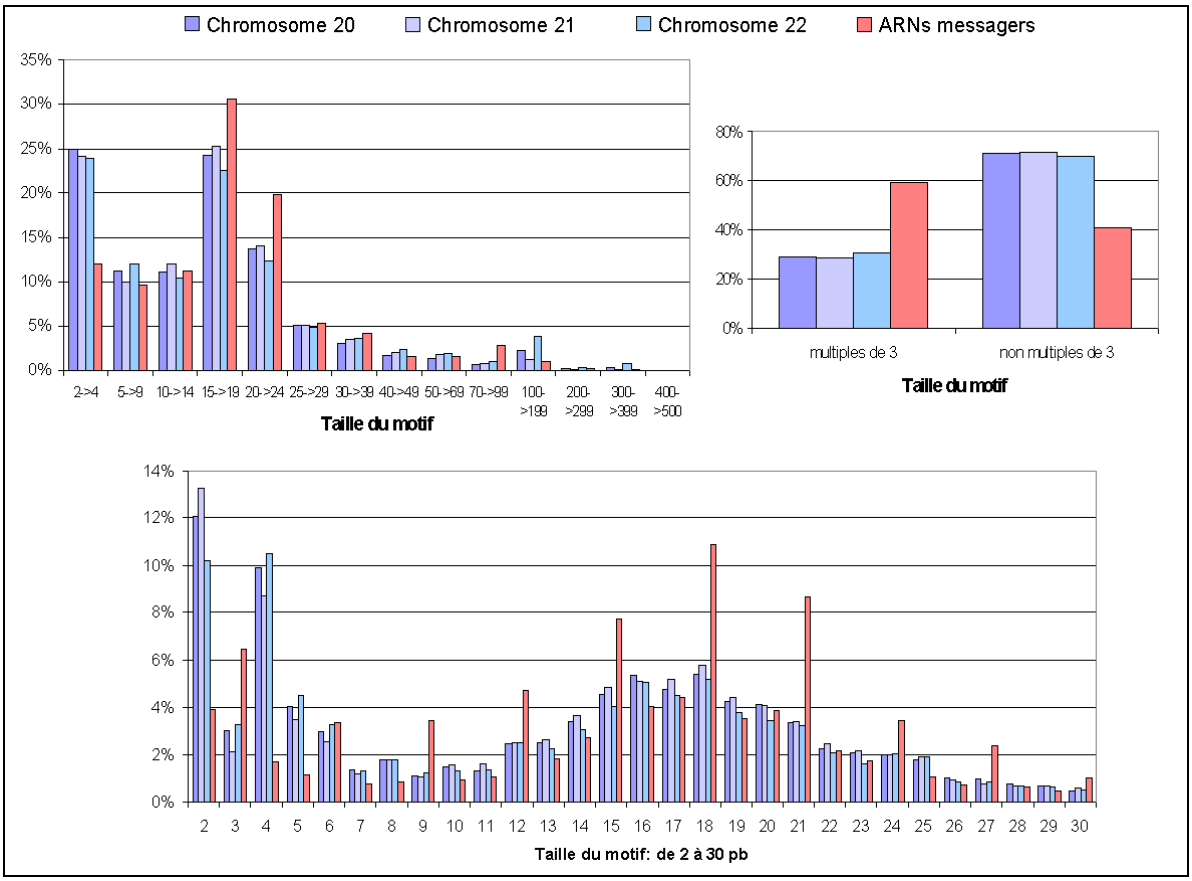


Figure 22 : Comparaison des distributions de la taille de motif pour les répétitions en tandem des chromosomes 20, 21 et 22, et des ARNs messagers.

2.3.3.4 Sélection des minisatellites

Le but de cette étude étant d'identifier les minisatellites potentiellement polymorphes dans les séquences codantes, nous avons utilisé les critères %GC et HistoryR qui ont été démontrés comme prédictifs sur le polymorphisme (Denoeud 2003). Les minisatellites que nous avons sélectionnés correspondent à la requête :

- Longueur totale supérieure ou égale à 200 paires de bases (pb)
- Unité répétée supérieure ou égale à 12 pb
- Nombre de copies supérieur ou égal à 4
- Pourcentage en GC supérieur ou égal à 70%
- HistoryR supérieur ou égal à 0.6

La requête utilisée sélectionne 50 minisatellites potentiellement polymorphes dans les ARNs messagers. La proportion de minisatellites répondant à la définition proposée ci-dessus parmi l'ensemble des répétitions en tandem est similaire à celle du chromosome 20 (Tableau 10) : 0,2% environ.

Les seuils sur la longueur totale et le nombre de répétitions sont peu stringents, car il est probable, compte-tenu de la pression de sélection, que les minisatellites présents dans les séquences codantes soient de taille relativement restreinte. Cette hypothèse est en accord avec la définition élargie des minisatellites proposée dans l'introduction et dans la revue du chapitre 2.3.1 : les répétitions en tandem de plus de 140 paires de bases peuvent être considérées comme des minisatellites.

Lors de l'étude menée sur les minisatellites des chromosomes 21 et 22, présentée au chapitre 2.3.2 (Denoeud 2003) une requête sélectionnant les minisatellites avec %GC ≥ 48% et HistoryR ≥ 0.54 produisait un enrichissement en minisatellites polymorphes de 43% à 59% (spécificité : 59%) dans le lot sélectionné, et ne manquait que 30% des minisatellites polymorphes (sensibilité : 70%). La requête utilisée ici est plus stringente : on peut donc espérer qu'au moins 60% des minisatellites sélectionnés (soit 36 sur 50) soient effectivement polymorphes, et que moins d'une vingtaine aient été manqués.

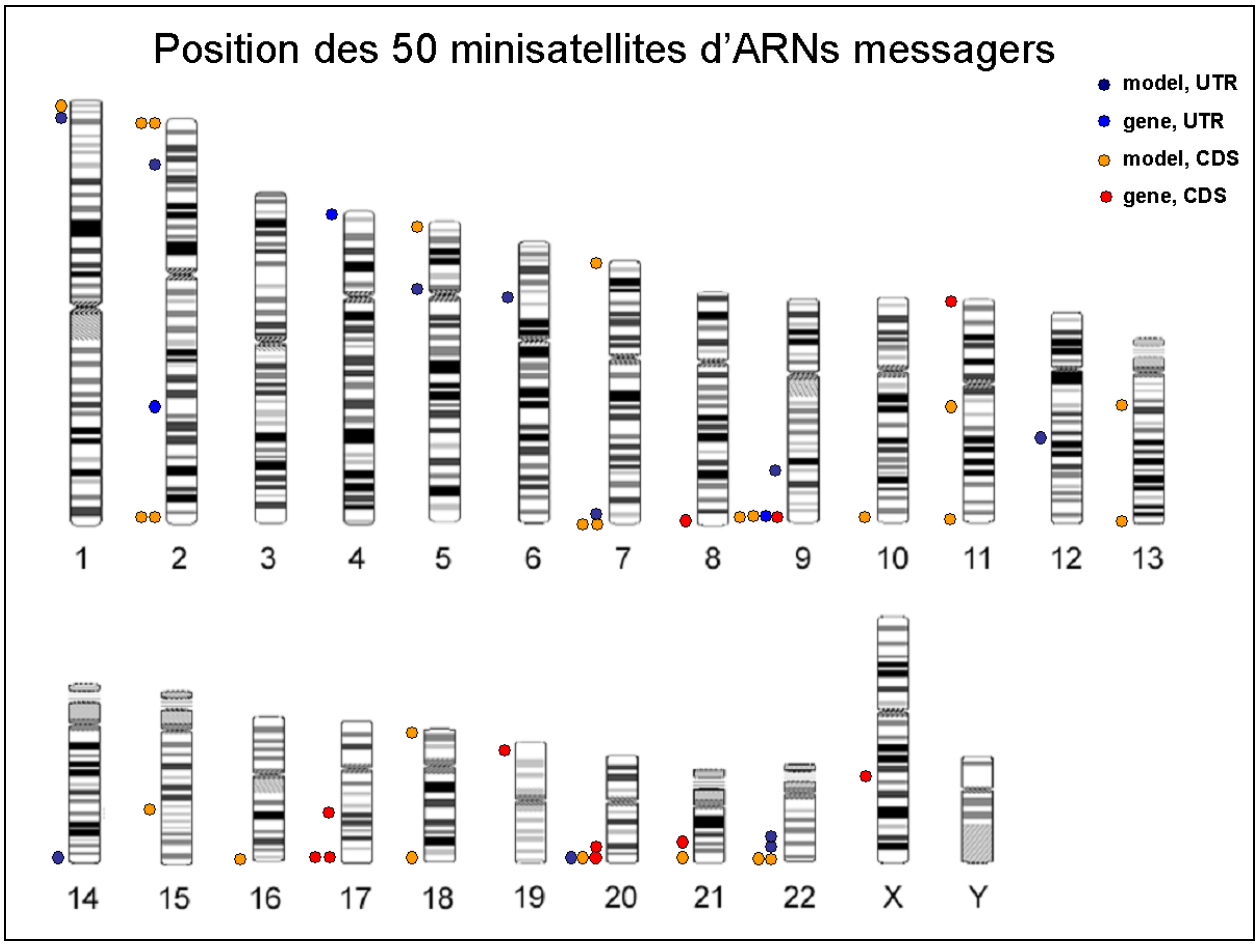


Figure 23 : Localisation chromosomique des minisatellites identifiés dans des séquences d'ARNm.

La Figure 23 représente la localisation des 50 minisatellites potentiellement polymorphes détectés dans des ARNs messagers, sur le génome humain. Les minisatellites des ARNs messagers montrent sensiblement la même répartition (télomérique) que celle indiquée par (Amarger 1998) pour la population générale des minisatellites.

La Figure 24 montre la distribution des minisatellites correspondant à la même requête sur le chromosome 20 (dans ou hors des ARNs messagers) : le biais de distribution télomérique est similaire (31 minisatellites parmi 50 sont situés aux extrémités des chromosomes pour les ARNs messagers, et 27/51 pour le chromosome 20).

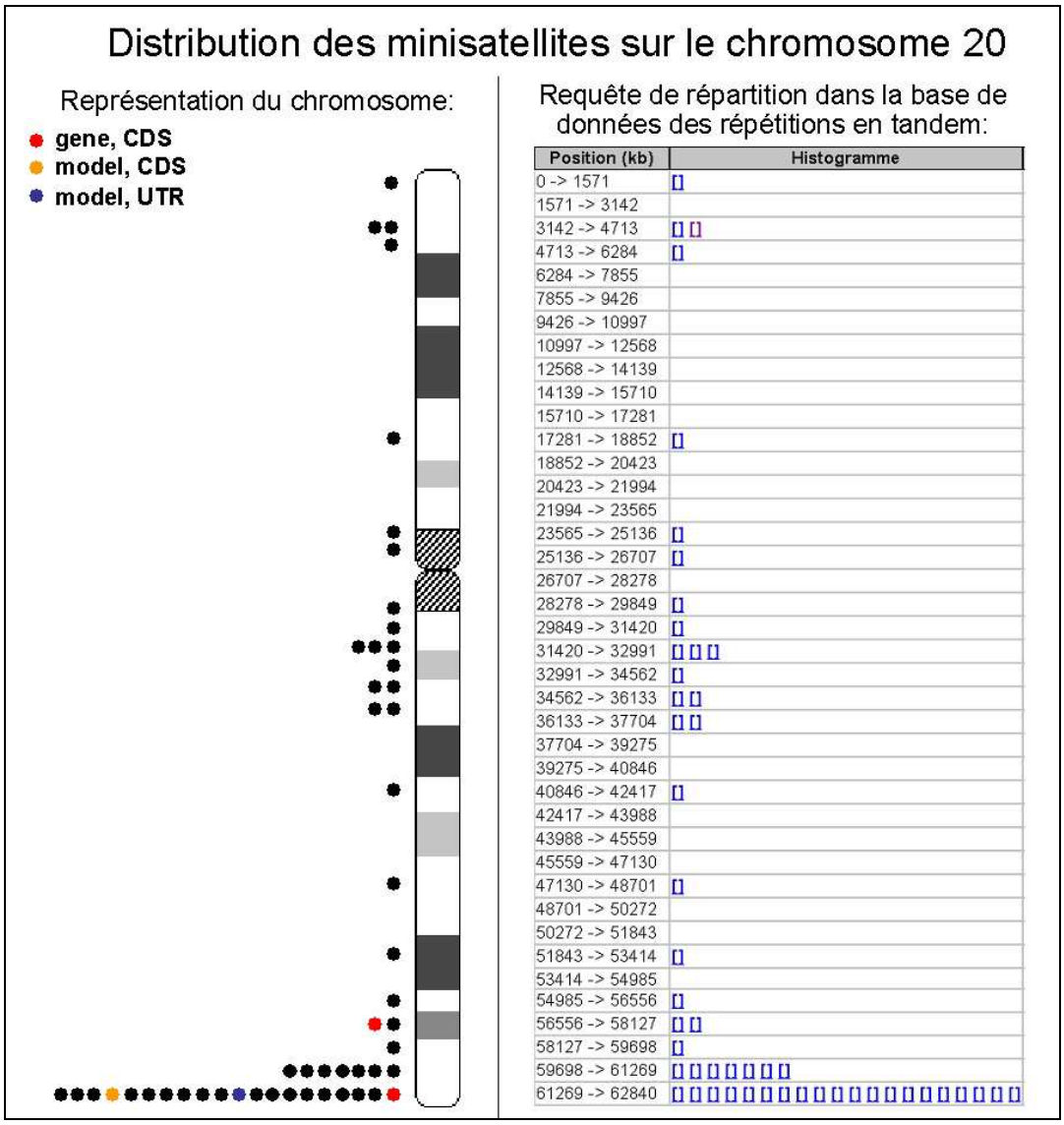


Figure 24 : Localisation des minisatellites identifiés dans la séquence complète du chromosome 20 (à droite : illutration d'une requête de répartition dans la base de données des répétitions en tandem).

2.3.3.5 Caractéristiques des 50 minisatellites potentiellement polymorphes sélectionnés

L'Annexe 4 présente la description des 50 minisatellites et des ARNs messagers qui leur sont associés. Pour les 50 minisatellites répondant à la requête, nous avons recherché si la répétition en tandem était située dans la région codante (CDS) ou dans les régions non traduites (UTRs), et si l'ARNm correspondait à un gène ou à un modèle de gène (prédiction). Les résultats sont présentés dans le Tableau 11. Parmi les 50 minisatellites, 36 sont situés dans

des ARNs messagers correspondant à des prédictions de gènes (modèles), soit 72%. Dans l'ensemble des ARNs messagers traités, seuls 51% correspondent à des modèles et 49% à des gènes avec des produits protéiques de fonction connue, inférée, ou inconnue.

Tableau 11 : Distribution des minisatellites d'ARNs messagers, de motif multiple de 3 ou non multiple de 3, dans les différentes classes de séquences (CDS, UTRs, gènes, modèles).

Total : 50 minisatellites potentiellement polymorphes	CDS		UTRs	
	gènes	modèles	gènes	modèles
Motif multiple de 3	10	11	1	2
Motif non multiple de 3	1	14	2	9
Total	11	25	3	11

Le Tableau 11 présente également le nombre de minisatellites d'unité répétée multiple ou non multiple de 3 situés dans des UTRs, dans des CDS, de gènes ou de modèles de gènes. On note que les motifs « non multiples de 3 » sont dans leur très grande majorité rencontrés dans les modèles (23 sur 27). Le seul gène mis en évidence possédant une répétition non multiple de 3 dans sa région codante correspond à une protéine hypothétique (certains ARNs messagers sont en effet classés parmi les « gènes » et non les « modèles de gènes », alors que la protéine correspondante est qualifiée d'hypothétique, de fonction inconnue). En principe, des répétitions nucléiques non multiples de 3 peuvent générer des répétitions d'acides aminés si le nombre de répétitions est suffisant : 3 unités répétées génèrent alors un « super-motif » nucléotidique multiple de 3, qui codera pour une répétition d'acides aminés. La Figure 25 présente la traduction correspondant au minisatellite du gène du récepteur à la dopamine (DRD4) , unité répétée : 48 pb, et du gène prédit « LOC349662 » (catégorie : « ab initio, with EST support »). Cependant, le fait que de telles structures (motifs non multiple de 3 pb codants) ne soient obtenues que pour des prédictions de gènes laisse présager que les modèles de gènes contenant des répétitions nucléiques codantes mais non multiples de 3 seraient des artéfacts ne correspondant pas à des gènes réels. D'une manière plus générale, on constate que, en accord avec les observations menées sur la répartition des répétitions en tandem dans les génomes (Li 2002), la majorité des minisatellites répondant à la requête d'enrichissement en locus polymorphes et appartenant à des CDS ont des unités multiples de 3 (21/36 soit 58%, ou si on ne considère pas les modèles, 10/11 soit 91%), et la majorité des minisatellites appartenant à des UTRs ont des motifs non multiples de 3 (11/15 soit 73%). Pour les UTRs, contrairement aux CDS, les proportions de minisatellites à motif multiple de 3 ne sont pas significativement différentes entre gènes et modèles.

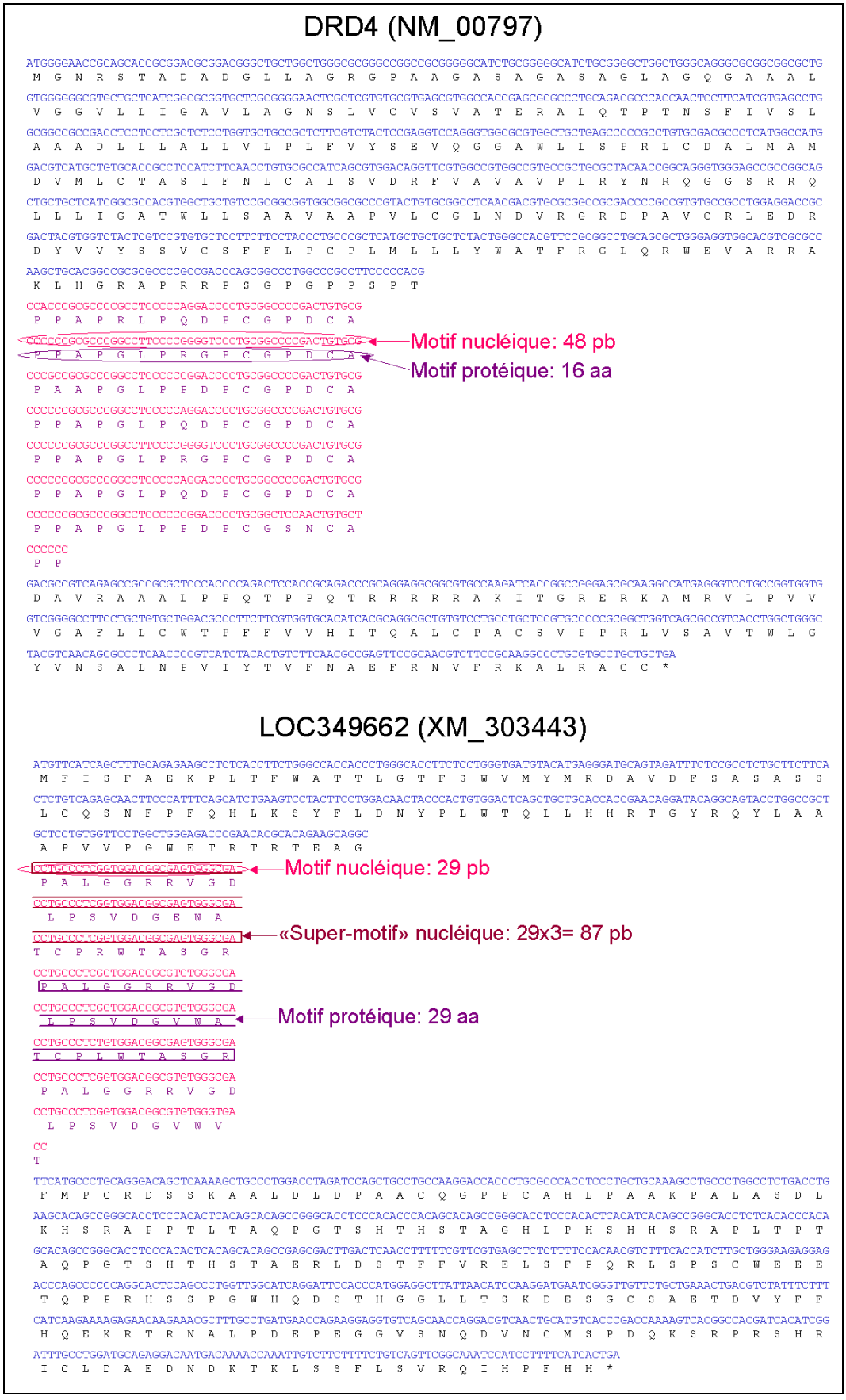


Figure 25 : Traduction des motifs répétés nucléiques en motifs répétés protéiques, pour deux minisatellites (l'un d'unité répétée de 48 pb, multiple de 3, et l'autre d'unité répétée de 29 pb).

2.3.3.6 Minisatellites appartenant à des séquences codantes

Nous avons identifié 36 minisatellites appartenant à des séquences codantes, dont 25 correspondent à des modèles de gènes (prédits), 2 à des gènes ayant des produits protéiques hypothétiques, et 9 à des gènes ayant des produits protéiques de fonction connue ou inférée : ces 9 protéines sont décrites dans le Tableau 12. Parmi ces répétitions en tandem, 5 sont déjà connues comme étant polymorphes : 3 d'entre elles (DRD4, CEL, OGFR) sont listées dans le Tableau 9 (paragraphe 1.2.3.3.3.2) car leur polymorphisme a déjà été largement étudié. Les deux autres correspondent à un minisatellite du chromosome 21 qui s'est révélé polymorphe (2 allèles) lors de l'étude que j'ai menée sur les chromosomes 21 et 22 (Denoeud 2003) (gène CLIC6 : ce minisatellite, CEB234, n'avait pas été localisé dans un gène lors d'analyses effectuées en 2001, alors qu'il fait partie d'une séquence codante selon la mise à jour du génome humain d'avril 2003), et à un ARN messager ayant plusieurs transcrits, provenant d'un phénomène d'épissage alternatif, différant par leur nombre de répétitions en tandem (gène MADCAM1) (Leung 1996). Quatre minisatellites potentiellement polymorphes, mais encore peu étudiés, situés dans des gènes connus ont donc été identifiés grâce à cette étude : ils appartiennent aux gènes GRINA, KRT10, GNAS1 et ESX1L, tous 4 impliqués dans des pathologies. L'étude de leur polymorphisme et de son influence sur ces pathologies mériterait donc d'être menée. De plus, les deux derniers gènes listés sont associés à des phénomènes d'empreinte parentale, ce qui les rend particulièrement intéressants à étudier.

Parmi les 25 minisatellites appartenant à des séquences codantes de modèles de gènes, 11 ont des motifs répétés de longueur multiple de 3 et sont donc plus susceptibles de correspondre à des gènes réels, et surtout d'être polymorphes (en effet, il est très peu vraisemblable que des répétitions polymorphes appartenant à des séquences codantes aient des motifs répétés non multiples de 3 : ce n'est en tout cas observé pour aucune de celles listés dans le Tableau 9, voir paragraphe 1.2.3.3.3.2, ni pour celles identifiées par De Fonzo *et al.* (De Fonzo 1998)). A ces 11 minisatellites, on peut ajouter le minisatellite identifié dans un gène codant pour une protéine hypothétique, de fonction inconnue, et qui a également un motif répété multiple de 3. Le Tableau 13 présente ces 12 gènes qui, s'ils ont été correctement prédits, contiennent des minisatellites potentiellement polymorphes dans leurs séquences codantes. Ces gènes, n'ayant jamais été étudiés, mériteraient donc une attention particulière, surtout ceux pour lesquels la répétition en tandem constitue la majeure partie de la protéine. Certaines de ces séquences codantes sont en effet constituées en majorité -58%, 84%, 86%- par la répétition en tandem, ce qui laisse supposer un rôle prépondérant de leurs répétitions d'acides aminés sur la structure tri-dimensionnelle de la protéine, et donc sur sa fonction. De façon intéressante, pour les 9 gènes présentés dans le Tableau 12, la proportion de la séquence protéique correspondant à la répétition en tandem n'atteint pas plus de 32% (des proportions bien plus importantes sont toutefois observées pour des protéines connues comme les mucines : voir paragraphe 1.2.3.3.3.2 (Vinall 1998 ; Debailleul 1998)).

Tableau 12 : Gènes de fonction connue contenant des minisatellites codants.

Numéro d'accession de l' ARNm	Position de la répétition dans l'ARN	Taille du motif	Nombre de répétitions	Longueur totale	% conservation	%GC	HistoryR	Chromosome	Répétition d'acides aminés (proportion de la protéine constituée par la répétition en tandem)	Nom du gène	ID locuslink **	Fonction de la protéine	Pathologie associée	Polymorphisme étudié ?	Références
XM_291268	253-495	15	16,2	243	59%	70%	0,78	8q24.3	5aa: YPQGP (23%)	GRINA	2907	« glutamate receptor, ionotropic, N-methyl D-asparate-associated protein 1 » (liaison au glutamate)	troubles psychiatriques	non	
NM_001807	1693-2218	33	15,9	526	92%	80%	0,92	9q34.3	11aa: PVPPTGDSGAP (23%)	CEL	1056	carboxyl ester lipase (homéostasie du cholestérol)	athérosclérose	oui: 6 allèles	(Higuchi 2002)
NM_000797	740-1162	48	8,81	423	91%	84%	0,8	11p15.5	16aa: LPQDPCGPDCAPPAPG (30%)	DRD4	1815	récepteur de la dopamine D4	troubles de l'attention et de la personnalité	oui: 9 allèles	(Lichter 1993)
NM_000421	1352-1626	15	18	275	62%	79%	0,63	17q21-q23	5aa: SSGGG (16%)	KRT10	3858	kératine 10	hyperkératose épidermolytique	non	
NM_130761	692-897	24	8,6	206	96%	71%	1	19p13.3	8aa: TSPESPDT (17%)	MADCAM1	8174	« mucosal vascular addressin cell adhesion molecule 1»	impliquée dans les processus inflammatoires	oui: 3 variants (1: N=8,6 ; 2: N=0 ; 3: N=5,6)	(Leung 1996)
NM_007346	1576-1996	60	7	421	94%	73%	0,85	20q13.3	20aa : PSETPGPRPAGPAGDEPAE (23%)	OGFR	11054	récepteur de facteur de croissance opioïde	rejet de tumeurs dirigé par l'antigène de l'angiogénèse	oui: 5 variants (0 à 8 répétitions)	(Mollick 2003)
NM_080425	805-1097	36	8,13	293	92%	79%	1	20q13.2-q13.3	12aa: PDAPADPDSGAA (11%)	GNAS1*	2778	« stimulatory guanine nucleotide-binding protein G »: sous-unité α	pseudo-hypoparathyroïdie, type 1a (PHP1a) ; osteodystrophie héréditaire d'Albright	non	(Lalande 2001)
NM_053277	465-860	30	13,2	396	82%	74%	0,7	21q22.12	20aa: GPAGDSVDAEGRVGDSVDAE (19%)	CLIC6	54102	canal chlore intracellulaire, 6	?	oui: 2 allèles, CEB234	(Denoeud 2003)
NM_153448	784-1177	27	14,6	394	90%	75%	0,86	Xq22.1	9aa: VPPGPPMAP (32%)	ESX1L	80712	« extraembryonic, spermatogenesis, homeobox 1-like » (facteur de transcription)	rôle supposé sur les phénomènes d'empreinte parentale	non	(Fohn 2001)

*épissage alternatif: la répétition en tandem n'est présente que dans le variant 3 qui contient un exon 5' différent ; ce gène est soumis à des phénomènes d'empreinte parentale

**Locuslink : http://www.ncbi.nlm.nih.gov/LocusLink/

Tableau 13 : Gènes hypothétiques contenant des minisatellites codants.

ARNm	Taille du motif	Nombre d'unités	Longueur totale	% conser-vation	%GC	HistoryR	Chromosome	Position de la répétition dans l'ARN	Position de la séquence codante dans l'ARN	Proportion de la protéine correspondant à la répétition	Répétition d'acides aminés	Catégorie de locus
XM_292899	24	32,41	778	94%	72%	0,77	2p25.3	522-1299 (atrophin-1 region)	1-1353	58%	8aa: DPTPVPSA	model, ab initio, with EST support
XM_293077	39	14,6	568	99%	70%	1	2q37.3	582-1149	13-1629	35%	13aa: RSSYTLGRGPVAP	model, ab initio, with EST support
XM_294605	24	115	2759	95%	70%	0,91	9q34.3	360-3118	195-3464	84%	8aa: DTPRPRDC	model, ab initio, with EST support
XM_305804	45	6,02	271	96%	70%	1	9q34.3	2100-2370	1-2412	11%	15aa: ATLIPVPPVLSPLST	model, ab initio, with EST support
XM_172821	33	6,75	223	94%	74%	1	10q26.3	62-284	41-517	47%	11aa: TPASSSSAAPP	model, supported by mRNA alignments
XM_062938	39	39,3	1530	92%	73%	0,71	13q34	197-1726	1-1770	86%	13aa:LGGDRTHGDAPSR	model, ab initio, with EST support
XM_301533	30	7,3	220	96%	74%	1	15q22.32	82-301	1-441	50%	10aa: EPGDGGKLPP	model, ab initio
XM_303976	33	7,1	234	96%	71%	1	16q24.3	1316-1549	1-2127	11%	11aa: GYRNSLFPPGE	model, ab initio, with EST support
NM_024510	48	4,8	230	98%	70%	1	17q25.3	871-1100	728-1411	34%	16aa: VPEPVHRPQDPWHIPG	**gene** with protein product, function unknown
XM_210355	39	5,3	207	95%	76%	1	18q23	378-584	1-846	24%	13aa: GSRSGEPRAPTDR	model, ab initio
XM_066408	33	32,6	1076	97%	71%	0,88	22q13.33	785-1860	1-1869	58%	11aa: GEAQSFPVEES	model, ab initio, with EST support
XM_301914	33	8,1	267	97%	73%	1	22q13.33	139-405	1-918	29%	11aa: HTAPDTPAQPF	model, ab initio

2.3.3.7 Conclusions

Différentes observations nous laissent penser que certains modèles de gènes, provenant de prédictions *ab initio*, ne doivent pas correspondre à de véritables gènes :

- On note un excédent de modèles de gènes parmi les ARNs messagers contenant des minisatellites : une hypothèse pour expliquer cette observation serait que certains de ces modèles sont faux.
- Seuls les modèles de gènes contiennent des répétitions nucléiques codantes d'unité non multiple de 3. Si de telles structures étaient utilisées pour coder pour des répétitions d'acides aminés (« super-motifs » de 3 motifs nucléiques codant pour un motif protéique : voir Figure 25), on observerait vraisemblablement des co-mutations au niveau de la séquence nucléique tous les 3 motifs répétés. Il est d'ailleurs probable que le Tandem Repeats Finder (Benson 1999) aurait identifié de façon redondante des répétitions avec des consensus 3 fois plus longs, et une meilleure conservation. L'analyse des séquences nucléiques correspondantes ne montre rien de tel. Il est donc très probable que les séquences codantes prédites, correspondant à des unités répétés non multiples de 3 paires de bases sont en fait des erreurs de prédiction. On ne peut pas exclure toutefois que ces répétitions en tandem soient monomorphes (et maintenues comme telles par la pression de sélection) auquel cas la taille du motif aurait moins d'importance.

On peut donc supposer que les séquences codantes prédites (modèles de gènes) contenant des minisatellites sont pour la plupart (en tout cas pour ceux ayant des motifs non multiples de 3) des erreurs de prédiction, tout du moins en ce qui concerne les frontières entre introns et exons. Il y a donc vraisemblablement moins de minisatellites appartenant à des séquences codantes que nous n'en avions identifié en première intention, c'est pourquoi nous n'avons décrit que les minisatellites à motif multiple de 3 (soit 21/36). Afin d'améliorer la prédiction de gènes, des critères tels que « absence de répétitions en tandem à motif non multiple de 3 » dans la séquence codante pourraient être appliqués. En effet, les minisatellites codants restent très rares : si de telles structures sont identifiées (quelle que soit la longueur du motif), il convient de vérifier la validité du gène prédit.

Les nouveaux minisatellites codants identifiés par cette étude sont potentiellement polymorphes : ils ont été sélectionnés selon des critères corrélés au polymorphisme (Denoeud 2003). Ce polymorphisme nécessite d'être confirmé. Le fait que parmi les 9 minisatellites codants isolés dans des gènes connus, 5 aient déjà été étudiés et caractérisés comme étant polymorphes, laisse présumer que la requête utilisée est efficace pour sélectionner des minisatellites polymorphes. Cependant, elle n'a permis d'identifier que 4 nouveaux minisatellites dans des gènes connus et 12 dans des gènes hypothétiques. De plus, les mucines, famille de protéines contenant des répétitions en tandem polymorphes (Vinall 1998), n'ont pas été identifiées par cette requête. Certains minisatellites codants ont donc été manqués : une requête moins stringente mériterait d'être testée. Par exemple, on pourrait

tâcher de trouver des caractéristiques communes aux minisatellites codants déjà étudiés (Tableau 9) afin de générer une requête plus générale. En tout cas, cette approche pour identifier des répétitions en tandem codantes, c'est-à-dire potentiellement fonctionnelles, est prometteuse. En particulier, cette étude semble cibler de façon préférentielle des locus impliqués dans des phénomènes d'empreinte parentale, comme le gène GNAS1, épissé de façon différentielle selon l'origine maternelle ou paternelle des allèles (Lalande 2001), et dont l'un des transcrits contient un minisatellite potentiellement polymorphe (Tableau 12).

3 Discussion et perspectives

3.1 La base de données des répétitions en tandem

Dans cette thèse, j'ai illustré l'utilité de la base de données des répétitions en tandem. Elle a été mise à profit au laboratoire pour :

- L'étude de la répartition des minisatellites sur des chromosomes eucaryotes (Vergnaud 2000).
- L'identification de marqueurs polymorphes pour l'épidémiologie bactérienne (Le Flèche 2001; Le Flèche 2002), en particulier grâce à la page de comparaison de souches (Denoeud 2004).
- La mise en évidence de critères prédictifs sur le polymorphisme de minisatellites humains (Denoeud 2003).

Cette base de données est accessible sur Internet, dans le but d'être utile à d'autres utilisateurs, ce qui semble être le cas au vu des nombreuses requêtes qui y sont menées quotidiennement, en particulier pour les génomes bactériens. Cependant, ce site reste « artisanal » (j'en suis pour l'instant la seule administratrice) et on ne peut qu'espérer qu'un des sites majeurs fournisseurs de services bioinformatiques trouve suffisamment d'intérêt à cet outil pour en réaliser des développements plus performants et plus durables.

Certaines améliorations mériteraient encore d'être apportées à la base de données. Comme nous l'avons décrit au paragraphe 2.1.1.1, l'élimination de la redondance dans les résultats du logiciel Tandem Repeats Finder (Benson 1999) a fait l'objet de choix automatiques sur des critères empiriques qui étaient dans certains cas peu judicieux. Par exemple, lorsqu'une répétition en tandem « émerge » à l'intérieur d'une autre (comme nous l'avons vu dans le cas des « VLTR » ou « variable length tandem repeat » (Hauth 2002) : Figure 11), le choix effectué va masquer cette émergence en présentant la plus grande répétition en tandem. Des ajustements pourraient permettre de compenser ce défaut : la base pourrait inclure par exemple une mesure de la distorsion de la conservation interne, ce qui permettrait d'identifier les répétitions en tandem qui en contiennent une autre, d'évolution plus récente, à motifs souvent beaucoup mieux conservés. Une amélioration a déjà été apportée par rapport à la première version de la base : les alignements retenus proposent un lien vers les autres alignements (non retenus) de la région de redondance, afin de permettre à l'utilisateur de choisir. Cette fonctionnalité n'est toutefois disponible que pour les génomes importés récemment. C'est également le cas d'autres options telles que l'accès à la séquence brute des répétitions en tandem (sans alignement des unités répétées avec le motif consensus).

Malgré ces quelques réserves, la base de données que j'ai développée reste à ce jour le seul site spécialement dédié à l'identification de répétitions en tandem dans les génomes, utilisable sans compétences ou goûts particuliers pour la bioinformatique. Au vu de l'intérêt croissant porté ces dernières années au génotypage de souches par les répétitions en tandem, on peut

espérer que la page consacrée aux génomes bactériens (et notamment la page de comparaison de souches) devienne un outil de référence pour les développeurs de tels moyens de typage.

A titre de comparaison, le site développé par Gary Benson (TRDB : Tandem Repeats Database [http://tandem.bu.edu/cgi-bin/trdb/trdb.exe]) dédié à la détection et à l'analyse des répétitions en tandem, n'a pas la même vocation que notre base de données. Il s'adresse plutôt à des « spécialistes » des répétitions en tandem, en leur proposant des utilitaires de visualisation et de classification des répétitions en tandem présentes dans les séquences qu'ils soumettent. Ce site est toutefois très utile pour analyser les séquences de différents allèles de répétitions en tandem polymorphes. En effet, pour le génotypage de bactéries, il est de plus en plus fréquent de pousser l'analyse jusqu'au séquençage des allèles. Dans notre laboratoire, différents locus polymorphes de bactéries pathogènes sont ainsi séquencés, et nous sommes des utilisateurs réguliers de la base de Gary Benson, avec qui nous entretenons une collaboration.

3.2 Répétitions en tandem et phylogénie

3.2.1 Intérêt phylogénique du typage de répétitions en tandem bactériennes

En épidémiologie bactérienne, les répétitions en tandem sont utilisées afin de distinguer différents isolats ou souches. Les génotypes obtenus par typage des répétitions en tandem permettent d'obtenir une classification des différentes souches, sous forme d'un arbre, ce qui donne une idée de leur proximité. Cette proximité vis à vis des répétitions en tandem n'a toutefois pas nécessairement une valeur phylogénique : l'arbre basé sur les répétitions en tandem peut ne pas correspondre à l'arbre évolutif réel de l'espèce. En effet, certains allèles peuvent être de taille identique sans avoir suivi le même chemin évolutif. Ce phénomène est appelé homoplasie. Lorsque la répétition en tandem est imparfaite (variations entre motifs), l'homoplasie peut se manifester sous la forme d'allèles de taille identique mais de séquence différente. Par exemple, pour le minisatellite de l'insuline analysé par Jeffreys et collègues, 39 allèles peuvent être distingués d'après leur taille, qui correspondent à 189 codes MVR distincts (Stead 2000). De plus, l'analyse des séquences ou des cartes MVR (Jeffreys 1991) de certains minisatellites montre que des allèles de taille différente peuvent être bien plus proches, du point de vue de l'agencement et des mutations internes des motifs, que des allèles de taille identique. Il arrive donc que des allèles de même taille soient retrouvés dans des branches distinctes de l'arbre provenant d'une analyse MLVA (multiple locus VNTR analysis). En général, les classifications produites par typage de taille de répétitions en tandem varient légèrement en fonction de la combinaison de marqueurs polymorphes utilisée. Cependant, si on emploie un nombre suffisant de marqueurs, l'arbre obtenu est robuste, et en

concordance avec la phylogénie connue pour l'espèce considérée (inférée grâce à des critères morphologiques, biochimiques, et/ou de spécificité d'hôte) : nous l'avons montré pour *Bacillus anthracis* (Le Flèche 2001), *Mycobacterium tuberculosis* (Le Flèche 2002), et d'autres espèces encore en cours d'analyse au laboratoire, comme *Brucella*. Ainsi, si le choix des marqueurs polymorphes est judicieux (élimination des marqueurs soumis au phénomène d'homoplasie et des marqueurs trop polymorphes donc peu informatifs d'un point de vue évolutif), l'arbre produit par l'approche MLVA aura de bonnes chances de correspondre à la phylogénie de l'espèce bactérienne.

3.2.2 Analyse des séquences de répétitions en tandem

Une façon simple de détecter des phénomènes d'homoplasie consiste à séquencer les différents allèles de même taille, à condition que le locus ne soit pas une répétition en tandem parfaite. Le séquençage des locus VNTR peut en outre s'avérer une source très puissante d'information sur la phylogénie des espèces : il suffirait dans les cas favorables de séquencer un ou quelques locus au lieu de typer une quinzaine de locus, pour aboutir au moins à un premier niveau de classification. Cette approche se heurte toutefois à une limitation. Actuellement, très peu de moyens existent pour analyser les séquences de répétitions en tandem, et en particulier pour appréhender leur évolution. Gary Benson a proposé en 1999 un logiciel de reconstruction de l'histoire des répétitions en tandem (Benson 1999). Ce programme, décrit dans l'article présenté au chapitre 2.3.2 (Denoeud 2003), cherche à contracter à moindre coût une répétition en tandem pour remonter à un motif ancestral. Cet algorithme se base sur l'hypothèse que lorsque plusieurs motifs ou groupes de motifs arborent les mêmes mutations internes, ils sont très probablement issus d'une duplication. Cette approche, même si elle a un intérêt pour quantifier l'hétérogénéité interne d'allèles de minisatellites considérés individuellement (nous le verrons au chapitre 3.3), ne peut pas s'appliquer à la comparaison de plusieurs allèles. Elemento et collègues se sont intéressés plus particulièrement à la reconstruction de l'histoire des duplications ayant conduit d'un gène ancestral à différents gènes paralogues répétés en tandem, comme c'est le cas pour la région «TRGV » (région variable de la chaîne gamma des récepteurs des lymphocytes T) constituée de 9 gènes (Elemento 2002a ; Elemento 2002b) : ces programmes, d'une grande efficacité, s'intéressent encore une fois aux événements de duplication survenus dans un seul allèle et ne s'appliquent donc pas à la comparaison d'allèles.

Actuellement, plusieurs projets de séquençage de minisatellites sont menés au laboratoire, et l'analyse des allèles est effectuée à la main, après un pré-traitement utilisant les outils de Gary Benson [http://tandem.bu.edu/cgi-bin/trdb/trdb.exe]. Il s'agit, comme pour les cartes MVR, de coder les minisatellites par la succession des différents types d'unités répétés qu'ils contiennent : voir Figure 26.

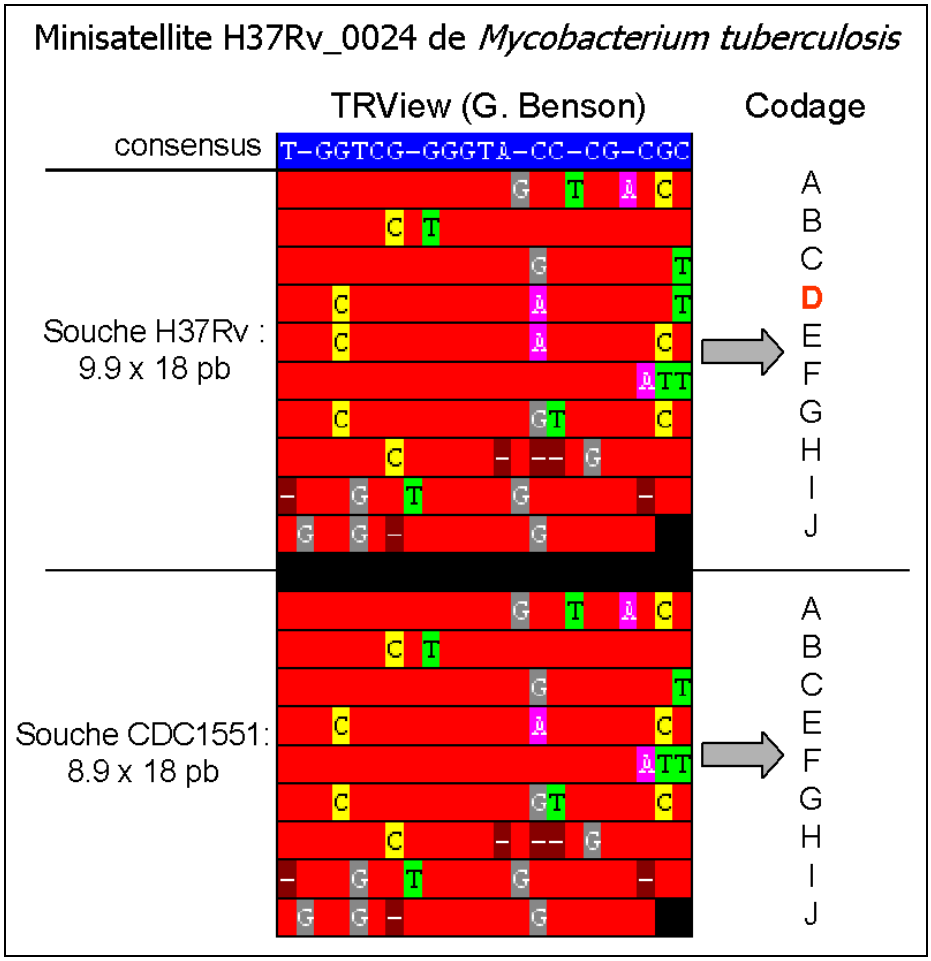


Figure 26 : Codage d'allèles d'un minisatellite de *M. tuberculosis* : à chaque motif différent est attribué un code, de A à J. Les deux allèles diffèrent par la présence/absence du motif D.

Ensuite, si le nombre de motifs possibles n'est pas trop grand et que le nombre d'allèles comparés n'est pas excessif, il est possible de reconstruire manuellement les événements évolutifs les plus probables ayant conduit d'un allèle à un autre (duplications, délétions, mutations ponctuelles dans des motifs), ou tout au moins de se faire une idée, visuellement, sur la proximité des allèles. De telles études devraient fournir des pistes sur les mécanismes de mutation des minisatellites bactériens, encore très peu étudiés. On peut espérer que l'analyse manuelle d'un certain nombre de locus minisatellites bactériens permettra de faire des généralisations sur les mécanismes d'évolution de ces structures, afin de créer un programme de comparaison automatique d'allèles. Il semble en effet prématuré de vouloir traiter automatiquement la phylogénie des répétitions en tandem, alors que les mécanismes sous-jacents restent énigmatiques et les données encore peu nombreuses.

Quelques tentatives ont toutefois été faites pour créer des algorithmes de phylogénie des répétitions en tandem, en utilisant des simplifications du problème. Ainsi, Bérard & Rivals proposent un algorithme d'alignement d'allèles de minisatellites (cartes MVR). Malgré le fait que ce programme n'autorise que des événements de nature simple (il ne considère que les délétions et amplifications d'une seule unité à la fois, et ne tient pas compte de la proximité

entre les différents types de motifs), les distances entre allèles qu'il a produites pour le minisatellite humain MSY1 ont permis de reconstruire une phylogénie concordante avec celle issue d'autres marqueurs (Berard 2003). Le minisatellite MSY1, situé sur le chromosome Y, correspond à une situation haploïde : seuls des événements intra-alléliques peuvent y survenir, ce qui est également le cas pour les minisatellites bactériens. Il serait intéressant d'appliquer cet algorithme aux séquences obtenues dans notre laboratoire pour des minisatellites de différentes espèces bactériennes (*Brucella*, *Staphylococcus aureus*…). Cependant, les études déjà faites manuellement pour certains locus montrent qu'une généralisation est difficile, et il est donc probable que ce modèle simple ne donne pas toujours des résultats satisfaisants. En outre, il n'est clairement pas adapté pour ceux des minisatellites humains localisés sur les autosomes qui peuvent subir des événements de recombinaison inter-allélique.

3.3 Prédiction du polymorphisme

3.3.1 Critères de séquence corrélés au polymorphisme

Pouvoir prédire le polymorphisme des répétitions en tandem à partir de la séquence d'un seul allèle constituerait une avancée majeure, dans la mesure où elle permettrait d'éliminer les étapes de typages préliminaires, parfois longues et coûteuses, visant à évaluer le polymorphisme des locus. Nous avons montré que certaines caractéristiques de séquence sont corrélées au polymorphisme des minisatellites humains : %GC et HistoryR (Denoeud 2003). Ces deux critères, même s'ils améliorent la sélection de minisatellites polymorphes, ne permettent pas de s'affranchir totalement de l'étape de typage préliminaire, puisque les prédictions ne produisent que 60% environ de minisatellites effectivement polymorphes dans le lot de minisatellites humains que nous avons testé. En outre, ils passent à côté d'un certain nombre de locus d'intérêt (30%). Par exemple, même si la majorité des minisatellites hypermutables actuellement connus est riche en GC (Vergnaud 2000), certains sont riches en AT et seraient donc éliminés par une telle sélection. Il faut noter que les quelques minisatellites riches en AT connus, comme FRA16B (Yu 1997), ne possèdent pas suffisamment de variants internes pour qu'on puisse étudier les événements de mutation qui s'y produisent. Il serait pourtant important de vérifier qu'ils sont bien soumis aux mêmes mécanismes que les minisatellites riches en GC. De façon intéressante, le meilleur critère de prédiction du polymorphisme des minisatellites humains, HistoryR, est aussi le plus complexe : il reflète la facilité avec laquelle on peut remonter depuis la répétition en tandem qu'on observe jusqu'à un motif « ancestral » unique, par des événements de contraction. Même si ce paramètre est obtenu à partir d'un modèle d'évolution des répétitions en tandem relativement simple, il reste plus informatif que des données de séquence brutes. On peut donc espérer que si des programmes plus élaborés voient le jour, ils fourniront de meilleurs prédicteurs du polymorphisme.

Chez les bactéries, nous avons également recherché des critères de séquence corrélés au polymorphisme des répétitions en tandem (Le Flèche 2001 ; Denoeud 2004). Nous avons effectué cette recherche sur des répétitions en tandem de longueur totale quelconque, donc généralement bien inférieure à celle des minisatellites que nous avions étudiés chez l'Homme. En effet, les minisatellites de grande taille (plusieurs centaines de paires de bases et plusieurs dizaines de motifs) sont en nombre trop faible dans certains génomes bactériens et nous avons donc étendu l'analyse aux répétitions ne comptant que quelques unités répétées. Pour des nombres d'unités faibles, la valeur de HistoryR a de fortes chances d'être égale à 1 (le coût maximal de reconstruction est égal au coût minimal) : ce critère est alors peu informatif, ce qui explique qu'il ne soit pas corrélé au polymorphisme dans ces analyses. Deux caractéristiques de séquence plus « classiques » sont en revanche corrélées au polymorphisme : la conservation interne et le nombre de copies. Etant donné que nous avons considéré des répétitions en tandem de taille assez restreinte, il n'est pas étonnant que des critères influant sur l'instabilité des microsatellites soient corrélés au polymorphisme dans cet échantillon. Cependant, la qualité prédictive de tels critères reste très variable selon les espèces bactériennes considérées : ils ne sont donc pas satisfaisants pour étudier de nouvelles espèces. Ainsi, la meilleure approche pour prédire le polymorphisme des répétitions en tandem bactériennes reste la comparaison de souches. Cette approche devrait être de plus en plus facile à mettre en œuvre car il est probable que nous disposions dans un avenir proche, pour la quasi-totalité des espèces bactériennes d'intérêt médical et/ou économique, de la séquence génomique de plusieurs souches.

3.3.2 Mécanismes de mutation

La prédiction du polymorphisme, comme la phylogénie basée sur les répétitions en tandem, profiterait d'une meilleure compréhension des mécanismes de mutation sous-jacents. Comme je l'ai décrit dans l'introduction, plusieurs types de mécanismes d'instabilité (méiotiques ou mitotiques) sont invoqués pour les minisatellites humains. Selon les locus, ces différents événements surviennent dans des proportions variables, encore impossibles à prédire. Il est probable que pour certains mécanismes, la séquence des répétitions en tandem joue un rôle, tandis que pour d'autres, seul l'environnement ait une importance. Ainsi, la compréhension des mécanismes de mutation des minisatellites humains, qui ont pourtant été largement étudiés, reste très partielle. Il semble que nous soyons encore dans une phase où les nouvelles données obtenues compliquent notre vision des choses plus qu'elles ne la simplifient.

Chez les bactéries, organismes haploïdes où les événements de recombinaison inter-alléliques ne peuvent pas se produire (sauf cas de transferts horizontaux), les mécanismes de mutation des minisatellites n'ont quasiment pas été étudiés. Ces mécanismes pourraient impliquer, comme chez l'Homme, des glissements lors de la réplication et la réparation de cassures double-brin par l'invasion de la chromatide sœur. On peut espérer que des investigations seront menées prochainement dans ce domaine.

Par ailleurs, lorsque les mécanismes d'évolution des répétitions en tandem seront mieux élucidés, il faudra encore être en mesure, afin d'améliorer l'étude des séquences pour la prédiction du polymorphisme et la phylogénie, de les modéliser *in silico*. Il s'agit là d'une tâche bioinformatique ambitieuse.

Bibliographie

Aach, J., Bulyk, M. L., Church, G. M., Comander, J., Derti, A. et Shendure, J. 2001. Computational comparison of two draft sequences of the human genome. *Nature* **409:** 856-9.

Achaz, G., Rocha, E. P., Netter, P. et Coissac, E. 2002. Origin and fate of repeats in bacteria. *Nucleic Acids Res* **30:** 2987-94.

Achtman, M., Zurth, K., Morelli, G., Torrea, G., Guiyoule, A. et Carniel, E. 1999. *Yersinia pestis*, the cause of plague, is a recently emerged clone of *Yersinia pseudotuberculosis*. *Proc Natl Acad Sci U S A* **96:** 14043-8.

Aharoni, A., Baran, N. et Manor, H. 1993. Characterization of a multisubunit human protein which selectively binds single stranded d(GA)n and d(GT)n sequence repeats in DNA. *Nucleic Acids Res* **21:** 5221-8.

Alm, R. A. *et al.* 1999. Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori*. *Nature* **397:** 176-80.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. et Lipman, D. J. 1990. Basic local alignment search tool. *J Mol Biol* **215:** 403-10.

Altschul, S. F., Boguski, M. S., Gish, W. et Wootton, J. C. 1994. Issues in searching molecular sequence databases. *Nature Genet.* **6:** 119-129.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. et Lipman, D. J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25:** 3389-402.

Amarger, V. *et al.* 1998. Analysis of the human, pig, and rat genomes supports a universal telomeric origin of minisatellite sequences. *Genomics* **52:** 62-71.

Amos, W. & Rubinstzein, D. C. 1996. Microsatellites are subject to directional evolution. *Nat Genet* **12:** 13-4.

Amos, W., Sawcer, S. J., Feakes, R. W. et Rubinsztein, D. C. 1996. Microsatellites show mutational bias and heterozygote instability. *Nat Genet* **13:** 390-1.

Appelgren, H., Cederberg, H. et Rannug, U. 1997. Mutations at the human minisatellite MS32 integrated in yeast occur with high frequency in meiosis and involve complex recombination events. *Mol. Gen. Genet.* **256:** 7-17.

Armour, J. A., Brinkworth, M. H. et Kamischke, A. 1999. Direct analysis by small-pool PCR of MS205 minisatellite mutation rates in sperm after mutagenic therapies. *Mutat Res* **445:** 73-80.

Armour, J. A. L., Patel, I., Thein, S. L., Fey, M. F. et Jeffreys, A. J. 1989. Analysis of somatic mutations at human minisatellite loci in tumors and cell lines. *Genomics* **4:** 328-334.

Ashley, T. 1994. Mammalian meiotic recombination: a reexamination. *Hum Genet* **94:** 587-93.

Atkin, N. B. 2001. Microsatellite instability. *Cytogenet Cell Genet* **92:** 177-81.

Aupetit, C., Drouet, M., Pinaud, E., Denizot, Y., Aldigier, J. C., Bridoux, F. et Cogne, M. 2000. Alleles of the alpha1 immunoglobulin gene 3' enhancer control evolution of IgA nephropathy toward renal failure. *Kidney Int* **58:** 966-71.

Autexier, C. & Greider, C. W. 1996. Telomerase and cancer: revisiting the telomere hypothesis. *Trends Biochem Sci* **21:** 387-91.

Awad, M., Pravica, V., Perrey, C., El Gamel, A., Yonan, N., Sinnott, P. J. et Hutchinson, I. V. 1999. CA repeat allele polymorphism in the first intron of the human interferon-gamma gene is associated with lung allograft fibrosis. *Hum Immunol* **60:** 343-6.

Bachtrog, D., Weiss, S., Zangerl, B., Brem, G. et Schlotterer, C. 1999. Distribution of dinucleotide microsatellites in the *Drosophila melanogaster* genome. *Mol Biol Evol* **16:** 602-10.

Bagli, M., Papassotiropoulos, A., Knapp, M., Jessen, F., Luise Rao, M., Maier, W. et Heun, R. 2000. Association between an interleukin-6 promoter and 3' flanking region haplotype and reduced Alzheimer's disease risk in a German population. *Neurosci Lett* **283:** 109-12.

Bailly, S., di Giovine, F. S. et Duff, G. W. 1993. Polymorphic tandem repeat region in interleukin-1 alpha intron 6. *Hum Genet* **91:** 85-6.

Bailly, S., Israel, N., Fay, M., Gougerot-Pocidalo, M. A. et Duff, G. W. 1996. An intronic polymorphic repeat sequence modulates interleukin-1 alpha gene regulation. *Mol Immunol* **33:** 999-1006.

Baldi, P. & Baisnee, P. F. 2000. Sequence analysis by additive scales: DNA structure for sequences and repeats of all lengths. *Bioinformatics* **16:** 865-89.

Bambara, R. A., Murante, R. S. et Henricksen, L. A. 1997. Enzymes and reactions at the eukaryotic DNA replication fork. *J Biol Chem* **272:** 4647-50.

Basil, J. B., Goodfellow, P. J., Rader, J. S., Mutch, D. G. et Herzog, T. J. 2000. Clinical significance of microsatellite instability in endometrial carcinoma. *Cancer* **89:** 1758-64.

Beckmann, J. S. & Weber, J. L. 1992. Survey of human and rat microsatellites. *Genomics* **12:** 627-631.

Bell, G. I., Horita, S. et Karam, J. H. 1984. A polymorphic locus near the insulin gene is associated with insulin-dependent diabetes mellitus. *Diabetes* **33:** 176-183.

Bennett, S. T. *et al.* 1995. Susceptibility to human type 1 diabetes at IDDM2 is determined by tandem repeat variation at the insulin gene minisatellite locus. *Nat Genet.* **9:** 284-292.

Bennett, P. 2000. Demystified ... microsatellites. *Mol Pathol* **53:** 177-83.

Benson, G. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27:** 573-580.

Benson, G. & Dong, L. 1999. Reconstructing the duplication history of a tandem repeat. *Proc Int Conf Intell Syst Mol Biol***:** 44-53.

Berard, S. & Rivals, E. 2003. Comparison of minisatellites. *J Comput Biol* **10:** 357-72.

Berg, E. S. & Olaisen, B. 1993. Characterization of the COL2A1 VNTR polymorphism. *Genomics* **16:** 350-354.

Berg, I., Cederberg, H. et Rannug, U. 2000. Tetrad analysis shows that gene conversion is the major mechanism involved in mutation at the human minisatellite MS1 integrated in *Saccharomyces cerevisiae*. *Genet Res* **75:** 1-12.

Berg, I., Neumann, R., Cederberg, H., Rannug, U. et Jeffreys, A. J. 2003. Two modes of germline instability at human minisatellite MS1 (locus D1S7): complex rearrangements and paradoxical hyperdeletion. *Am J Hum Genet* **72:** 1436-47.

Bernal, A., Ear, U. et Kyrpides, N. 2001. Genomes OnLine Database (GOLD): a monitor of genome projects world-wide. *Nucleic Acids Res* **29:** 126-7.

Biet, E., Sun, J. et Dutreix, M. 1999. Conserved sequence preference in DNA binding among recombination proteins: an effect of ssDNA secondary structure. *Nucleic Acids Res* **27:** 596-600.

Bikker, H., Baas, F. et de Vijlder, J. J. 1992. Structure and characterization of a 50 bp repeat in intron 10 of the human thyroid peroxidase gene. *Mol Cell Endocrinol* **83:** 21-8.

Bingen, E. H., Denamur, E. et Elion, J. 1994. Use of ribotyping in epidemiological surveillance of nosocomial outbreaks. *Clin Microbiol Rev* **7:** 311-27.

Bishop, A. J., Louis, E. J. et Borts, R. H. 2000. Minisatellite Variants Generated in Yeast Meiosis Involve DNA Removal During Gene Conversion. *Genetics* **156:** 7-20.

Boan, F., Rodriguez, J. M. et Gomez-Marquez, J. 1998. A non-hypervariable human minisatellite strongly stimulates in vitro intramolecular homologous recombination. *J Mol Biol* **278:** 499-505.

Bobek, L. A., Tsai, H., Biesbrock, A. R. et Levine, M. J. 1993. Molecular cloning, sequence, and specificity of expression of the gene encoding the low molecular weight human salivary mucin (MUC7). *J Biol Chem* **268:** 20563-9.

Bois, P. R. 2003. Hypermutable minisatellites, a human affair? *Genomics* **81:** 349-55.

Bonthron, D. T., Smith, S. J. et Campbell, R. 1999. Complex patterns of intragenic polymorphism at the PDGFA locus. *Hum Genet* **105:** 452-9.

Borensztajn, K., Sobrier, M. L., Fischer, A. M., Chafa, O., Amselem, S. et Tapon-Bretaudiere, J. 2003. Factor VII gene intronic mutation in a lethal factor VII deficiency: effects on splice-site selection. *Blood* **102:** 561-3.

Borstnik, B. & Pumpernik, D. 2002. Tandem repeats in protein coding regions of primate genes. *Genome Res* **12:** 909-15.

Bouzekri, N., Taylor, P. G., Hammer, M. F. et Jobling, M. A. 1998. Novel mutation processes in the evolution of a haploid minisatellite, MSY1: array homogenization without homogenization. *Hum Mol Genet* **7:** 655-9.

Bowcock, A. M., Ray, A., Erlich, H. et Sehgal, P. B. 1989. Rapid detection and sequencing of alleles in the 3' flanking region of the interleukin-6 gene. *Nucleic Acids Res* **17:** 6855-64.

Boyer, J. C., Umar, A., Risinger, J. I., Lipford, J. R., Kane, M., Yin, S., Barrett, J. C., Kolodner, R. D. et Kunkel, T. A. 1995. Microsatellite instability, mismatch repair deficiency, and genetic defects in human cancer cell lines. *Cancer Res* **55:** 6063-70.

Brinkmann, B., Klintschar, M., Neuhuber, F., Huhne, J. et Rolf, B. 1998. Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat. *Am J Hum Genet* **62:** 1408-15.

Britten, R. J. & Kohne, D. E. 1968. Repeated sequences in DNA. Hundreds of thousands of copies of DNA sequences have been incorporated into the genomes of higher organisms. *Science* **161:** 529-40.

Brohede, J. & Ellegren, H. 1999. Microsatellite evolution: polarity of substitutions within repeats and neutrality of flanking sequences. *Proc R Soc Lond B Biol Sci* **266:** 825-33.

Brook, J. D. *et al.* 1992. Molecular basis of myotonic dystrophy: expansion of a trinucleotide (CTG) repeat at the 3' end of a transcript encoding a protein kinase family member. *Cell* **68:** 799-808.

Brooks, B. P. & Fischbeck, K. H. 1995. Spinal and bulbar muscular atrophy: a trinucleotide-repeat expansion neurodegenerative disease. *Trends Neurosci* **18:** 459-61.

Brown, S. M. 2003. Bioinformatics becomes respectable. *Biotechniques* **34:** 1124-7.

Brusco, A., Saviozzi, S., Cinque, F., Bottaro, A. et DeMarchi, M. 1999. A recurrent breakpoint in the most common deletion of the Ig heavy chain locus (del A1-GP-G2-G4-E). *J Immunol* **163:** 4392-8.

Buard, J. & Vergnaud, G. 1994. Complex recombination events at the hypermutable minisatellite CEB1 (D2S90). *EMBO J.* **13:** 3203-3210.

Buard, J., Bourdet, A., Yardley, J., Dubrova, Y. et Jeffreys, A. J. 1998. Influences of array size and homogeneity on minisatellite mutation. *Embo J* **17:** 3495-502.

Buard, J., Collick, A., Brown, J. et Jeffreys, A. J. 2000. Somatic versus germline mutation processes at minisatellite CEB1 (D2S90) in humans and transgenic mice. *Genomics* **65:** 95-103.

Buchs, N., Silvestri, T., di Giovine, F. S., Chabaud, M., Vannier, E., Duff, G. W. et Miossec, P. 2000. IL-4 VNTR gene polymorphism in chronic polyarthritis. The rare allele is associated with protection against destruction. *Rheumatology (Oxford)* **39:** 1126-31.

Bugert, P., Kenck, C. et Kovacs, G. 1998. A 33 bp minisatellite repeat upstream of the 'mutated in colon cancer' gene at chromosome 5q21. *Electrophoresis* **19:** 1362-5.

Burn, T. C. *et al.* 1995. Analysis of the genomic sequence for the autosomal dominant polycystic kidney disease (PKD1) gene predicts the presence of a leucine-rich repeat. The American PKD1 Consortium (APKD1 Consortium). *Hum Mol Genet* **4:** 575-82.

Caligo, M. A., Ghimenti, C., Sensi, E., Piras, A. et Rainaldi, G. 1999. Microsatellite instability is co-selectable with gene amplification in a mammalian mutator phenotype. *Anticancer Res* **19:** 1271-5.

Campbell, T. A., Palmer, M. S., Will, R. G., Gibb, W. R., Luthert, P. J. et Collinge, J. 1996. A prion disease with a novel 96-base pair insertional mutation in the prion protein gene. *Neurology* **46:** 761-6.

Campuzano, V. *et al.* 1996. Friedreich's ataxia: autosomal recessive disease caused by an intronic GAA triplet repeat expansion. *Science* **271:** 1423-7.

Cancel, G. *et al.* 1995. Marked phenotypic heterogeneity associated with expansion of a CAG repeat sequence at the spinocerebellar ataxia 3/Machado-Joseph disease locus. *Am J Hum Genet* **57:** 809-16.

Castelo, A. T., Martins, W. et Gao, G. R. 2002. TROLL--tandem repeat occurrence locator. *Bioinformatics* **18:** 634-6.

Chaillet, J. R., Bader, D. S. et Leder, P. 1995. Regulation of genomic imprinting by gametic and embryonic processes. *Genes Dev* **9:** 1177-87.

Chang, D. K., Metzgar, D., Wills, C. et Boland, C. R. 2001. Microsatellites in the eukaryotic DNA mismatch repair genes as modulators of evolutionary mutation rate. *Genome Res* **11:** 1145-6.

Charlesworth, B., Sniegowski, P. et Stephan, W. 1994. The evolutionary dynamics of repetitive DNA in eukaryotes. *Nature* **371:** 215-220.

Chinen, K., Takahashi, E. et Nakamura, Y. 1996. Isolation and mapping of a human gene (SEC14L), partially homologous to yeast SEC14, that contains a variable number of tandem repeats (VNTR) site in its 3' untranslated region. *Cytogenet Cell Genet* **73:** 218-23.

Chou, P. Y. & Fasman, G. D. 1978. Prediction of the secondary structure of proteins from their amino acid sequence. *Adv Enzymol Relat Areas Mol Biol* **47:** 45-148.

Citti, C. & Rosengarten, R. 1997. *Mycoplasma* genetic variation and its implication for pathogenesis. *Wien Klin Wochenschr* **109:** 562-8.

Cohen, H., Sears, D. D., Zenvirth, D., Hieter, P. et Simchen, G. 1999. Increased instability of human CTG repeat tracts on yeast artificial chromosomes during gametogenesis. *Mol Cell Biol* **19:** 4153-8.

Cole, S. T. *et al.* 1998. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* **393:** 537-544.

Cole, S. T., Supply, P. et Honore, N. 2001. Repetitive sequences in *Mycobacterium leprae* and their impact on genome plasticity. *Lepr Rev* **72:** 449-61.

Csink, A. K. & Henikoff, S. 1998. Something from nothing: the evolution and utility of satellite repeats. *Trends Genet* **14:** 200-4.

Davies, P. A., Pistis, M., Hanna, M. C., Peters, J. A., Lambert, J. J., Hales, T. G. et Kirkness, E. F. 1999. The 5-HT3B subunit is a major determinant of serotonin-receptor function. *Nature* **397:** 359-63.

De Bolle, X., Bayliss, C. D., Field, D., van de Ven, T., Saunders, N. J., Hood, D. W. et Moxon, E. R. 2000. The length of a tetranucleotide repeat tract in *Haemophilus influenzae* determines the phase variation rate of a gene with homology to type III DNA methyltransferases. *Mol Microbiol* **35:** 211-22.

De Fonzo, V., Bersani, E., Aluffi-Pentini, F., Castrignano, T. et Parisi, V. 1998. Are only repeated triplets guilty? *J Theor Biol* **194:** 125-42.

Debailleul, V., Laine, A., Huet, G., Mathon, P., d'Hooghe, M. C., Aubert, J. P. et Porchet, N. 1998. Human mucin genes MUC2, MUC3, MUC4, MUC5AC, MUC5B, and MUC6 express stable and extremely large mRNAs and exhibit a variable length polymorphism. An improved method to analyze large mRNAs. *J Biol Chem* **273:** 881-90.

Debrauwère, H., Buard, J., Tessier, J., Aubert, D., Vergnaud, G. et Nicolas, A. 1999. Meiotic instability of human minisatellite CEB1 in yeast requires DNA double-strand breaks. *Nat Genet* **23:** 367-71.

Deloukas, P. *et al.* 2001. The DNA sequence and comparative analysis of human chromosome 20. *Nature* **414:** 865-71.

Denizot, Y., Pinaud, E., Aupetit, C., Le Morvan, C., Magnoux, E., Aldigier, J. C. et Cogne, M. 2001. Polymorphism of the human alpha1 immunoglobulin gene 3' enhancer hs1,2 and its relation to gene expression. *Immunology* **103:** 35-40.

Denney, R. M., Koch, H. et Craig, I. W. 1999. Association between monoamine oxidase A activity in human male skin fibroblasts and genotype of the MAOA promoter-associated variable number tandem repeat. *Hum Genet* **105:** 542-51.

Denoeud, F., Vergnaud, G. et Benson, G. 2003. Predicting Human Minisatellite Polymorphism. *Genome Res* **13:** 856-867.

Denoeud, F. & Vergnaud, G. 2004. Identification of polymorphic tandem repeats by direct comparison of genome sequence from different bacterial strains : a Web-based ressource. *BMC Bioinformatics* **5:** 4.

Desseyn, J. L., Guyonnet-Duperat, V., Porchet, N., Aubert, J. P. et Laine, A. 1997. Human mucin gene MUC5B, the 10.7-kb large central exon encodes various alternate subdomains resulting in a super-repeat. Structural evidence for a 11p15.5 gene family. *J Biol Chem* **272:** 3168-78.

Desseyn, J. L., Buisine, M. P., Porchet, N., Aubert, J. P., Degand, P. et Laine, A. 1998. Evolutionary history of the 11p15 human mucin gene family. *J Mol Evol* **46:** 102-6.

Desseyn, J. L., Rousseau, K. et Laine, A. 1999. Fifty-nine bp repeat polymorphism in the uncommon intron 36 of the human mucin gene MUC5B. *Electrophoresis* **20:** 493-6.

Destro-Bisol, G., Belledi, M., Capelli, C., Maviglia, R. et Spedini, G. 2000. Genetic variation at the ApoB 3' HVR minisatellite locus in the Mbenzele Pygmies from the Central African Republic. *Am J Human Biol* **12:** 588-592.

Dib, C. *et al.* 1996. A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature* **380:** 152-154.

Ding, S., Larson, G. P., Foldenauer, K., Zhang, G. et Krontiris, T. G. 1999. Distinct mutation patterns of breast cancer-associated alleles of the HRAS1 minisatellite locus. *Hum Mol Genet* **8:** 515-21.

Doege, K. J., Coulter, S. N., Meek, L. M., Maslen, K. et Wood, J. G. 1997. A human-specific polymorphism in the coding region of the aggrecan gene. Variable number of tandem repeats produce a range of core protein sizes in the general population. *J Biol Chem* **272:** 13974-9.

Doucette-Stamm, L. A., Blakely, D. J., Tian, J., Mockus, S. et Mao, J. I. 1995. Population genetic study of the human dopamine transporter gene (DAT1). *Genet Epidemiol* **12:** 303-8.

Dubrova, Y. E., Jeffreys, A. J. J. et Malashenko, A. M. 1993. Mouse minisatellite mutations induced by ionizing radiation. *Nat Genet.* **5:** 92-94.

Dubrova, Y. E., Nesterov, V. N., Krouchinsky, N. G., Ostapenko, V. A., Neumann, R., Neil, D. L. et Jeffreys, A. J. 1996. Human minisatellite mutation rate after the Chernobyl accident. *Nature* **380:** 683-686.

Dubrova, Y. E., Nesterov, V. N., Krouchinsky, N. G., Ostapenko, V. A., Vergnaud, G., Giraudeau, F., Buard, J. et Jeffreys, A. J. 1997. Further evidence for elevated human minisatellite mutation rate in Belarus eight years after the Chernobyl accident. *Mut. Res.* **381:** 267-278.

Dubrova, Y. E., Plumb, M., Brown, J., Fennelly, J., Bois, P., Goodhead, D. et Jeffreys, A. J. 1998. Stage specificity, dose response, and doubling dose for mouse minisatellite germ-line mutation induced by acute radiation. *Proc Natl Acad Sci U S A* **95:** 6251-5.

Dubrova, Y. E., Grant, G., Chumak, A. A., Stezhka, V. A. et Karakasian, A. N. 2002. Elevated minisatellite mutation rate in the post-chernobyl families from ukraine. *Am J Hum Genet* **71:** 801-9.

Dubrova, Y. E., Bersimbaev, R. I., Djansugurova, L. B., Tankimanova, M. K., Mamyrbaeva, Z., Mustonen, R., Lindholm, C., Hulten, M. et Salomaa, S. 2002. Nuclear weapons tests and human germline mutation rate. *Science* **295:** 1037.

Dunham, I. *et al.* 1999. The DNA sequence of human chromosome 22. *Nature* **402:** 489-95.

Dutreix, M. 1997. (GT)n repetitive tracts affect several stages of RecA-promoted recombination. *J Mol Biol* **273:** 105-13.

Eichler, E. E., Holden, J. J. A., Popovich, B. W., Reiss, A. L., Snow, K., Thibodeau, S. N., Richards, C. S., Ward, P. A. et Nelson, D. L. 1994. Length of uninterrupted CGG repeats determines instability in the FMR1 gene. *Nature Genet.* **8:** 88-94.

Elemento, O. & Gascuel, O. 2002a. An efficient and accurate distance based algorithm to reconstruct tandem duplication trees. *Bioinformatics* **18 Suppl 2:** S92-S99.

Elemento, O., Gascuel, O. et Lefranc, M. P. 2002b. Reconstructing the duplication history of tandemly repeated genes. *Mol Biol Evol* **19:** 278-88.

Ellegren, H., Lindgren, G., Primmer, C. R. et Moller, A. P. 1997. Fitness loss and germline mutations in barn swallows breeding in Chernobyl. *Nature* **389:** 593-6.

Ellegren, H. 2000a. Microsatellite mutations in the germline: implications for evolutionary inference. *Trends Genet* **16:** 551-8.

Ellegren, H. 2000b. Heterogeneous mutation processes in human microsatellite DNA sequences. *Nat Genet* **24:** 400-2.

Engelmann, K., Baldus, S. E. et Hanisch, F. G. 2001. Identification and topology of variant sequences within individual repeat domains of the human epithelial tumor mucin MUC1. *J Biol Chem* **276:** 27764-9.

Epplen, C., Melmer, G., Siedlaczck, I., Schwaiger, F. W., Maueler, W. et Epplen, J. T. 1993. On the essence of "meaningless" simple repetitive DNA in eukaryote genomes. *Exs* **67:** 29-45.

Escande, F., Aubert, J. P., Porchet, N. et Buisine, M. P. 2001. Human mucin gene MUC5AC: organization of its 5'-region and central repetitive region. *Biochem J* **358:** 763-72.

Feng, D. F. & Doolittle, R. F. 1987. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J Mol Evol* **25:** 351-60.

Fleischmann, R. D. *et al.* 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269:** 496-512.

Fohn, L. E. & Behringer, R. R. 2001. ESX1L, a novel X chromosome-linked human homeobox gene expressed in the placenta and testis. *Genomics* **74:** 105-8.

Foster, P. L. & Trimarchi, J. M. 1994. Adaptive reversion of a frameshift mutation in *Escherichia coli* by simple base deletions in homopolymeric runs. *Science* **265:** 407-409.

Fraser, C. M. *et al.* 1995. The minimal gene complement of *Mycoplasma genitalium*. *Science* **270:** 397-403.

Fraser, C. M. *et al.* 1997. Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*. *Nature* **390:** 580-6.

Fraser, C. M. *et al.* 1998. Complete genome sequence of *Treponema pallidum*, the syphilis spirochete. *Science* **281:** 375-88.

Fraser, C. M., Eisen, J. A. et Salzberg, S. L. 2000. Microbial genome sequencing. *Nature* **406:** 799-803.

Fraser, C. M., Eisen, J. A., Nelson, K. E., Paulsen, I. T. et Salzberg, S. L. 2002. The value of complete microbial genome sequencing (you get what you pay for). *J Bacteriol* **184:** 6403-5; discusion 6405.

Freudenreich, C. H., Kantrow, S. M. et Zakian, V. A. 1998. Expansion and length-dependent fragility of CTG repeats in yeast. *Science* **279:** 853-6.

Frothingham, R. & Meeker-O'Connell, W. A. 1998. Genetic diversity in the *Mycobacterium tuberculosis* complex based on variable numbers of tandem DNA repeats. *Microbiology* **144:** 1189-1196.

Fu, Y.-H. *et al.* 1991. Variation of the CGG repeat at the Fragile X site results in genetic instability: resolution of the Sherman paradox. *Cell* **67:** 1047-1058.

Fujisawa, T., Ikegami, H., Kawaguchi, Y., Yamato, E., Nakagawa, Y., Shen, G. Q., Fukuda, M. et Ogihara, T. 1999. Length rather than a specific allele of dinucleotide repeat in the 5' upstream region of the aldose reductase gene is associated with diabetic retinopathy. *Diabet Med* **16:** 1044-7.

Gallegos-Arreola, M., Rivas-Solis, F., Flores-Martinez, S., Zuniga-Gonzalez, G., Sandoval-Ramirez, L., Cantu-Garza, J. M., Ranaji, C., Figuera, L., Moran-Moguel, M. C. et Sanchez Corona, J. 1999. Linkage disequilibrium between IDUA kpnI-VNTR haplotype in Mexican patients with MPS-I. *Arch Med Res* **30:** 375-9.

Garcia, E., Carvalho, F., Amorim, A. et David, L. 1997. MUC6 gene polymorphism in healthy individuals and in gastric cancer patients from northern Portugal. *Cancer Epidemiol Biomarkers Prev* **6:** 1071-4.

Gardner, M. J. *et al.* 1998. Chromosome 2 sequence of the human malaria parasite *Plasmodium falciparum*. *Science* **282:** 1126-32.

Garza, J. C., Slatkin, M. et Freimer, N. B. 1995. Microsatellite allele frequencies in humans and chimpanzees, with implications for constraints on allele size. *Mol Biol Evol* **12:** 594-603.

Gebhardt, F., Zanker, K. S. et Brandt, B. 1999. Modulation of epidermal growth factor receptor gene transcription by a polymorphic dinucleotide repeat in intron 1. *J Biol Chem* **274:** 13176-80.

Gebhardt, F., Burger, H. et Brandt, B. 2000. Modulation of EGFR gene transcription by secondary structures, a polymorphic repetitive sequence and mutations--a link between genetics and epigenetics. *Histol Histopathol* **15:** 929-36.

Gendler, S., Taylor-Papadimitriou, J., Duhig, T., Rothbard, J. et Burchell, J. 1988. A highly immunogenic region of a human polymorphic epithelial mucin expressed by carcinomas is made up of tandem repeats. *J Biol Chem* **263:** 12820-3.

Gendrel, C. G., Boulet, A. et Dutreix, M. 2000. (CA/GT)(n) microsatellites affect homologous recombination during yeast meiosis. *Genes Dev* **14:** 1261-8.

Genereux, D. P. & Logsdon, J. M., Jr. 2003. Much ado about bacteria-to-vertebrate lateral gene transfer. *Trends Genet* **19:** 191-5.

Giraud, T., Fortini, D., Levis, C. et Brygoo, Y. 1998. The minisatellite MSB1, in the fungus *Botrytis cinerea*, probably mutates by slippage. *Mol Biol Evol* **15:** 1524-31.

Giraudeau, F., Aubert, D., Young, I., Horsley, S., Knight, S., Kearney, L., Vergnaud, G. et Flint, J. 1997. Molecular-cytogenetic detection of a deletion of 1p36.3 leads to a revised estimate of the frequency of subtelomeric rearrangements in idiopathic mental retardation. *J. Med. Genet.* **34:** 314-317.

Giraudeau, F. *et al.* 2001. Use of a set of highly polymorphic minisatellite probes for the identification of cryptic 1p36.3 deletions in a large collection of patients with idiopathic mental retardation : three new cases. *J Med Genet* **38:** 121-125.

Glew, M. D., Baseggio, N., Markham, P. F., Browning, G. F. et Walker, I. D. 1998. Expression of the pMGA genes of *Mycoplasma gallisepticum* is controlled by variation in the GAA trinucleotide repeat lengths within the 5' noncoding regions. *Infect Immun* **66:** 5833-41.

Goltsov, A. A., Eisensmith, R. C., Konecki, D. S., Lichter-Konecki, U. et Woo, S. L. 1992. Associations between mutations and a VNTR in the human phenylalanine hydroxylase gene. *Am J Hum Genet* **51:** 627-36.

Gonzalez-Conejero, R., Lozano, M. L., Rivera, J., Corral, J., Iniesta, J. A., Moraleda, J. M. et Vicente, V. 1998. Polymorphisms of platelet membrane glycoprotein Ib associated with arterial thrombotic disease. *Blood* **92:** 2771-6.

Goodier, J. L. & Davidson, W. S. 1998. Characterization of novel minisatellite repeat loci in Atlantic salmon (*Salmo salar*) and their phylogenetic distribution. *J Mol Evol* **46:** 245-55.

Gordenin, D. A., Kunkel, T. A. et Resnick, M. A. 1997. Repeat expansion--all in a flap? *Nat Genet* **16:** 116-8.

Gravekamp, C., Rosner, B. et Madoff, L. C. 1998. Deletion of repeats in the alpha C protein enhances the pathogenicity of group B streptococci in immune mice. *Infect Immun* **66:** 4347-54.

Grillot, I. 2003. Small-angle neutron scattering study of a world-wide known emulsion: Le Pastis. *Colloids and Surfaces A: Physicochem. Eng. Aspects* **225:** 153-60.

Guerin, M., Robichon, N., Geiselmann, J. et Rahmouni, A. R. 1998. A simple polypyrimidine repeat acts as an artificial Rho-dependent terminator in vivo and in vitro. *Nucleic Acids Res* **26:** 4895-900.

Gum, J. R., Byrd, J. C., Hicks, J. W., Toribara, N. W., Lamport, D. T. et Kim, Y. S. 1989. Molecular cloning of human intestinal mucin cDNAs. Sequence analysis and evidence for genetic polymorphism. *J Biol Chem* **264:** 6480-7.

Gum, J. R., Hicks, J. W., Swallow, D. M., Lagace, R. L., Byrd, J. C., Lamport, D. T., Siddiki, B. et Kim, Y. S. 1990. Molecular cloning of cDNAs derived from a novel human intestinal mucin gene. *Biochem Biophys Res Commun* **171:** 407-15.

Gyapay, G., Morisette, J., Vignal, A., Dib, C., Fizames, C., Milasseau, P., Marc, S., Bernardi, G., Lathrop, M. et Weissenbach, J. 1994. The 1993-94 Généthon human genetic linkage map. *Nature Genet.* **7:** 246-339.

Gyllensten, U. B., Jakobsson, S., Temrin, H. et Wilson, A. C. 1989. Nucleotide sequence and genomic organization of bird minisatellites. *Nucleic Acids Res* **17:** 2203-14.

Haber, J. E. & Louis, E. J. 1998. Minisatellite origins in yeast and humans. *Genomics* **48:** 132-5.

Hancock, J. M. & Santibanez-Koref, M. F. 1998. Trinucleotide expansion diseases in the context of micro- and minisatellite evolution, Hammersmith Hospital, April 1-3, 1998. *Embo J* **17:** 5521-4.

Handt, O., Sutherland, G. R. et Richards, R. I. 2000. Fragile sites and minisatellite repeat instability. *Mol Genet Metab* **70:** 99-105.

Harr, B. & Schlotterer, C. 2000. Long microsatellite alleles in *Drosophila melanogaster* have a downward mutation bias and short persistence times, which cause their genome-wide underrepresentation. *Genetics* **155:** 1213-20.

Hartmann, C., Johnk, L., Sasaki, H., Jenkins, R. B. et Louis, D. N. 2002. Novel PLA2G4C polymorphism as a molecular diagnostic assay for 19q loss in human gliomas. *Brain Pathol* **12:** 178-82.

Hattori, M. *et al.* 2000. The DNA sequence of human chromosome 21. The chromosome 21 mapping and sequencing consortium. *Nature* **405:** 311-9.

Hauth, A. M. & Joseph, D. A. 2002. Beyond tandem repeats: complex pattern structures and distant regions of similarity. *Bioinformatics* **18 Suppl 1:** S31-7.

Hayette, S., Gadoux, M., Martel, S., Bertrand, S., Tigaud, I., Magaud, J. P. et Rimokh, R. 1998. FLRG (follistatin-related gene), a new target of chromosomal rearrangement in malignant blood disorders. *Oncogene* **16:** 2949-54.

He, Q., Cederberg, H., Armour, J. A., May, C. A. et Rannug, U. 1999. Cis-regulation of inter-allelic exchanges in mutation at human minisatellite MS205 in yeast. *Gene* **232:** 143-53.

He, Q., Cederberg, H. et Rannug, U. 2002. The influence of sequence divergence between alleles of the human MS205 minisatellite incorporated into the yeast genome on length-mutation rates and lethal recombination events during meiosis. *J Mol Biol* **319:** 315-27.

Hedenskog, M., Sjogren, M., Cederberg, H. et Rannug, U. 1997. Induction of germline-length mutations at the minisatellites PC-1 and PC-2 in male mice exposed to polychlorinated biphenyls and diesel exhaust emissions. *Environ Mol Mutagen* **30:** 254-9.

Heilig, R. *et al.* 2003. The DNA sequence and analysis of human chromosome 14. *Nature* **421:** 601-7.

Heise, C. E. *et al.* 2000. Characterization of the human cysteinyl leukotriene 2 receptor. *J Biol Chem* **275:** 30531-6.

Henderson, E., Hardin, C. C., Walk, S. K., Tinoco, I., Jr. et Blackburn, E. H. 1987. Telomeric DNA oligonucleotides form novel intramolecular structures containing guanine-guanine base pairs. *Cell* **51:** 899-908.

Henderson, S. T. & Petes, T. D. 1992. Instability of simple sequence DNA in *Saccharomyces cerevisiae*. *Molecular and Cellular Biology* **12:** 2749-2757.

Henderson, I. R., Owen, P. et Nataro, J. P. 1999. Molecular switches--the ON and OFF of bacterial phase variation. *Mol Microbiol* **33:** 919-32.

Henikoff, S. 2001. Chromosomes on the move. *Trends Genet* **17:** 689-90.

Heringa, J. 1998. Detection of internal repeats: how common are they? *Curr Opin Struct Biol* **8:** 338-45.

Hewett, D. R., Handt, O., Hobson, L., Mangelsdorf, M., Eyre, H. J., Baker, E., Sutherland, G. R., Schuffenhauer, S., Mao, J. I. et Richards, R. I. 1998. FRA10B structure reveals common elements in repeat expansion and chromosomal fragile site genesis. *Mol Cell* **1:** 773-81.

Heyer, E., Puymirat, J., Dieltjes, P., Bakker, E. et de Knijff, P. 1997. Estimating Y chromosome specific microsatellite mutation frequencies using deep rooting pedigrees. *Hum Mol Genet* **6:** 799-803.

Higuchi, S., Nakamura, Y. et Saito, S. 2002. Characterization of a VNTR polymorphism in the coding region of the CEL gene. *J Hum Genet* **47:** 213-5.

Hillier, L. W. *et al.* 2003. The DNA sequence of human chromosome 7. *Nature* **424:** 157-64.

Himmelreich, R., Hilbert, H., Plagens, H., Pirkl, E., Li, B. C. et Herrmann, R. 1996. Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. *Nucleic Acids Res* **24:** 4420-49.

Hofferbert, S., Schanen, N. C., Chehab, F. et Francke, U. 1997. Trinucleotide repeats in the human genome : size distributions for all possible triplets and detection of expanded disease alleles in a group of Huntington disease individuals by the Repeat Expansion Detection method. *Human Molec. Genet.* **6:** 77-83.

Hogenesch, J. B., Ching, K. A., Batalov, S., Su, A. I., Walker, J. R., Zhou, Y., Kay, S. A., Schultz, P. G. et Cooke, M. P. 2001. A comparison of the Celera and Ensembl predicted gene sets reveals little overlap in novel genes. *Cell* **106:** 413-5.

Hollingshead, S. K., Fischetti, V. A. et Scott, J. R. 1987. Size variation in group A streptococcal M protein is generated by homologous recombination between intragenic repeats. *Mol Gen Genet* **207:** 196-203.

Holmlund, G. & Lindblom, B. 1998. Different ancestor alleles: a reason for the bimodal fragment size distribution in the minisatellite D2S44 (YNH24). *Eur J Hum Genet* **6:** 597-602.

Hood, D. W., Deadman, M. E., Jennings, M. P., Bisercic, M., Fleischmann, R. D., Venter, J. C. et Moxon, E. R. 1996. DNA repeats identify novel virulence genes in Haemophilus influenzae. *Proc Natl Acad Sci U S A* **93:** 11121-5.

Horowitz, H. & Haber, J. E. 1984. Subtelomeric regions of yeast chromosomes contain a 36 base-pair tandemly repeated sequence. *Nucleic Acids Res* **12:** 7105-21.

Huang, L.-S. & Breslow, J. L. 1987. A unique AT-rich hypervariable minisatellite 3' to the ApoB gene defines a high information restriction fragment length polymorphism. *J. Biol. Chem.* **262:** 8952-8955.

Imbert, G. *et al.* 1996. Cloning of the gene for spinocerebellar ataxia 2 reveals a locus with high sensitivity to expanded CAG/glutamine repeats. *Nat Genet* **14:** 285-91.

Irshaid, N. M., Chester, M. A. et Olsson, M. L. 1999. Allele-related variation in minisatellite repeats involved in the transcription of the blood group ABO gene. *Transfus Med* **9:** 219-26.

Ito, K. *et al.* 2002. A variable number of tandem repeats in the serotonin transporter gene does not affect the antidepressant response to fluvoxamine. *Psychiatry Res* **111:** 235-9.

Jackson, P. J. *et al.* 1997. Characterization of the variable-number tandem repeats in vrrA from different *Bacillus anthracis* isolates. *Appl Environ Microbiol* **63:** 1400-5.

Jacob, S. & Praz, F. 2002. DNA mismatch repair defects: role in colorectal carcinogenesis. *Biochimie* **84:** 27-47.

Jakupciak, J. P. & Wells, R. D. 2000. Gene conversion (recombination) mediates expansions of CTG.CAG repeats. *J Biol Chem* **275:** 40003-13.

Jankowski, C., Nasar, F. et Nag, D. K. 2000. Meiotic instability of CAG repeat tracts occurs by double-strand break repair in yeast. *Proc Natl Acad Sci U S A* **97:** 2134-2139.

Jansen, R., Embden, J. D., Gaastra, W. et Schouls, L. M. 2002. Identification of genes that are associated with DNA repeats in prokaryotes. *Mol Microbiol* **43:** 1565-75.

Jeffreys, A. J., Wilson, V. et Thein, S. L. 1985a. Hypervariable 'minisatellite' regions in human DNA. *Nature* **314:** 67-73.

Jeffreys, A. J., Wilson, V. et Thein, S. L. 1985b. Individual-specific 'fingerprints' of human DNA. *Nature* **316:** 76-79.

Jeffreys, A. J., Royle, N. J., Wilson, V. et Wong, Z. 1988. Spontaneous mutation rates to new length alleles at tandem-repetitive hypervariable loci in human DNA. *Nature* **332:** 278-281.

Jeffreys, A. J., MacLeod, A., Tamaki, K., Neil, D. L. et Monckton, D. G. 1991. Minisatellite repeat coding as a digital approach to DNA typing. *Nature* **354:** 204-209.

Jeffreys, A. J., Tamaki, K., MacLeod, A., Monckton, D. G., Neil, D. L. et Armour, J. A. L. 1994. Complex gene conversion events in germline mutation at human minisatellites. *Nat. Genet.* **6:** 136-145.

Jeffreys, A. J. & Neumann, R. 1997. Somatic mutation processes at a human minisatellite. *Hum. Mol. Genet.* **6:** 129-136.

Jeffreys, A. J. *et al.* 1999. Human minisatellites, repeat DNA instability and meiotic recombination. *Electrophoresis* **20:** 1665-75.

Jeffreys, A. J. & Neumann, R. 2002. Reciprocal crossover asymmetry and meiotic drive in a human recombination hot spot. *Nat Genet* **31:** 267-71.

Jernigan, D. B. *et al.* 2002. Investigation of bioterrorism-related anthrax, United States, 2001: epidemiologic findings. *Emerg Infect Dis* **8:** 1019-28.

Jilma-Stohlawetz, P., Homoncik, M., Jilma, B., Knechtelsdorfer, M., Unger, P., Mannhalter, C., Santoso, S. et Panzer, S. 2003. Glycoprotein Ib polymorphisms influence platelet plug formation under high shear rates. *Br J Haematol* **120:** 652-5.

Jin, L., Macaubas, C., Hallmayer, J., Kimura, A. et Mignot, E. 1996. Mutation rate varies among alleles at a microsatellite locus: phylogenetic evidence. *Proc Natl Acad Sci U S A* **93:** 15285-8.

Jobling, M. A., Bouzekri, N. et Taylor, P. G. 1998. Hypervariable digital DNA codes for human paternal lineages: MVR-PCR at the Y-specific minisatellite, MSY1 (DYF155S1). *Hum Mol Genet* **7:** 643-53.

Johannsdottir, J. T., Jonasson, J. G., Bergthorsson, J. T., Amundadottir, L. T., Magnusson, J., Egilsson, V. et Ingvarsson, S. 2000. The effect of mismatch repair deficiency on tumourigenesis; microsatellite instability affecting genes containing short repeated sequences. *Int J Oncol* **16:** 133-9.

Jurado, L. A., Coloma, A. et Cruces, J. 1999. Identification of a human homolog of the Drosophila rotated abdomen gene (POMT1) encoding a putative protein O-mannosyl-transferase, and assignment to human chromosome 9q34.1. *Genomics* **58:** 171-80.

Kalikin, L. M., Bugeaud, E. M., Palmbos, P. L., Lyons, R. H., Jr. et Petty, E. M. 2001. Genomic characterization of human SEC14L1 splice variants within a 17q25 candidate tumor suppressor gene region and identification of an unrelated embedded expressed sequence tag. *Mamm Genome* **12:** 925-9.

Kamerbeek, J. *et al.* 1997. Simultaneous detection and strain differentiation of *Mycobacterium tuberculosis* for diagnosis and epidemiology. *J Clin Microbiol* **35:** 907-14.

Karlin, S. 1998. Global dinucleotide signatures and analysis of genomic heterogeneity. *Curr Opin Microbiol* **1:** 598-610.

Katsuyama, Y., Inoko, H., Imanishi, T., Mizuki, N., Gojobori, T. et Ota, M. 1998. Genetic relationships among Japanese, northern Han, Hui, Uygur, Kazakh, Greek, Saudi Arabian, and Italian populations based on allelic frequencies at four VNTR (D1S80, D4S43, COL2A1, D17S5) and one STR (ACTBP2) loci. *Hum Hered* **48:** 126-37.

Kayser, M. *et al.* 2000. Characteristics and frequency of germline mutations at microsatellite loci from the human Y chromosome, as revealed by direct observation in father/son pairs. *Am J Hum Genet* **66:** 1580-8.

Keim, P., Price, L. B., Klevytska, A. M., Smith, K. L., Schupp, J. M., Okinaka, R., Jackson, P. J. et Hugh-Jones, M. E. 2000. Multiple-Locus Variable-Number Tandem Repeat Analysis Reveals Genetic Relationships within *Bacillus anthracis*. *J Bacteriol* **182:** 2928-2936.

Kinarsky, L., Suryanarayanan, G., Prakash, O., Paulsen, H., Clausen, H., Hanisch, F. G., Hollingsworth, M. A. et Sherman, S. 2003. Conformational studies on the MUC1 tandem repeat glycopeptides: implication for the enzymatic O-glycosylation of the mucin protein core. *Glycobiology*.

Kirkbride, H. J., Bolscher, J. G., Nazmi, K., Vinall, L. E., Nash, M. W., Moss, F. M., Mitchell, D. M. et Swallow, D. M. 2001. Genetic polymorphism of MUC7: allele frequencies and association with asthma. *Eur J Hum Genet* **9:** 347-54.

Kodaira, M., Satoh, C., Hiyama, K. et Toyama, K. 1995. Lack of effects of atomic bomb radiation on genetic instability of tandem-repetitive elements in human germ cells. *Am. J. Hum. Genet.* **57:** 1275-1283.

Kohl, S. *et al.* 2000. Mutations in the CNGB3 gene encoding the beta-subunit of the cone photoreceptor cGMP-gated channel are responsible for achromatopsia (ACHM3) linked to chromosome 8q21. *Hum Mol Genet* **9:** 2107-16.

Kokoska, R. J., Stefanovic, L., Tran, H. T., Resnick, M. A., Gordenin, D. A. et Petes, T. D. 1998. Destabilization of yeast micro- and minisatellite DNA sequences by mutations affecting a nuclease involved in Okazaki fragment processing (rad27) and DNA polymerase delta (pol3-t). *Mol Cell Biol* **18:** 2779-88.

Kokoska, R. J., Stefanovic, L., Buermeyer, A. B., Liskay, R. M. et Petes, T. D. 1999. A mutation of the yeast gene encoding PCNA destabilizes both microsatellite and minisatellite DNA sequences. *Genetics* **151:** 511-9.

Kolpakov, R., Bana, G. et Kucherov, G. 2003. mreps: Efficient and flexible detection of tandem repeats in DNA. *Nucleic Acids Res* **31:** 3672-8.

Kominato, Y., Tsuchiya, T., Hata, N., Takizawa, H. et Yamamoto, F. 1997. Transcription of human ABO histo-blood group genes is dependent upon binding of transcription factor CBF/NF-Y to minisatellite sequence. *J Biol Chem* **272:** 25890-8.

Korotkov, E. V., Korotkova, M. A. et Tulko, J. S. 1997. Latent sequence periodicity of some oncogenes and DNA-binding protein genes. *Comput Appl Biosci* **13:** 37-44.

Kovalchuk, O., Dubrova, Y. E., Arkhipov, A., Hohn, B. et Kovalchuk, I. 2000. Wheat mutation rate after Chernobyl. *Nature* **407:** 583-4.

Krasilnikova, M. M., Samadashwily, G. M., Krasilnikov, A. S. et Mirkin, S. M. 1998. Transcription through a simple DNA repeat blocks replication elongation. *Embo J* **17:** 5095-102.

Krontiris, T. G., Devlin, B., Karp, D. D., Robert, N. J. et Risch, N. 1993. An association between the risk of cancer and mutations in the HRAS1 minisatellite locus. *N. Engl. J. Med.* **329:** 517-523.

Kruglyak, S., Durrett, R. T., Schug, M. D. et Aquadro, C. F. 1998. Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. *Proc Natl Acad Sci U S A* **95:** 10774-8.

Kurtz, S. & Schleiermacher, C. 1999. REPuter: fast computation of maximal repeats in complete genomes. *Bioinformatics* **15:** 426-7.

Kurtz, S., Ohlebusch, E., Schleiermacher, C., Stoye, J. et Giegerich, R. 2000. Computation and visualization of degenerate repeats in complete genomes. *Proc Int Conf Intell Syst Mol Biol* **8:** 228-38.

Kyte, J. & Doolittle, R. F. 1982. A simple method for displaying the hydropathic character of a protein. *J Mol Biol* **157:** 105-32.

La Spada, A. R. and J. P. Taylor. 2003. Polyglutamines placed into context. *Neuron* **38**: 681-4.

Laken, S. J. *et al.* 1997. Familial colorectal cancer in Ashkenazim due to a hypermutable tract in APC. *Nat Genet* **17:** 79-83.

Lalande, M. 2001. Imprints of disease at GNAS1. *J Clin Invest* **107:** 793-4.

Lalioti, M. D., Scott, H. S., Buresi, C., Rossier, C., Bottani, A., Morris, M. A., Malafosse, A. et Antonarakis, S. E. 1997. Dodecamer repeat expansion in cystatin B gene in progressive myoclonus epilepsy. *Nature* **386:** 847-51.

Landau, G. M., Schmidt, J. P. et Sokol, D. 2001. An algorithm for approximate tandem repeats. *J Comput Biol* **8:** 1-18.

Lander, E. S. *et al.* 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860-921.

Langdon, J. A. & Armour, J. A. 2003. Evolution and population genetics of the H-ras minisatellite and cancer predisposition. *Hum Mol Genet* **12:** 891-900.

Lanz, R. B., Wieland, S., Hug, M. et Rusconi, S. 1995. A transcriptional repressor obtained by alternative translation of a trinucleotide repeat. *Nucleic Acids Res.* **23:** 138-145.

Laplanche, J. L., Delasnerie-Laupretre, N., Brandel, J. P., Dussaucy, M., Chatelain, J. et Launay, J. M. 1995. Two novel insertions in the prion protein gene in patients with late-onset dementia. *Hum Mol Genet* **4:** 1109-11.

Larson, G. P., Ding, S., Lafreniere, R. G., Rouleau, G. A. et Krontiris, T. G. 1999. Instability of the EPM1 minisatellite. *Hum Mol Genet* **8:** 1985-8.

Le Flèche, P., Hauck, Y., Onteniente, L., Prieur, A., Denoeud, F., Ramisse, V., Sylvestre, P., Benson, G., Ramisse, F. et Vergnaud, G. 2001. A tandem repeats database for bacterial genomes: application to the genotyping of *Yersinia pestis* and *Bacillus anthracis*. *BMC Microbiol* **1:** 2.

Le Flèche, P., Fabre, M., Denoeud, F., Koeck, J. L. et Vergnaud, G. 2002. High resolution, on-line identification of strains from the *Mycobacterium tuberculosis* complex based on tandem repeat typing. *BMC Microbiol* **2:** 37.

Lee, S. & Park, M. S. 2002. Human FEN-1 can process the 5'-flap DNA of CTG/CAG triplet repeat derived from human genetic diseases by length and sequence dependent manner. *Exp Mol Med* **34:** 313-7.

Leem, S. H. *et al.* 2002. The human telomerase gene: complete genomic sequence and analysis of tandem repeat polymorphisms in intronic regions. *Oncogene* **21:** 769-77.

Leung, E., Greene, J., Ni, J., Raymond, L. G., Lehnert, K., Langley, R. et Krissansen, G. W. 1996. Cloning of the mucosal addressin MAdCAM-1 from human brain: identification of novel alternatively spliced transcripts. *Immunol Cell Biol* **74:** 490-6.

Leung, W. K., Kim, J. J., Kim, J. G., Graham, D. Y. et Sepulveda, A. R. 2000. Microsatellite instability in gastric intestinal metaplasia in patients with and without gastric cancer. *Am J Pathol* **156:** 537-43.

Levinson, G. & Gutman, G. A. 1987. High frequency of short frameshifts in poly-CA/TG tandem repeats borne by bacteriophage M13 in *Escherichia coli* K-12. *Nucleic Acids Res.* **15:** 5323-5338.

Li, X., Li, J., Harrington, J., Lieber, M. R. et Burgers, P. M. 1995. Lagging strand DNA synthesis at the eukaryotic replication fork involves binding and stimulation of FEN-1 by proliferating cell nuclear antigen. *J Biol Chem* **270:** 22109-12.

Li, T. *et al.* 1997. Association analysis of the dopamine D4 gene exon III VNTR and heroin abuse in Chinese subjects. *Mol Psychiatry* **2:** 413-6.

Li, Y., Fahima, T., Korol, A. B., Peng, J., Roder, M. S., Kirzhner, V., Beiles, A. et Nevo, E. 2000. Microsatellite diversity correlated with ecological-edaphic and genetic factors in three microsites of wild emmer wheat in North Israel. *Mol Biol Evol* **17:** 851-62.

Li, Y. C., Korol, A. B., Fahima, T., Beiles, A. et Nevo, E. 2002. Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. *Mol Ecol* **11:** 2453-65.

Lichter, J. B., Barr, C. L., Kennedy, J. L., Van Tol, H. H., Kidd, K. K. et Livak, K. J. 1993. A hypervariable segment in the human dopamine receptor D4 (DRD4) gene. *Hum Mol Genet* **2:** 767-73.

Lievers, K. J., Kluijtmans, L. A., Heil, S. G., Boers, G. H., Verhoef, P., van Oppenraay-Emmerzaal, D., den Heijer, M., Trijbels, F. J. et Blom, H. J. 2001. A 31 bp VNTR in the cystathionine beta-synthase (CBS) gene is associated with reduced CBS activity and elevated post-load homocysteine levels. *Eur J Hum Genet* **9:** 583-9.

Lin, X. *et al.* 1999. Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana*. *Nature* **402:** 761-8.

Lin, J. J., Yueh, K. C., Chang, D. C., Chang, C. Y., Yeh, Y. H. et Lin, S. Z. 2003. The homozygote 10-copy genotype of variable number tandem repeat dopamine transporter gene may confer protection against Parkinson's disease for male, but not to female patients. *J Neurol Sci* **209:** 87-92.

Liu, B. *et al.* 1995. Mismatch repair gene defects in sporadic colorectal cancers with microsatellite instability. *Nature Genet.* **9:** 48-55.

Livshits, L. A. *et al.* 2001. Children of Chernobyl Cleanup Workers do not Show Elevated Rates of Mutations in Minisatellite Alleles. *Radiat Res* **155:** 74-80.

Lopes, J., Debrauwere, H., Buard, J. et Nicolas, A. 2002. Instability of the human minisatellite CEB1 in rad27Delta and dna2-1 replication-deficient yeast cells. *Embo J* **21:** 3201-11.

Lopez, J. A., Ludwig, E. H. et McCarthy, B. J. 1992. Polymorphism of human glycoprotein Ib alpha results from a variable number of tandem repeats of a 13-amino acid sequence in the mucin-like macroglycopeptide region. Structure/function implications. *J Biol Chem* **267:** 10055-61.

Ma, P., Chen, D., Pan, J. et Du, B. 2002. Genomic polymorphism within interleukin-1 family cytokines influences the outcome of septic patients. *Crit Care Med* **30:** 1046-50.

MacNeill, S. A. 2001. DNA replication: partners in the Okazaki two-step. *Curr Biol* **11:** R842-4.

Madoff, L. C., Michel, J. L., Gong, E. W., Kling, D. E. et Kasper, D. L. 1996. Group B streptococci escape host immunity by deletion of tandem repeat elements of the alpha C protein. *Proc Natl Acad Sci U S A* **93:** 4131-6.

Maeng, J. H. & Yoon, J. B. 1998. The human PTFgamma/SNAP43 gene: structure, chromosomal location, and identification of a VNTR in 5'-UTR. *J Biochem (Tokyo)* **124:** 23-7.

Mahtani, M. M. & Willard, H. F. 1993. A polymorphic X-linked tetranucleotide repeat locus displaying a high rate of new mutation: implications for mechanisms of mutation at short tandem repeat loci. *Human Molecular Genetics* **2:** 431-437.

Maiden, M. C. *et al.* 1998. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc Natl Acad Sci U S A* **95:** 3140-5.

Majewski, J. & Ott, J. 2000. GT repeats are associated with recombination on human chromosome 22. *Genome Res* **10:** 1108-14.

Maleki, S., Cederberg, H. et Rannug, U. 1997. Mutations occurring at the human minisatellite MS1 integrated in haploid yeast are similar to MS1 mutations in humans. *Mol Gen Genet* **254:** 37-42.

Malik, H. S. & Henikoff, S. 2001. Adaptive evolution of Cid, a centromere-specific histone in Drosophila. *Genetics* **157:** 1293-8.

Manuck, S. B., Flory, J. D., Ferrell, R. E., Mann, J. J. et Muldoon, M. F. 2000. A regulatory polymorphism of the monoamine oxidase-A gene may be associated with variability in aggression, impulsivity, and central nervous system serotonergic responsivity. *Psychiatry Res* **95:** 9-23.

Marcotte, E. M., Pellegrini, M., Yeates, T. O. et Eisenberg, D. 1999. A census of protein repeats. *J Mol Biol* **293:** 151-60.

Markowitz, S. *et al.* 1995. Inactivation of the type II TGF-beta receptor in colon cancer cells with microsatellite instability. *Science* **268:** 1336-8.

Martin-Farmer, J. & Janssen, G. R. 1999. A downstream CA repeat sequence increases translation from leadered and unleadered mRNA in *Escherichia coli*. *Mol Microbiol* **31:** 1025-38.

Masepohl, B., Gorlitz, K. et Bohme, H. 1996. Long tandemly repeated repetitive (LTRR) sequences in the filamentous cyanobacterium *Anabaena sp.* PCC 7120. *Biochim Biophys Acta* **1307:** 26-30.

Matsunami, H., Montmayeur, J. P. et Buck, L. B. 2000. A family of candidate taste receptors in human and mouse. *Nature* **404:** 601-4.

May, C. A., Jeffreys, A. J. et Armour, J. A. L. 1996. Mutation rate heterogeneity and the generation of allele diversity at the human minisatellite MS205 (D16S309). *Human Molecular Genetics* **5:** 1823-1833.

May, C. A., Tamaki, K., Neumann, R., Wilson, G., Zagars, G., Pollack, A., Dubrova, Y. E., Jeffreys, A. J. et Meistrich, M. L. 2000. Minisatellite mutation frequency in human sperm following radiotherapy. *Mutat Res* **453:** 67-75.

Mays, P. K., Tromp, G., Kuivaniemi, H., Ryynanen, M. et Prockop, D. J. 1992. A 15 base-pair AT-rich variable number tandem repeat in the type III procollagen gene (COL3A1) as an informative marker for 2q31-2q32.3. *Matrix* **12:** 44-9.

McMurray, C. T. 1999. DNA secondary structure: a common and causative factor for expansion in human disease. *Proc Natl Acad Sci U S A* **96:** 1823-5.

Meloni, R., Albanese, V., Ravassard, P., Treilhou, F. et Mallet, J. 1998. A tetranucleotide polymorphic microsatellite, located in the first intron of the tyrosine hydroxylase gene, acts as a transcription regulatory element in vitro. *Hum Mol Genet* **7:** 423-8.

Meneveri, R., Agresti, A. et Ginelli, E. 1984. Distribution of repeated DNA families in the human genome. *Biochem Biophys Res Commun* **124:** 400-6.

Messier, W., Li, S. H. et Stewart, C. B. 1996. The birth of microsatellites. *Nature* **381:** 483.

Metzgar, D., Bytof, J. et Wills, C. 2000. Selection against frameshift mutations limits microsatellite expansion in coding DNA. *Genome Res* **10:** 72-80.

Meulenbelt, I., Bijkerk, C., De Wildt, S. C., Miedema, H. S., Breedveld, F. C., Pols, H. A., Hofman, A., Van Duijn, C. M. et Slagboom, P. E. 1999. Haplotype analysis of three polymorphisms of the COL2A1 gene and associations with generalised radiological osteoarthritis. *Ann Hum Genet* **63 (Pt 5):** 393-400.

Mill, J., Asherson, P., Browes, C., D'Souza, U. et Craig, I. 2002. Expression of the dopamine transporter gene is regulated by the 3' UTR VNTR: Evidence from brain and lymphocytes using quantitative RT-PCR. *Am J Med Genet* **114:** 975-9.

Mitas, M. 1997. Trinucleotide repeats associated with human disease. *Nucleic Acids Res* **25:** 2245-54.

Miyahara, K. *et al.* 1994. Cloning and structural characterization of the human endothelial nitric-oxide-synthase gene. *Eur J Biochem* **223:** 719-26.

Mojica, F. J., Diez-Villasenor, C., Soria, E. et Juez, G. 2000. Biological significance of a family of regularly spaced repeats in the genomes of Archaea, Bacteria and mitochondria. *Mol Microbiol* **36:** 244-6.

Mollick, J. A., Hodi, F. S., Soiffer, R. J., Nadler, L. M. et Dranoff, G. 2003. MUC1-like tandem repeat proteins are broadly immunogenic in cancer patients. *Cancer Immun* **3:** 3.

Monckton, D. G., Neumann, R., Guram, T., Fretwell, N., Tamaki, K., MacLeod, A. et Jeffreys, A. J. 1994. Minisatellite mutation rate variation associated with a flanking DNA sequence polymorphism. *Nat genetics* **8:** 162-170.

Monckton, D. G. & Caskey, C. T. 1995. Unstable triplet repeat diseases. *Circulation* **91:** 513-20.

Moore, H., Greenwell, P. W., Liu, C. P., Arnheim, N. et Petes, T. D. 1999. Triplet repeats form secondary structures that escape DNA repair in yeast. *Proc Natl Acad Sci U S A* **96:** 1504-9.

Morgante, M., Hanafey, M. et Powell, W. 2002. Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. *Nat Genet* **30:** 194-200.

Morral, N., Nunes, V., Casals, T. et Estivill, X. 1991. CA/GT microsatellite alleles within the cystic fibrosis transmembrane conductance regulator (CFTR) gene are not generated by unequal crossingover. *Genomics* **10:** 692-8.

Mout, R., Willemze, R. et Landegent, J. E. 1991. Repeat polymorphisms in the interleukin-4 gene (IL4). *Nucleic Acids Res* **19:** 3763.

Moxon, E. R., Rainey, P. B., Nowak, M. A. et Lenski, R. E. 1994. Adaptive evolution of highly mutable loci in pathogenic bacteria. *Curr Biol* **4:** 24-33.

Muckian, C., Fitzgerald, A., O'Neill, A., O'Byrne, A., Fitzgerald, D. J. et Shields, D. C. 2002. Genetic variability in the extracellular matrix as a determinant of cardiovascular risk: association of type III collagen COL3A1 polymorphisms with coronary artery disease. *Blood* **100:** 1220-3.

Murray, R. E., McGuigan, F., Grant, S. F., Reid, D. M. et Ralston, S. H. 1997. Polymorphisms of the interleukin-6 gene are associated with bone mineral density. *Bone* **21:** 89-92.

Murray, J., Buard, J., Neil, D. L., Yeramian, E., Tamaki, K., Hollies, C. R. et Jeffreys, A. J. 1999. Comparative sequence analysis of human minisatellites showing meiotic repeat instability. *Genome Res.* **9:** 130-136.

Myers, E. W., Sutton, G. G., Smith, H. O., Adams, M. D. et Venter, J. C. 2002. On the sequencing and assembly of the human genome. *Proc Natl Acad Sci U S A* **99:** 4145-6.

Nagafuchi, S. *et al.* 1994. Dentatorubral and pallidoluysian atrophy expansion of an unstable CAG trinucleotide on chromosome 12p. *Nature Genet.* **6:** 14-18.

Nakamura, Y., Lathrop, M., O'Connell, P., Leppert, M., Lalouel, J.-M. et White, R. 1988. A primary map of ten DNA markers and two serological markers for human chromosome 19. *Genomics* **3:** 67-71.

Nakamura, Y., Koyama, K. et Matsushima, M. 1998. VNTR (variable number of tandem repeat) sequences as transcriptional, translational, or functional regulators. *J Hum Genet* **43:** 149-52.

Needleman, S. B. & Wunsch, C. D. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* **48:** 443-53.

Nelson, K. E., Paulsen, I. T., Heidelberg, J. F. et Fraser, C. M. 2000. Status of genome projects for nonpathogenic bacteria and archaea. *Nat Biotechnol* **18:** 1049-54.

Neumann, B., Kubicka, P. et Barlow, D. P. 1995. Characteristics of imprinted genes. *Nat Genet* **9:** 12-13.

Nollet, S., Moniaux, N., Maury, J., Petitprez, D., Degand, P., Laine, A., Porchet, N. et Aubert, J. P. 1998. Human mucin gene MUC4: organization of its 5'-region and polymorphism of its central tandem repeat array. *Biochem J* **332 (Pt 3):** 739-48.

Ogilvie, A. D., Battersby, S., Bubb, V. J., Fink, G., Harmar, A. J., Goodwim, G. M. et Smith, C. A. 1996. Polymorphism in serotonin transporter gene associated with susceptibility to major depression. *Lancet* **347:** 731-3.

O'Hara, P. J. & Grant, F. J. 1988. The human factor VII gene is polymorphic due to variation in repeat copy number in a minisatellite. *Gene* **66:** 147-158.

Okladnova, O., Syagailo, Y. V., Tranitz, M., Stober, G., Riederer, P., Mossner, R. et Lesch, K. P. 1998. A promoter-associated polymorphic repeat modulates PAX-6 expression in human brain. *Biochem Biophys Res Commun* **248:** 402-5.

Olive, D. M. & Bean, P. 1999. Principles and applications of methods for DNA-based typing of microbial organisms. *J Clin Microbiol* **37:** 1661-9.

Olivieri, N. F. & Weatherall, D. J. 1998. The therapeutic reactivation of fetal haemoglobin. *Hum Mol Genet* **7:** 1655-8.

Onteniente, L., S. Brisse, P. T. Tassios and G. Vergnaud. 2003. Evaluation of the polymorphisms associated with tandem repeats for Pseudomonas aeruginosa strain typing. *J Clin Microbiol* **41**: 4991-7.

Orr, H. T., Chung, M., Banfi, S., Kwiatkowski, T. J., Servadio, A., Beaudet, A. L., McCall, A. E., Duvick, L. A., Ranum, L. P. W. et Zoghbi, H. Y. 1993. Expansion of an unstable trinucleotide CAG repeat in spinocerebellar ataxia type 1. *Nature Genet.* **4:** 221-226.

Paques, F., Richard, G. F. et Haber, J. E. 2001. Expansions and contractions in 36-bp minisatellites by gene conversion in yeast. *Genetics* **158:** 155-66.

Parniewski, P. & Staczek, P. 2002. Molecular mechanisms of TRS instability. *Adv Exp Med Biol* **516:** 1-25.

Pausova, Z., Morgan, K., Fujiwara, M., Bourdon, J., Goltzman, D. et Hendy, G. N. 1993. Molecular characterization of an intragenic minisatellite (VNTR) polymorphism in the human parathyroid hormone-related peptide gene in chromosome region 12p12.1-p11.2. *Genomics* **17:** 243-244.

Pearson, W. R. & Lipman, D. J. 1988. Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A* **85:** 2444-8.

Pellegrini, M., Marcotte, E. M. et Yeates, T. O. 1999. A fast algorithm for genome-wide analysis of proteins with repeated sequences. *Proteins* **35:** 440-6.

Perez-Vilar, J. & Hill, R. L. 1999. The structure and assembly of secreted mucins. *J Biol Chem* **274:** 31751-4.

Petes, T. D., Greenwell, P. W. et Dominska, M. 1997. Stabilization of microsatellite sequences by variant repeats in the yeast *Saccharomyces cerevisiae*. *Genetics* **146:** 491-8.

Pigny, P. *et al.* 1996. Human mucin genes assigned to 11p15.5: identification and organization of a cluster of genes. *Genomics* **38:** 340-52.

Pinaud, E., Aupetit, C., Chauveau, C. et Cogne, M. 1997. Identification of a homolog of the C alpha 3'/hs3 enhancer and of an allelic variant of the 3'IgH/hs1,2 enhancer downstream of the human immunoglobulin alpha 1 gene. *Eur J Immunol* **27:** 2981-5.

Pizza, M. *et al.* 2000. Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing [see comments]. *Science* **287:** 1816-20.

Pourcel, C., Vidgop, Y., Ramisse, F., Vergnaud, G. et Tram, C. 2003. Characterization of a Tandem Repeat Polymorphism in *Legionella pneumophila* and Its Use for Genotyping. *J Clin Microbiol* **41:** 1819-1826.

Read, T. D. *et al.* 2002. Comparative genome sequencing for discovery of novel polymorphisms in Bacillus anthracis. *Science* **296:** 2028-33.

Rice, P., Longden, I. et Bleasby, A. 2000. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* **16:** 276-7.

Richard, G. F. & Paques, F. 2000. Mini- and microsatellite expansions: the recombination connection. *EMBO Rep* **1:** 122-6.

Robitaille, Y., I. Lopes-Cendes, M. Becher, G. Rouleau and A. W. Clark. 1997. The neuropathology of CAG repeat diseases: review and update of genetic and molecular features. *Brain Pathol* **7**: 901-26.

Roversi, G., Beghini, A., Zambruno, G., Paradisi, M. et Larizza, L. 2003. Identification of two novel RECQL4exonic SNPs and genomic characterization of the IVS12 minisatellite. *J Hum Genet* **48:** 107-9.

Royle, N. J., Clarkson, R. E., Wong, Z. et Jeffreys, A. J. 1988. Clustering of hypervariable minisatellites in the proterminal regions of human autosomes. *Genomics* **3:** 352-360.

Sabol, S. Z., Hu, S. et Hamer, D. 1998. A functional polymorphism in the monoamine oxidase A gene promoter. *Hum Genet* **103:** 273-9.

Sagot, M. F. & Myers, E. W. 1998. Identifying satellites and periodic repetitions in biological sequences. *J Comput Biol* **5:** 539-53.

Saitou, N. & Nei, M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* **4:** 406-25.

Sandberg, G. & Schalling, M. 1997. Effect of in vitro promoter methylation and CGG repeat expansion on FMR-1 expression. *Nucleic Acids Res* **25:** 2883-7.

Saunders, N. J., Peden, J. F., Hood, D. W. et Moxon, E. R. 1998. Simple sequence repeats in the *Helicobacter pylori* genome. *Mol Microbiol* **27:** 1091-8.

Scharf, S. J., Bowcock, A. M., McClure, G., Klitz, W., Yandell, D. W. et Erlich, H. A. 1992. Amplification and characterization of the retinoblastoma gene VNTR by PCR. *Am J Hum Genet* **50:** 371-81.

Scherer, S. W. *et al.* 2003. Human chromosome 7: DNA sequence and biology. *Science* **300:** 767-72.

Schlotterer, C. & Tautz, D. 1992. Slippage synthesis of simple sequence DNA. *Nucleic Acids Res* **20:** 211-5.

Schlotterer, C., Ritter, R., Harr, B. et Brem, G. 1998. High mutation rate of a long microsatellite allele in *Drosophila melanogaster* provides evidence for allele-specific mutation rates. *Mol Biol Evol* **15:** 1269-74.

Schueler, M. G., Higgins, A. W., Rudd, M. K., Gustashaw, K. et Willard, H. F. 2001. Genomic and genetic definition of a functional human centromere. *Science* **294:** 109-15.

Schweitzer, J. K. & Livingston, D. M. 1998. Expansions of CAG repeat tracts are frequent in a yeast mutant defective in Okazaki fragment maturation. *Hum Mol Genet* **7:** 69-74.

Scott, H. S., Nelson, P. V., MacDonald, M. E., Gusella, J. F., Hopwood, J. J. et Morris, C. P. 1992. An 86-bp VNTR within IDUA is the basis of the D4S111 polymorphic locus. *Genomics* **14:** 1118-20.

Sehouli, J. & Mustea, A. 2002. Interleukin-1 receptor antagonist gene polymorphism and cancer. *Clin Infect Dis* **34:** 1535-6.

Semple, C. A., Morris, S. W., Porteous, D. J. et Evans, K. L. 2002. Computational comparison of human genomic sequence assemblies for a region of chromosome 4. *Genome Res* **12:** 424-9.

Seznec, H., Lia-Baldini, A. S., Duros, C., Fouquet, C., Lacroix, C., Hofmann-Radvanyi, H., Junien, C. et Gourdon, G. 2000. Transgenic mice carrying large human genomic sequences with expanded CTG repeat mimic closely the DM CTG repeat intergenerational and somatic instability. *Hum Mol Genet* **9:** 1185-94.

Shaikh, T. H. *et al.* 2000. Chromosome 22-specific low copy repeats and the 22q11.2 deletion syndrome: genomic organization and deletion endpoint analysis. *Hum Mol Genet* **9:** 489-501.

Shopsin, B., Gomez, M., Waddington, M., Riehman, M. et Kreiswirth, B. N. 2000. Use of coagulase gene (coa) repeat region nucleotide sequences for typing of methicillin-resistant *Staphylococcus aureus* strains. *J Clin Microbiol* **38:** 3453-6.

Sia, E. A., Kokoska, R. J., Dominska, M., Greenwell, P. et Petes, T. D. 1997. Microsatellite instability in yeast: dependence on repeat unit size and DNA mismatch repair genes. *Mol Cell Biol* **17:** 2851-8.

Silva, F. *et al.* 2001. MUC1 gene polymorphism in the gastric carcinogenesis pathway. *Eur J Hum Genet* **9:** 548-52.

Simon, M., Phillips, M. et Green, H. 1991. Polymorphism due to variable number of repeats in the human involucrin gene. *Genomics* **9:** 576-580.

Smith, T. F. & Waterman, M. S. 1981. Identification of common molecular subsequences. *J Mol Biol* **147:** 195-7.

Song, J., Yoon, Y., Park, K. U., Park, J., Hong, Y. J., Hong, S. H. et Kim, J. Q. 2003. Genotype-specific influence on nitric oxide synthase gene expression, protein concentrations, and enzyme activity in cultured human endothelial cells. *Clin Chem* **49:** 847-52.

Spire-Vayron de la Moureyre, C., Debuysere, H., Fazio, F., Sergent, E., Bernard, C., Sabbagh, N., Marez, D., Lo Guidice, J. M., D'Halluin J, C. et Broly, F. 1999. Characterization of a variable number tandem repeat region in the thiopurine S-methyltransferase gene promoter. *Pharmacogenetics* **9:** 189-98.

Staden, R. 1979. A strategy of DNA sequencing employing computer programs. *Nucleic Acids Res* **6:** 2601-10.

Staden, R. 1980. A new computer method for the storage and manipulation of DNA gel reading data. *Nucleic Acids Res* **8:** 3673-94.

Stead, J. D. & Jeffreys, A. J. 2000. Allele diversity and germline mutation at the insulin minisatellite. *Hum Mol Genet* **9:** 713-23.

Stothard, D. R., Van Der Pol, B., Smith, N. J. et Jones, R. B. 1998. Effect of serial passage in tissue culture on sequence of omp1 from *Chlamydia trachomatis* clinical isolates. *J Clin Microbiol* **36:** 3686-8.

Strand, M., Prolla, T. A., Liskay, R. M. et Petes, T. D. 1993. Destabilization of tracts of simple repetitive DNA in yeast by mutations affecting DNA mismatch repair. *Nature* **365:** 274-276.

Sun, X., Wahlstrom, J. et Karpen, G. 1997. Molecular structure of a functional Drosophila centromere. *Cell* **91:** 1007-19.

Supply, P., Magdalena, J., Himpens, S. et Locht, C. 1997. Identification of novel intergenic repetitive units in a mycobacterial two-component system operon. *Mol Microbiol* **26:** 991-1003.

Supply, P., Mazars, E., Lesjean, S., Vincent, V., Gicquel, B. et Locht, C. 2000. Variable human minisatellite-like regions in the *Mycobacterium tuberculosis* genome. *Mol Microbiol* **36:** 762-771.

Sutherland, G. R., Baker, E. et Richards, R. I. 1998. Fragile sites still breaking. *Trends Genet* **14:** 501-6.

Swanson, J. *et al.* 2000. Attention deficit/hyperactivity disorder children with a 7-repeat allele of the dopamine receptor D4 gene have extreme behavior but normal performance on critical neuropsychological tests of attention. *Proc Natl Acad Sci U S A* **97:** 4754-9.

Sybenga, J. 1999. What makes homologous chromosomes find each other in meiosis? A review and an hypothesis. *Chromosoma* **108:** 209-19.

Sylvestre, P., Couture-Tosi, E. et Mock, M. 2003. Polymorphism in the collagen-like region of the *Bacillus anthracis* BclA protein leads to variation in exosporium filament length. *J Bacteriol* **185:** 1555-63.

Tamaki, K., May, C. A., Dubrova, Y. E. et Jeffreys, A. J. 1999. Extremely complex repeat shuffling during germline mutation at human minisatellite B6.7. *Hum Mol Genet* **8:** 879-88.

Tammi, M. T., Arner, E. et Andersson, B. 2003. TRAP: Tandem Repeat Assembly Program produces improved shotgun assemblies of repetitive sequences. *Comput Methods Programs Biomed* **70:** 47-59.

Taylor, J. S., Sanny, J. S. et Breden, F. 1999. Microsatellite allele size homoplasy in the guppy (*Poecilia reticulata*). *J Mol Evol* **48:** 245-7.

Taylor, J. S. & Breden, F. 2000. Slipped-strand mispairing at noncontiguous repeats in *Poecilia reticulata*: a model for minisatellite birth. *Genetics* **155:** 1313-20.

Tenover, F. C., Arbeit, R. D., Goering, R. V., Mickelsen, P. A., Murray, B. E., Persing, D. H. et Swaminathan, B. 1995. Interpreting chromosomal DNA restriction patterns produced by pulsed-field gel electrophoresis: criteria for bacterial strain typing. *J Clin Microbiol* **33:** 2233-9.

Timchenko, L. T. & Caskey, C. T. 1999. Triplet repeat disorders: discussion of molecular mechanisms. *Cell Mol Life Sci* **55:** 1432-47.

Tishkoff, D. X., Filosi, N., Gaida, G. M. et Kolodner, R. D. 1997. A novel mutation avoidance mechanism dependent on S. cerevisiae *RAD27* is distinct from DNA mismatch repair. *Cell* **88:** 253-263.

Tomb, J. F. *et al.* 1997. The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature* **388:** 539-47.

Tonjum, T., Caugant, D. A., Dunham, S. A. et Koomey, M. 1998. Structure and function of repetitive sequence elements associated with a highly polymorphic domain of the *Neisseria meningitidis* PilQ protein. *Mol Microbiol* **29:** 111-24.

Toribara, N. W., Gum, J. R., Jr., Culhane, P. J., Lagace, R. E., Hicks, J. W., Petersen, G. M. et Kim, Y. S. 1991. MUC-2 human small intestinal mucin gene structure. Repeated arrays and polymorphism. *J Clin Invest* **88:** 1005-13.

Toribara, N. W., Roberton, A. M., Ho, S. B., Kuo, W. L., Gum, E., Hicks, J. W., Gum, J. R., Jr., Byrd, J. C., Siddiki, B. et Kim, Y. S. 1993. Human gastric mucin. Identification of a unique species by expression cloning. *J Biol Chem* **268:** 5879-85.

Toth, G., Gaspari, Z. et Jurka, J. 2000. Microsatellites in different eukaryotic genomes: survey and analysis. *Genome Res* **10:** 967-81.

Tran, H. T., Gordenin, D. A. et Resnick, M. A. 1996. The prevention of repeat-associated deletions in *Saccharomyces cerevisiae* by mismatch repair depends on size and origin of deletions. *Genetics* **143:** 1579-87.

Treco, D. & Arnheim, N. 1986. The evolutionarily conserved repetitive sequence d(TG.AC)n promotes reciprocal exchange and generates unusual recombinant tetrads during yeast meiosis. *Mol Cell Biol* **6:** 3934-47.

Turri, M. G., Cuin, K. A. et Porter, A. C. 1995. Characterisation of a novel minisatellite that provides multiple splice donor sites in an interferon-induced transcript. *Nucleic Acids Res* **23:** 1854-1861.

Ugarkovic, D. & Plohl, M. 2002. Variation in satellite DNA profiles--causes and effects. *Embo J* **21:** 5955-9.

Urquhart, A. & Gill, P. 1993. Tandem-repeat internal mapping (TRIM) of the involucrin gene: repeat number and repeat-pattern polymorphism within a coding region in human populations. *Am. J. Hum. Genet.* **53:** 279-286.

Vamvakopoulos, J. E., Taylor, C. J., Morris-Stiff, G. J., Green, C. et Metcalfe, S. 2002. The interleukin-1 receptor antagonist gene: a single-copy variant of the intron 2 variable number tandem repeat (VNTR) polymorphism. *Eur J Immunogenet* **29:** 337-40.

van Belkum, A., Scherer, S., van Alphen, L. et Verbrugh, H. 1998. Short-sequence DNA repeats in prokaryotic genomes. *Microbiol Mol Biol Rev* **62:** 275-93.

van Belkum, A. 1999a. The role of short sequence repeats in epidemiologic typing. *Curr Opin Microbiol* **2:** 306-11.

van Belkum, A. 1999b. Short sequence repeats in microbial pathogenesis and evolution. *Cell Mol Life Sci* **56:** 729-34.

van Belkum, A., Scherer, S., van Leeuwen, W., Willemse, D., van Alphen, L. et Verbrugh, H. 1997. Variable number of tandem repeats in clinical strains of *Haemophilus influenzae*. *Infect Immun* **65:** 5017-27.

van Belkum, A., Struelens, M., de Visser, A., Verbrugh, H. et Tibayrenc, M. 2001. Role of genomic typing in taxonomy, evolutionary genetics, and microbial epidemiology. *Clin Microbiol Rev* **14:** 547-60.

van der Ende, A., Hopman, C. T., Zaat, S., Essink, B. B., Berkhout, B. et Dankert, J. 1995. Variable expression of class 1 outer membrane protein in *Neisseria meningitidis* is caused by variation in the spacing between the -10 and -35 regions of the promoter. *J Bacteriol* **177:** 2475-80.

van Embden, J. D., van Gorkom, T., Kremer, K., Jansen, R., van Der Zeijst, B. A. et Schouls, L. M. 2000. Genetic variation and evolutionary origin of the direct repeat locus of *Mycobacterium tuberculosis* complex bacteria. *J Bacteriol* **182:** 2393-401.

van Ham, S. M., van Alphen, L., Mooi, F. R. et van Putten, J. P. 1993. Phase variation of *H. influenzae* fimbriae: transcriptional control of two divergent genes through a variable combined promoter region. *Cell* **73:** 1187-96.

Van Klinken, B. J., Van Dijken, T. C., Oussoren, E., Buller, H. A., Dekker, J. et Einerhand, A. W. 1997. Molecular cloning of human MUC3 cDNA reveals a novel 59 amino acid tandem repeat region. *Biochem Biophys Res Commun* **238:** 143-8.

Vandenbergh, D. J., Persico, A. M., Hawkins, A. L., Griffin, C. A., Li, X., Jabs, E. W. et Uhl, G. R. 1992. Human dopamine transporter gene (DAT1) maps to chromosome 5p15.3 and displays a VNTR. *Genomics* **14:** 1104-6.

Vandenbroeck, K., Fiten, P., Ronsse, I., Goris, A., Porru, I., Melis, C., Rolesu, M., Billiau, A., Marrosu, M. G. et Opdenakker, G. 2000. High-resolution analysis of IL-6 minisatellite polymorphism in Sardinian multiple sclerosis: effect on course and onset of disease. *Genes Immun* **1:** 460-3.

Venter, J. C. *et al.* 2001. The sequence of the human genome. *Science* **291:** 1304-51.

Vergnaud, G. 1989. Polymers of random short oligonucleotides detect polymorphic loci in the human genome. *Nucleic Acids Res.* **17:** 7623-7630.

Vergnaud, G., Mariat, D., Apiou, F., Aurias, A., Lathrop, M. et Lauthier, V. 1991. The use of synthetic tandem repeats to isolate new VNTR loci: cloning of a human hypermutable sequence. *Genomics* **11:** 135-144.

Vergnaud, G. & Denoeud, F. 2000. Minisatellites: Mutability and Genome Architecture. *Genome Res* **10:** 899-907.

Verkerk, A. J. M. H. *et al.* 1991. Identification of a gene (FMR-1) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome. *Cell* **65:** 905-914.

Vinall, L. E., Hill, A. S., Pigny, P., Pratt, W. S., Toribara, N., Gum, J. R., Kim, Y. S., Porchet, N., Aubert, J. P. et Swallow, D. M. 1998. Variable number tandem repeat polymorphism of the mucin genes located in the complex on 11p15.5. *Hum Genet* **102:** 357-66.

Vinall, L. E. *et al.* 2000. Polymorphism of human mucin genes in chest disease: possible significance of MUC2. *Am J Respir Cell Mol Biol* **23:** 678-86.

Vogt, P. H. *et al.* 1997. Report of the Third International Workshop on Y Chromosome Mapping 1997. Heidelberg, Germany, April 13-16, 1997. *Cytogenet Cell Genet* **79:** 1-20.

Volfovsky, N., Haas, B. J. et Salzberg, S. L. 2001. A clustering method for repeat analysis in DNA sequences. *Genome Biol* **2:** RESEARCH0027.

Vos, P. *et al.* 1995. AFLP: a new technique for DNA fingerprinting. *Nucleic Acids Res* **23:** 4407-14.

Wahls, W. P., Wallace, L. J. et Moore, P. D. 1990. The Z-DNA motif d(TG)30 promotes reception of information during gene conversion events while stimulating homologous recombination in human cells in culture. *Mol. Cell. Biol.* **10:** 785.

Wahls, W. P. & Moore, P. D. 1998. Recombination hotspot activity of hypervariable minisatellite DNA requires minisatellite DNA binding proteins. *Somat Cell Mol Genet* **24:** 41-51.

Warpeha, K. M., Xu, W., Liu, L., Charles, I. G., Patterson, C. C., Ah-Fat, F., Harding, S., Hart, P. M., Chakravarthy, U. et Hughes, A. E. 1999. Genotyping and functional analysis of a polymorphic (CCTTT)(n) repeat of NOS2A in diabetic retinopathy. *Faseb J* **13:** 1825-32.

Waterston, R. H., Lander, E. S. et Sulston, J. E. 2002. On the sequencing of the human genome. *Proc Natl Acad Sci U S A* **99:** 3712-6.

Weber, J. L. 1990. Informativeness of human (dC-dA)n (dG-dT)n polymorphisms. *Genomics* **7:** 524-530.

Weiser, J. N., Love, J. M. et Moxon, E. R. 1989. The molecular mechanism of phase variation of *H. influenzae* lipopolysaccharide. *Cell* **59:** 657-65.

Weiser, J. N., Williams, A. et Moxon, E. R. 1990. Phase-variable lipopolysaccharide structures enhance the invasive capacity of *Haemophilus influenzae*. *Infect Immun* **58:** 3455-7.

Weitzel, J. N., Ding, S., Larson, G. P., Nelson, R. A., Goodman, A., Grendys, E. C., Ball, H. G. et Krontiris, T. G. 2000. The HRAS1 minisatellite locus and risk of ovarian cancer. *Cancer Res* **60:** 259-61.

Wells, R. D. 1996. Molecular basis of genetic instability of triplet repeats. *J Biol Chem* **271:** 2875-8.

Wierdl, M., Dominska, M. et Petes, T. D. 1997. Microsatellite instability in yeast: dependence on the length of the microsatellite. *Genetics* **146:** 769-79.

Williams, J. G. K., Kubelik, A. R., Livak, K. J., Rafalski, J. A. et Tingey, S. V. 1990. DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. *Nucleic Acids Res.* **18:** 6531-6535.

Winter, E. & Varshavsky, A. 1989. A DNA binding protein that recognizes oligo(dA).oligo(dT) tracts. *Embo J* **8:** 1867-77.

Woerner, S. M., Benner, A., Sutter, C., Schiller, M., Yuan, Y. P., Keller, G., Bork, P., Doeberitz, M. K. et Gebert, J. F. 2003. Pathogenesis of DNA repair-deficient cancers: a statistical meta-analysis of putative Real Common Target genes. *Oncogene* **22:** 2226-35.

Wong, Z., Wilson, V., Patel, I., Povey, S. et Jeffreys, A. J. 1987. Characterization of a panel of highly variable minisatellites cloned from human DNA. *Annu. Hum. Genet.* **51:** 269-288.

Wooster, R., Bignell, G., Lancaster, J., Swift, S., Seal, S., Mangion, J., Collins, N., Gregory, S., Gumbs, C. et Micklem, G. 1995. Identification of the breast cancer susceptibility gene BRCA2. *Nature* **378:** 789-92.

Wren, J. D., Forgacs, E., Fondon, J. W., 3rd, Pertsemlidis, A., Cheng, S. Y., Gallardo, T., Williams, R. S., Shohet, R. V., Minna, J. D. et Garner, H. R. 2000. Repeat polymorphisms within gene regions: phenotypic and evolutionary implications. *Am J Hum Genet* **67:** 345-56.

Wyman, A. R. & White, R. 1980. A highly polymorphic locus in human DNA. *Proc. Natl. Acad. Sci. USA* **77:** 6754-6758.

Xu, G. & Goodridge, A. G. 1998. A CT repeat in the promoter of the chicken malic enzyme gene is essential for function at an alternative transcription start site. *Arch Biochem Biophys* **358:** 83-91.

Xu, X., Peng, M. et Fang, Z. 2000. The direction of microsatellite mutations is dependent upon allele length. *Nat Genet* **24:** 396-9.

Yamada, N. A., Smith, G. A., Castro, A., Roques, C. N., Boyer, J. C. et Farber, R. A. 2002. Relative rates of insertion and deletion mutations in dinucleotide repeats of various lengths in mismatch repair proficient mouse and mismatch repair deficient human cells. *Mutat Res* **499:** 213-25.

Yamauchi, M., Tsuji, S., Mita, K., Saito, T. et Morimyo, M. 2000. A novel minisatellite repeat expansion identified at FRA16B in a Japanese carrier. *Genes Genet Syst* **75:** 149-54.

Yan, L., Zhang, S., Eiff, B., Szumlanski, C. L., Powers, M., O'Brien, J. F. et Weinshilboum, R. M. 2000. Thiopurine methyltransferase polymorphic tandem repeat: genotype-phenotype correlation analysis. *Clin Pharmacol Ther* **68:** 210-9.

Yang, F., Hanson, N. Q., Schwichtenberg, K. et Tsai, M. Y. 2000. Variable number tandem repeat in exon/intron border of the cystathionine beta-synthase gene: a single nucleotide substitution in the second repeat prevents multiple alternate splicing. *Am J Med Genet* **95:** 385-90.

Yauk, C. L. & Quinn, J. S. 1996. Multilocus DNA fingerprinting reveals high rate of heritable genetic mutation in herring gulls nesting in an industrialized urban site. *Proc. Natl. Acad. Sci. USA* **93:** 12137-12141.

Yeramian, E. & Buc, H. 1999. Tandem repeats in complete bacterial genome sequences: sequence and structural analyses for comparative studies. *Res Microbiol* **150:** 745-54.

Yoshida, T., Obata, N. et Oosawa, K. 2000. Color-coding reveals tandem repeats in the Escherichia coli genome. *J Mol Biol* **298:** 343-9.

Young, E. T., Sloan, J. S. et Van Riper, K. 2000. Trinucleotide repeats are clustered in regulatory genes in *Saccharomyces cerevisiae*. *Genetics* **154:** 1053-68.

Yousef, G. M., Bharaj, B. S., Yu, H., Poulopoulos, J. et Diamandis, E. P. 2001. Sequence analysis of the human kallikrein gene locus identifies a unique polymorphic minisatellite element. *Biochem Biophys Res Commun* **285:** 1321-9.

Yu, S. *et al.* 1997. Human chromosomal fragile site FRA16B is an amplified AT-rich minisatellite repeat. *Cell* **88:** 367-374.

Zagon, I. S., Verderame, M. F., Allen, S. S. et McLaughlin, P. J. 2000. Cloning, sequencing, chromosomal location, and function of cDNAs encoding an opioid growth factor receptor (OGFr) in humans. *Brain Res* **856:** 75-83.

Zagursky, R. J., Olmsted, S. B., Russell, D. P. et Wooters, J. L. 2003. Bioinformatics: how it is being used to identify bacterial vaccine candidates. *Expert Rev Vaccines* **2:** 417-36.

Zhuchenko, O., Bailey, J., Bonnen, P., Ashizawa, T., Stockton, D. W., Amos, C., Dobyns, W. B., Subramony, S. H., Zoghbi, H. Y. et Lee, C. C. 1997. Autosomal dominant cerebellar ataxia (SCA6) associated with small polyglutamine expansions in the alpha 1A-voltage-dependent calcium channel. *Nat Genet* **15:** 62-9.

Zivanovic, Y., Lopez, P., Philippe, H. et Forterre, P. 2002. *Pyrococcus* genome comparison evidences chromosome shuffling-driven evolution. *Nucleic Acids Res* **30:** 1902-10.

Zorkol'tseva, I. V., Liubinskii, O. A., Sharipov, R. N., Zaidman, A. M., Aksenovich, T. I. et Dymshits, G. M. 2002. Analysis of polymorphism of the number of tandem repeats in the aggrecan gene exon G3 in the families with idiopathic scoliosis. *Genetika* **38:** 259-63.

Annexes

Annexe 1 : Extrait du programme Perl utilisé pour générer les tables à importer dans la base de données des répétitions en tandem

```
sub detection_redondance
{
# Détecte la redondance : le tableau groupe contiendra 0 aux indices
correspondant à des sorties non-redondantes, et il contiendra le numéro du
groupe de redondance aux indices correspondant à des sorties redondantes (au
moins 2 sorties auront le même numéro de groupe)

#taille_tabl = dimension de nos tableaux
#pos_left, pos_right = tableaux contenant les indices de début et de fin
#groupe = tableau contenant les numéros de groupes
#deb_plage et fin_plage = positions de début et de fin de la plage union des
plages chevauchantes

my ($i, $j);

# initialisation du tableau "groupe":
for ($i=0; $i<= $taille_tabl; $i++)
{
      $groupe[$i] = 0;
}

$d1 = $pos_left[0];
$f1 = $pos_right[0];

$j = 1; #numéro de groupe
$dans_groupe = 0;   # vaudra 1 si on est dans un groupe

for ($i = 1; $i <= $taille_tabl; $i++)
{
    $d2 = $pos_left[$i];
    $f2 = $pos_right[$i];

    if ($d2 > $f1)
    {   #plages 1 et 2 non chevauchantes
        $d1 = $d2;
        $f1 = $f2;
        if ($dans_groupe == 1)
        {   # Au tour précédent, nous étions dans un groupe : on vient de
quitter un groupe
            $groupe[$i-1] = $j;  # On remplit groupe(indice du dernier du
groupe)
            $j++;
            $dans_groupe = 0;
        }
    }
    else
    {   #plages 1 et 2 chevauchantes
        $d1 = min($d1, $d2);
        $f1 = max($f1, $f2);
        #on prend l'union des 2 plages
        $deb_plage[$j] = $d1;
        $fin_plage[$j] = $f1;
        $groupe[$i - 1] = $j;
        if ($i == $taille_tabl)
```

```
        {   #on ne rentrera plus dans la boucle : il faut donc remplir
groupe(indice du dernier du groupe) soit groupe(taille_tabl)
            $groupe[$i] = $j;
            $j++;    # pr calculer Nb_groupes
        }
        $dans_groupe = 1;      #on signale qu'on était dans un groupe à ce tour
    }
}
}
#############################################################################

sub traitement_redondance
{
# Traite la redondance en sélectionnant le plus petit motif répété parmi ceux
qui ont une longueur totale maximale à 20% près et ayant les meilleurs taux
d'alignement. On conservera l'union des étendues chevauchantes.
# Recalcule aussi la longueur totale et le nombre de répétitions et modifie les
cases correspondantes.

# On est à l'intérieur d'un groupe de redondance
# ideb = indice du début du groupe
# ifin = indice de la fin du groupe
# debut_pl et fin_pl = début et fin de plage

my($i, $ind);

# 1. On recherche l'indice correspondant à la longueur totale maximale :
ind_maxlong et l'indice correspondant au %matches maximal pour une longueur
totale maximale, à 20% près

$ind_maxlong=max_tabl($ideb, $ifin);

$ind_maxmatch = $ind_maxlong;  #il faut initialiser à un indice répondant au
critère de longueur totale

for ($i = $ideb; $i <= $ifin; $i++)
{
    if ($long_tot[$i] >= $long_tot[$ind_maxlong] * 0.8)
    {
        if ($matches[$i] > $matches[$ind_maxmatch])
        {
            $ind_maxmatch = $i;
        }
    }
}

# 2. On recherche le motif minimal, pour des séquences répondant aux deux
critères suivants: longueur totale maximale à 20% près, et %matches maximal à
10 près

$ind_motifmin = $ind_maxmatch;
#initialise dans la plage considérée (conditions sur Ltot et %M vérifiées)

for ($i = $ideb; $i <= $ifin; $i++)
{
    if ($long_tot[$i] >= $long_tot[$ind_maxlong] * 0.8 && $matches[$i] >=
$matches[$ind_maxmatch]- 10)
    {
        if ($U[$i] < $U[$ind_motifmin])
        {
            $ind_motifmin = $i;
```

```
        }
        if ($U[$i] == $U[$ind_motifmin])  #on privilegie la plus grande plage
        {
            if ($long_tot[$i] > $long_tot[$ind_motifmin])
            {
            $ind_motifmin=$i;
            }
        }
    }
}
$ind=$ind_motifmin;
#pour raccourcir la ligne à écrire

#3. Modification de la sequence: sequence_cor

$ileft=$ideb;
#ileft est l'indice de la séquence la plus à gauche: on commencera par elle
# (correspond à ideb car pos_left classées dans l'ordre croissant)
$fin_seq=$pos_right[$ileft];
$seq=$sequence[$ileft];
$i=$ideb;
while ($fin_seq < $fin_pl)
      {
            if ($pos_right[$i] > $fin_seq && $pos_left[$i] <= $fin_seq)
            {
               $ecart=$pos_right[$i]-$fin_seq;
               $deb = length($sequence[$i]) - $ecart;
               $seq=$seq.substr($sequence[$i],$deb);
               $fin_seq = $pos_right[$i];
            }
            $i++;
      }

$L= $fin_pl - $debut_pl + 1;
if (length($seq) == $L)
{
      $sequence_cor[$ind] = $seq;
}
else
{
      $sequence_cor[$ind] = "probleme!";
      print "pb seq_cor!_n";
}

#4. On écrit la ligne correspondante

$plage = $debut_pl."--".$fin_pl;
$N = $L / $U[$ind];
$N=int($N*100)/100; #arrondit à deux chiffres après la virgule
if ($hist == 1)
{
$ligne = $plage."\t".
$U[$ind]."\t".$N."\t".$cons_size[$ind]."\t".$matches[$ind]."\t".$indel[$ind]."\
t".$score[$ind]."\t".$pA[$ind]."\t".$pC[$ind]."\t".$pG[$ind]."\t".$pT[$ind]."\t
".$ent[$ind]."\t".$L."\t".$nom_seq."\t".$chemin_align[$ind]."\ta\ta\ta\t".$B_GC
[$ind]."\t".$B_AT[$ind]."\t".$B_pp[$ind]."\t".$consensus[$ind]."\t".$pos_left[$
ind]."\t".$pos_right[$ind]."\t".$sequence[$ind]."\ty\t".$sequence_cor[$ind]."\t
".$avg_ent[$ind]."\t".$historyR[$ind]."\n";
}
else
{
```

```
$ligne = $plage."\t".
$U[$ind]."\t".$N."\t".$cons_size[$ind]."\t".$matches[$ind]."\t".$indel[$ind]."\
t".$score[$ind]."\t".$pA[$ind]."\t".$pC[$ind]."\t".$pG[$ind]."\t".$pT[$ind]."\t
".$ent[$ind]."\t".$L."\t".$nom_seq."\t".$chemin_align[$ind]."\ta\ta\ta\t".$B_GC
[$ind]."\t".$B_AT[$ind]."\t".$B_pp[$ind]."\t".$consensus[$ind]."\t".$pos_left[$
ind]."\t".$pos_right[$ind]."\t".$sequence[$ind]."\ty\t".$sequence_cor[$ind]."\t
-1\t-1\n";
}
print OUT_FILE $ligne;

#5. Modification du fichier d'alignement

@al=split(/\/+/, $chemin_align[$ind]);
$align=join("\/","./alignements",$al[4],$al[5]);

open (ALIGN_FILE, ">>".$align) || die "Unable to open alignment file $align";
print ALIGN_FILE  "\nThe Tandem Repeat Finder software suggested alternative
ways to present\nthe alignment for this (or part of this) tandem repeat (see <a
href=\"http://iech5.igmors.u-psud.fr/ALIGNEMENTS/base_ms/overlapping.html\">
explanation</a>)\nOther alignments:\n";

for ($i = $ideb; $i <= $ifin; $i++)
{
            if ($i != $ind)
            {
                  print ALIGN_FILE  "<a href = $chemin_align[$i]>positions
$pos_left[$i] to $pos_right[$i]: $N[$i]x$U[$i] bp<\/a>\n";
            }
}
print ALIGN_FILE "\nEntire sequence: from positions $debut_pl to
$fin_pl\n$sequence_cor[$ind]\n";
close (ALIGN_FILE);
}
```

Annexe 2 : Extrait du script Perl utilisé pour le Blast de couples d'amorces PCR

```
#!/usr/bin/perl

# Réception argument d'entrée

$dossier_blast = $ARGV[0];

# Ce nom est celui du dossier avec les seq blast, du fichier .nt contenu
dans ce dossier, ET de la table correspondante dans base_ms.mdb
# les fichiers blasts .nt devront contenir les seq nommées:
>nom_base.posdebTR--posfinTR.descriptif
# descriptif=left flanking sequence ou right flanking sequence ou TR

if (substr($dossier_blast, 0, 7) ne "GENOMES")
{
     $seq=$dossier_blast;
     $base_blast="D:/denoeud/projet_blast/".$dossier_blast."/".$seq.".nt";
}
else
{
     $seq=substr($dossier_blast, 8);
     $base_blast="D:/denoeud/projet_blast/".$dossier_blast."/".$seq.".nt";
}
$fichier_input_left="C:/Inetpub/wwwroot/ASPSamp/base_ms/Blast/tmp/query_lef
t.txt";
$fichier_input_right="C:/Inetpub/wwwroot/ASPSamp/base_ms/Blast/tmp/query_ri
ght.txt";

#lancement des blasts:

$fichier_temp_left="C:/Inetpub/wwwroot/ASPSamp/base_ms/Blast/tmp/temp_blast
_left.txt";

$fichier_temp_right="C:/Inetpub/wwwroot/ASPSamp/base_ms/Blast/tmp/temp_blas
t_right.txt";

`blastall -p blastn -d $base_blast -i $fichier_input_left -F F -o
$fichier_temp_left -e 0.3`;

`blastall -p blastn -d $base_blast -i $fichier_input_right -F F -o
$fichier_temp_right -e 0.3`;

############ Ecriture entete fichier HTMl sortie #############

print "<HTML><TITLE>BLAST Search Results</TITLE><BODY BGCOLOR=\"#ffffc0\"
LINK=\"#0000FF\" VLINK=\"#660099\" ALINK=\"#660099\"><FONT face=\'courier
new\' size=2>";

if (substr($dossier_blast, 0, 7) ne "GENOMES")
{
print "<b><font face =arial size +2 color=#FF6600> BLAST OF PCR PRIMERS IN
THE <a href=\"http://minisatellites.u-psud.fr\">TANDEM REPEATS DATABASE</a>
</font><BR>";
}
else
{
print "<b><font face =arial size +2 color=#FF6600> BLAST of PCR primers in:
$seq </font><BR>";
}
```

```
  print "<br>  BLASTN 2.2.1 [Apr-13-2001]</b><b><a
href=\"http://www.ncbi.nlm.nih.gov/htbin-
post/Entrez/query?uid=9254694&form=6&db=m&Dopt=r\"><BR>";
  print "Reference</a>:</b><BR>Altschul, Stephen F., Thomas L. Madden,
Alejandro A. Sch&auml;ffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and
David J. Lipman (1997), \"Gapped BLAST and PSI-BLAST: a new generation of
protein database search programs\",  Nucleic Acids Res. 25:3389-
3402.</b><BR><BR>";
  print "<BR><font face=arial size +1><b> List of primer pairs
matches:</B></font><BR><BR>";

  ########## Analyse des résultats du blast ##########

  blast($fichier_temp_left);
  # retourne la pos 1 du primer même si le match ne va pas jusqu'à 1 et qq
soit le sens +/+ ou +/-
  # les positions de match sont rangées dans les tableaux matchP (+/+) et
matchM (+/-)

  positionne(); ## Calcule pos réelle match sur la séquence (corrige @matchP
  et @matchM)

  @matchLP = @matchP;
  @matchLM = @matchM;
  @exactLP=@match_exactP;
  @exactLM=@match_exactM;
  @nomLP=@nom_seqP;
  @nomLM=@nom_seqM;
  @redondLP=@groupe_redondP;
  @redondLM=@groupe_redondM;
  $nb_groupesLP=$num_groupeP;
  $nb_groupesLM=$num_groupeM;

  #print "matchs left +: @matchLP\t exact: @exactLP\tseq: @nomLP <BR> matchs
left -: @matchLM\texact:@exactLM\tseq: @nomLM <BR>";

  blast($fichier_temp_right);
  positionne(); ## Calcule pos réelle match sur la séquence

  @matchRP = @matchP;
  @matchRM = @matchM;
  @exactRP=@match_exactP;
  @exactRM=@match_exactM;
  @nomRP=@nom_seqP;
  @nomRM=@nom_seqM;
  @redondRP=@groupe_redondP;
  @redondRM=@groupe_redondM;
  $nb_groupesRP=$num_groupeP;
  $nb_groupesRM=$num_groupeM;

  ################ Analyse des matchs ##################
  match();

  ############ Ecriture fin fichier HTML ################
  print "</BODY></HTML>";

  ######################################################
   END
  {
  unlink ($fichier_temp_left, $fichier_input_left, $fichier_temp_right,
  $fichier_input_right) or die "Pb suppr fichiers tmp!\n"
```

Annexe 3 : Tableaux supplémentaires associés à l'article (Denoeud, 2003)

Tableau supplémentaire 1 :

IDENTIFIER INFORMATION				SEQUENCE CHARACTERISTICS									POLYMORPHISM MEASURES				ALLELE SIZE INFORMATION			PREDICTIVE CRITERIA			
chr	SET	Name	Physical position (kb)	Unit length (bp)	Copy number	Total length (bp)	percent matches	%GC	GC bias	PurPyr bias	avg entropy	HistoryR	number of alleles (28ind)	heterozygosity (28ind)	number of alleles (76ind)	heterozygosity (76ind)	Length predicted by Human Genome Project sequence (bp)	Length predicted by Celera sequence (bp)	observed size range (bp)	crit1	crit2	crit3	crit4
21	training	CEB256	220	26	49	1277	85%	53%	0.4	0.1	38.35	0.705	3	0.57	/	/	1454	?	900-1450	+	+	+	?
21	training	CEB255	3948	38	25	962	83%	8%	0.5	0.03	44.25	0.469	1	0	1	0	1045	682	1050	-	-	-	+
21	test	CEB258	4445	22	29	633	91%	27%	0.41	0.29	26.94	0.706	1	0	/	/	720	720	720	+	-	-	-
21	test	CEB259	6028	26	97	2523	73%	10%	0.6	0.08	45.32	0.556	failed	failed	/	/	2638	2461	failed	+	-	-	+
21	test	CEB260	7748	73	10	748	93%	35%	0.43	0.2	13.84	0.286	1	0	/	/	891	891	890	-	-	-	-
21	test	CEB261	10501	25	15	370	82%	1%	1	0.04	41.84	0.563	4	0.72	/	/	543	491	490-550	+	-	-	+
21	test	CEB262	19540	30	16	481	85%	5%	1	0.22	37.36	0.455	failed	failed	/	/	669	779	failed	-	-	-	+
21	test	CEB263	19963	29	28	811	71%	34%	0.47	0.28	63.1	0.272	1	0	/	/	917	917	920	-	-	-	-
21	test	CEB264	21556	30	32	959	79%	52%	0.69	0.13	46.92	0.387	1	0	/	/	1086	1026	1090	-	+	-	+
21	training	CEB215	22291	28	97	2705	87%	25%	0.44	0.22	39.74	0.788	failed	failed	failed	failed	2834	?	failed	+	-	-	?
21	test	CEB265	23006	54	13	680	74%	55%	0.38	0.24	54.57	0.429	2	0.14	/	/	830	831	830-880	-	+	-	-
21	training	CEB234	23639	30	13	396	82%	74%	0.38	0.41	34.99	0.7	2	0.36	2	0.28	636	636	610-640	+	+	+	-
21	test	CEB266	24912	45	13	565	72%	41%	0.9	0.27	46.95	0.438	2	0.36	/	/	821	819	820-870	-	-	-	-
21	test	CEB267	25371	24	18	419	90%	3%	1	0.16	28.15	0.4	failed	failed	/	/	1109	1112	failed	-	-	-	-
21	test	CEB268	26682	22	17	383	72%	60%	1	0.98	47.01	0.516	6	0.64	/	/	498	357	500-690	-	+	-	+
21	test	CEB269	28940	74	14	1021	97%	48%	0.37	0.46	4.75	0.6	14*	0.88*	/	/	1192	1044	1600-4500	+	+	+	+

21	training	CEB235	29314	39	17	665	90%	41%	0.56	0.21	28.38	0.684	2	0.04	2	0.03	813	813	1500-1850	+	-	-	-
21	training	CEB257	29529	18	25	446	75%	11%	1	0.11	55.31	0.552	failed	failed	failed	failed	618	546	failed	+	-	-	+
21	test	CEB270	30145	22	24	518	83%	61%	0.7	0.42	35.8	0.833	2	0.39	/	/	642	642	640-690	+	+	+	-
21	test	CEB271	30440	20	104	2084	75%	64%	0.97	0.66	35.05	0.391	10*	0.83*	/	/	2227	2038	1100-2900	-	+	-	+
21	training	CEB236	30516	23	28	639	74%	47%	0.53	0.37	49.78	0.595	1	0	1	0	899	904	900	+	-	-	-
21	test	CEB272	30670	34	12	410	98%	54%	0.37	0.18	4.87	1	2	0.47	/	/	600	498	600-640	+	+	+	+
21	test	CEB273	30771	28	14	379	80%	63%	0.9	0.69	42.28	0.368	1	0	/	/	580	580	580	-	+	-	-
21	test	CEB274	30829	29	36	1045	77%	64%	0.59	0.19	34.95	0.639	8	0.6	/	/	1241	1017	660-1420	+	+	+	+
21	test	CEB275	31147	34	37	1284	77%	61%	0.54	0.35	52.56	0.44	1	0	/	/	1508	1511	1510	-	+	-	-
21	test	CEB276	31252	22	35	769	78%	56%	0.39	0.16	35.26	0.516	2	0.19	/	/	838	541	840-870	-	+	-	+
21	training	CEB237	31420	44	19	752	81%	27%	0.78	0.22	35.29	0.633	2	0.29	2	0.27	912	690	910-960	+	-	-	+
21	training	CEB238	31572	51	59	3015	94%	61%	0.44	0.37	14.04	0.91	7*	0.79*	10*	0.79*	3109	7305	830-3100	+	+	+	+
21	test	CEB277	31619	114	14	1603	78%	41%	0.71	0.27	45.39	0.479	1	0	/	/	1654	1657	1660	-	-	-	-
21	training	CEB239	31675	37	12	446	93%	52%	0.62	0.54	25.72	0.842	3	0.56	3	0.57	650	539	650-850	+	+	+	+
21	test	CEB278	31833	31	30	921	83%	64%	0.41	0.06	32.28	0.548	4	0.69	/	/	1033	1000	810-1200	+	+	+	+
21	training	CEB240	32177	17	53	883	80%	56%	0.46	0.17	48.86	0.53	3	0.51	3	0.49	1040	1040	900-1060	-	+	-	-
21	test	CEB279	32246	34	18	617	76%	61%	0.54	0.23	52.7	0.378	1	0	/	/	761	763	760	-	+	-	-
21	training	CEB241	32529	20	122	2432	86%	61%	0.51	0.38	36.5	0.805	4	0.57	4	0.55	2553	1241	1980-2550	+	+	+	+
21	training	CEB242	32690	51	12	594	93%	70%	0.49	0.36	18.65	0.769	7	0.66	7	0.66	737	483	650-1200	+	+	+	+
21	test	CEB280	33109	19	22	422	80%	75%	0.39	0.34	50.05	0.48	2	0.07	/	/	589	590	590-670	-	+	-	-
21	test	CEB281	33286	35	11	393	83%	82%	0.59	0.63	26.48	0.6	3	0.37	/	/	635	635	600-670	+	+	+	-
21	training	CEB243	33317	60	11	653	93%	72%	0.58	0.34	14.42	1	3	0.27	4	0.32	765	765	410-770	+	+	+	-
21	training	CEB244	33318	36	10	376	97%	71%	0.38	0.48	9.49	1	2	0.17	3	0.3	643	644	640-860	+	+	+	-
21	training	CEB245	33383	52	12	631	88%	61%	0.64	0.28	26.9	0.333	2	0.04	2	0.04	901	901	850-900	-	+	-	-
21	test	CEB282	33481	20	21	416	77%	71%	0.75	0.4	51.79	0.519	3	0.23	/	/	645	647	610-710	-	+	-	-
21	test	CEB283	33711	17	55	936	80%	70%	0.37	0.05	41.41	0.589	4	0.74	/	/	1114	1045	1020-1110	+	+	+	+
21	test	CEB284	33832	21	55	1164	86%	59%	0.73	0.5	34.56	0.77	3	0.55	/	/	1322	545	1300-1340	+	+	+	+
21	training	CEB246	33920	43	12	496	87%	68%	0.38	0.34	32.47	0.889	5	0.66	5	0.69	758	758	600-1350	+	+	+	-

21	training	CEB247	34282	19	27	522	84%	73%	0.45	0.44	37.72	0.571	1	0	1	0	649	539	650	+	+	+	+
21	test	CEB285	34327	36	25	895	86%	64%	0.47	0.27	37.25	0.444	6	0.3	/	/	1092	1092	880-1090	-	+	-	-
21	training	CEB248	34384	36	37	1339	70%	53%	0.66	0.15	32.33	0.5	5	0.64	6	0.63	1510	1514	830-1800	-	+	-	-
21	test	CEB286	34390	18	34	609	82%	72%	0.61	0.22	31.44	0.776	5	0.66	/	/	713	713	560-850	+	+	+	-
21	test	CEB287	34454	44	10	448	77%	73%	0.45	0.31	44.91	0.227	3	0.51	/	/	570	570	430-570	-	+	-	-
21	test	CEB288	34474	66	11	694	86%	70%	0.4	0.31	31.47	0.667	3	0.57	/	/	822	394	690-820	+	+	+	+
21	training	CEB249	34683	51	22	1132	88%	63%	0.49	0.46	27.01	0.486	2	0.5	2	0.48	1305	430	1300-1350	-	+	-	+
21	test	CEB289	34741	45	45	2030	75%	44%	0.55	0.33	52.83	0.678	1	0	/	/	2144	3451	5000	+	-	-	+
21	test	CEB290	34830	52	12	627	91%	63%	0.43	0.28	18.19	0.682	4	0.64	/	/	833	374	670-920	+	+	+	+
21	test	CEB291	34854	28	22	602	85%	66%	0.52	0.22	37.67	0.706	10	0.87	12	0.88	772	775	770-1300	+	+	+	-
21	training	CEB250	34932	23	32	744	94%	72%	0.53	0.46	12.11	0.8	19	0.93	25	0.94	1045	487	540-1850	+	+	+	+
21	test	CEB292	34950	33	17	552	91%	79%	0.44	0.4	19.43	0.727	5	0.55	/	/	647	648	580-1100	+	+	+	-
21	training	CEB251	34992	39	40	1575	88%	43%	0.72	0.12	29.91	0.435	3	0.39	3	0.38	1737	?	1740-1960	-	-	-	?
21	training	CEB252	35084	50	17	843	88%	56%	0.36	0.14	25.59	0.757	4	0.61	5	0.62	983	1033	980-1330	+	+	+	+
21	training	CEB253	35145	45	14	629	97%	57%	0.37	0.29	7.46	0.833	6	0.49	6	0.62	798	663	870-1100	+	+	+	+
21	training	CEB254	35165	24	20	483	80%	64%	0.44	0.34	43.33	0.784	3	0.2	3	0.18	656	539	630-680	+	+	+	+
22	training	CEB224	1528	41	21	855	91%	63%	0.71	0.46	26.31	0.609	6	0.75	7	0.77	1066	1019	940-1500	+	+	+	+
22	training	CEB213	1571	62	25	1548	94%	62%	0.39	0.24	13.36	0.538	5	0.53	/	/	1401	1092	900-1700	-	+	-	+
22	test	CEB293	3522	36	20	719	86%	40%	0.4	0.34	33.64	0.64	3	0.35	/	/	858	858	820-910	+	-	-	-
22	training	CEB208	3635	20	149	2987	77%	69%	0.48	0.31	32.46	0.538	failed	failed	failed	failed	3126	3125	failed	-	+	-	-
22	training	CEB225	4095	19	24	459	77%	61%	0.57	0.25	42.31	0.538	2	0.23	2	0.22	555	555	550-600	-	+	-	-
22	training	CEB216	10395	23	59	1362	79%	57%	0.65	0.41	46.71	0.6	2	0.07	3	0.09	1563	1290	1300-1600	+	+	+	+
22	test	CEB294	11741	43	12	506	73%	49%	0.39	0.08	40.85	0.3	3	0.5	/	/	624	628	630-800	-	+	-	-
22	training	CEB226	12766	21	26	521	79%	55%	0.53	0.09	43.89	0.647	1	0	1	0	650	652	650	+	+	+	-
22	test	CEB295	13055	29	23	681	87%	52%	0.88	0.74	32.07	0.829	9*	0.8*	10*	0.73*	805	809	1020-4700	+	+	+	-
22	training	CEB222	15802	23	26	587	84%	57%	0.96	0.5	41.32	0.423	1	0	1	0	822	822	1200	-	+	-	-
22	test	CEB296	16228	21	83	1735	72%	57%	0.65	0.04	33.99	0.375	5	0.52	/	/	1866	1868	1860-2200	-	+	-	-
22	training	CEB227	16333	37	14	515	82%	17%	0.41	0.22	35.32	0.304	1	0	1	0	670	671	670	-	-	-	-

22	test	CEB297	16658	21	28	602	85%	71%	0.63	0.22	33.42	0.649	2	0.24	/	/	738	738	740-880	+	+	+	-
22	test	CEB298	18295	20	20	395	86%	58%	0.97	0.96	37.6	0.778	8	0.78	9	0.77	569	567	1080-1300	+	+	+	-
22	test	CEB299	19972	22	28	621	86%	53%	1	0.98	39.06	0.567	8	0.79	9	0.78	1016	1993	1020-2500	+	+	+	+
22	training	CEB228	20986	18	33	588	80%	45%	0.38	0.01	43.63	0.567	1	0	2	0.01	723	723	700-720	+	-	-	-
22	test	CEB300	21044	18	47	851	80%	63%	0.84	0.45	41.92	0.491	3	0.56	/	/	968	758	1420-1670	-	+	-	+
22	training	CEB201	21327	59	11	668	78%	70%	0.4	0.3	40.26	0.75	3	0.58	7	0.63	820	2301	320-900	+	+	+	+
22	test	CEB301	21541	147	12	1752	88%	57%	0.4	0.06	37.64	0.762	3	0.33	/	/	1790	1796	1650-2090	+	+	+	-
22	training	CEB214	22157	25	19	480	91%	4%	0.5	0.08	23.42	0.727	1	0	/	/	668	418	660	+	-	-	+
22	test	CEB302	22447	57	10	552	97%	51%	0.45	0.29	6.91	1	2	0.04	/	/	737	739	680-740	+	+	+	-
22	test	CEB303	23529	30	15	455	81%	57%	0.82	0.52	47.86	0.577	2	0.13	/	/	713	713	710-770	+	+	+	-
22	test	CEB304	25144	29	12	356	87%	55%	0.64	0.53	27.94	0.889	1	0	/	/	533	533	530	+	+	+	-
22	training	CEB229	26130	22	16	353	88%	35%	0.54	0.01	27.89	0.733	2	0.5	2	0.5	500	500	500-520	+	-	-	-
22	test	CEB305	27086	47	28	1313	76%	61%	0.74	0.61	45.56	0.727	14	0.89	16	0.92	1557	1109	1060-2150	+	+	+	+
22	test	CEB306	27209	20	21	426	74%	62%	0.71	0.64	47.17	0.389	1	0	/	/	629	629	630	-	+	-	-
22	test	CEB307	27382	26	15	401	85%	74%	0.57	0.48	28.33	0.667	4	0.58	/	/	528	528	430-640	+	+	+	-
22	training	CEB217	27398	20	68	1364	82%	47%	0.36	0.14	36.11	0.529	1	0	1	0	1527	1527	1530	-	-	-	-
22	test	CEB308	27654	28	33	918	91%	77%	0.43	0.39	18.94	0.872	7*	0.78*	7*	0.8*	1191	1191	600-1200	+	+	+	-
22	test	CEB309	27773	35	11	387	80%	28%	0.5	0.1	42.69	0.654	1	0	/	/	513	513	513	+	-	-	-
22	training	CEB212	28490	36	29	1037	88%	37%	0.89	0.56	28.54	0.652	4	0.49	4	0.52	1201	1203	1100-1400	+	-	-	-
22	training	CEB202	29067	41	18	703	82%	59%	0.8	0.33	35.83	0.476	19*	0.92*	21*	0.93*	907	909	630-1900	-	+	-	-
22	training	CEB230	30257	35	15	513	71%	65%	0.69	0.46	52.56	0.591	5	0.77	7	0.79	683	684	720-1200	+	+	+	-
22	training	CEB231	30336	20	88	1761	84%	43%	0.77	0.15	33.19	0.545	3	0.45	3	0.57	1862	1844	1620-1850	+	-	-	+
22	training	CEB209	30404	46	16	755	90%	74%	0.38	0.4	23.66	0.6	failed	failed	failed	failed	906	904	failed	+	+	+	-
22	test	CEB310	30541	24	25	595	98%	20%	0.6	0.02	8.79	1	12	0.85	12	0.85	998	998	900-1550	+	-	-	-
22	test	CEB311	30860	67	14	960	82%	59%	0.49	0.25	36.41	0.5	2	0.46			1115	1111	1110-1180	-	+	-	-
22	training	CEB232	31534	17	22	373	70%	39%	0.64	0.05	65.06	0.324	1	0	3	0.04	572	572	500-605	-	-	-	-
22	test	CEB312	31865	19	23	439	81%	64%	0.78	0.54	45.61	0.565	2	0.04	/	/	615	615	610-640	+	+	+	-
22	test	CEB313	32217	40	17	685	79%	51%	0.45	0.25	50.49	0.6	1	0	/	/	811	811	810	+	+	+	-

22	test	CEB314	32267	58	15	872	77%	72%	0.53	0.39	49.29	0.55	4	0.56	/	/	1037	1036	630-1040	+	+	+	-
22	training	CEB203	32298	21	25	529	92%	63%	0.62	0.46	16.84	0.833	1	0	1	0	695	695	700	+	+	+	-
22	test	CEB315	32458	23	43	996	80%	70%	0.63	0.63	31.95	0.788	5	0.72	/	/	1228	1245	1400-1600	+	+	+	+
22	training	CEB218	32693	34	12	418	88%	56%	0.43	0.04	24.97	0.727	1	0	1	0	590	590	600	+	+	+	-
22	test	CEB316	32741	18	33	592	78%	64%	0.87	0.63	40.96	0.364	5	0.62	/	/	757	759	700-810	-	+	-	-
22	training	CEB219	32911	30	14	411	92%	61%	0.38	0.06	24.51	0.611	3	0.51	4	0.53	559	559	500-700	+	+	+	-
22	test	CEB317	32915	26	19	489	81%	54%	0.7	0.41	44.76	0.444	1	0	/	/	594	594	600	-	+	-	-
22	training	CEB204	32948	32	21	671	82%	55%	0.6	0.31	41.49	0.679	5	0.66	5	0.61	875	867	850-1500	+	+	+	-
22	test	CEB318	32980	43	13	559	90%	56%	0.68	0.43	24.33	0.727	3	0.43	/	/	777	777	730-860	+	+	+	-
22	training	CEB205	33057	33	33	1086	96%	71%	0.38	0.36	8.75	0.882	19*	0.93*	23*	0.94*	1318	1317	550-2500	+	+	+	-
22	training	CEB206	33318	45	22	1024	88%	64%	0.47	0.32	27.93	0.581	2	0.44	3	0.47	1227	1226	820-1270	+	+	+	-
22	training	CEB233	33400	45	16	736	82%	43%	0.4	0.08	39.87	0.143	1	0	1	0	856	856	850	-	-	-	-
22	test	CEB319	33414	38	11	420	93%	62%	0.39	0.16	16.51	0.4	1	0	/	/	572	572	570	-	+	-	-
22	test	CEB320	33419	32	18	563	86%	67%	0.58	0.39	32.65	0.75	2	0.04	/	/	691	691	690-760	+	+	+	-
22	test	CEB321	33434	31	38	1172	77%	51%	1	0.48	38.15	0.321	4	0.57	/	/	1272	1272	1240-1500	-	+	-	-
22	test	CEB322	33545	33	11	359	91%	57%	0.4	0.22	23.67	0.75	2	0.04	/	/	578	578	540-580	+	+	+	-
22	training	CEB220	33592	47	13	611	98%	60%	0.4	0.27	3.81	0.667	3	0.28	4	0.21	785	785	600-930	+	+	+	-
22	test	CEB323	33618	37	19	702	71%	71%	0.41	0.34	57.46	0.5	1	0	/	/	872	872	870	-	+	-	-
22	training	CEB207	33817	19	129	2459	76%	51%	0.41	0.3	48.19	0.701	2	0.13	2	0.08	2617	2663	2600-3500	+	+	+	+
22	test	CEB324	33825	43	21	890	93%	56%	0.64	0.17	14.25	0.621	19*	0.93*	27*	0.94*	985	244	400-3500	+	+	+	+
22	test	CEB325	33854	18	42	751	81%	61%	0.64	0.21	35.2	0.381	4	0.53	/	/	926	928	930-1400	-	+	-	-
22	training	CEB221	33864	23	27	625	91%	58%	0.45	0.22	26.43	0.852	3	0.51	4	0.46	802	686	750-830	+	+	+	+
22	test	CEB326	33965	75	16	1224	83%	49%	0.63	0.14	36.35	0.508	3	0.25	/	/	1408	802	1250-1680	-	+	-	+
22	test	CEB327	33983	52	10	543	74%	70%	0.46	0.24	53.47	0.25	1	0	/	/	793	793	800	-	+	-	-
22	training	CEB223	34031	27	54	1456	75%	60%	0.53	0.2	47.81	0.428	2	0.5	2	0.5	1836	1864	1850-1950	-	+	-	+
22	training	CEB210	34182	62	28	1732	93%	71%	0.41	0.29	11.72	0.87	failed	failed	failed	failed	1880	1886	failed	+	+	+	-
22	training	CEB211	34393	43	17	695	95%	46%	0.35	0.16	18.22	0.889	failed	failed	failed	failed	1048	1041	failed	+	-	-	-

Tableau supplémentaire 2 :

1331 12	(1331 01)	(1347 02)	1347 15	1416 11	(1416 01)	45 01	45 02	1345 01	1345 02	1420 01	1420 02
1331 13	1331 14	1362 13	(1362 01)	1416 12	1416 13	66 01	66 02	1346 01	1346 02	1421 01	1421 02
(1331 02)	1331 15	1362 14	1362 15	(1416 02)	1416 14	102 01	102 02	1349 01	1349 02	1423 01	1423 02
1332 13	(1332 01)	(1362 02)	1362 16	884 15	(884 01)	104 01	(104 02)	1350 01	1350 02	1424 01	1424 02
1332 14	1332 15	1413 18	(1413 01)	884 16	884 17	1334 01	1334 02	1375 01	1375 02	13291 01	13291 02
(1332 02)	1332 16	(1413 02)	1413 19	(884 02)	884 18	1340 01	1340 02	1377 01	1377 02	(13292 01)	13292 02
1347 12	(1347 01)	23 01	23 02	35 01	35 02	1341 01	1341 02	1408 01	1408 02	(13293 01)	13293 02
1347 13	1347 14	28 01	28 02	37 01	37 02	1344 01	(1344 02)	1418 01	1418 02	(13294 01)	(13294 02)

Annexe 4 : description des 50 minisatellites isolés dans des ARNs messagers

ARNm	Taille du motif	Nombre d'unités	Longueur totale	% conservation	%GC	HistoryR	Chromosome	Position de la répétition dans l'ARN	5', 3', Codant ?	Répétition d'acides aminés ?	Catégorie de locus
XM_114067	16	14,75	236	63%	81%	0,7	1p36.22	15-236	5'-codant	non	model, supported by EST alignments
XM_303126	49	8,22	403	97%	72%	1	1p36.32	900-1302	codant	49aa: VRQGWCWQLCGGAGLLGTAGVVLAALW RCRAAGYGRGGAGSSVAVPGCR	model, ab initio, with EST support
NM_152384	175	5,38	942	87%	77%	1	2q31.1	1851-2792	3'	/	**gene** with protein product, function unknown
XM_209485	127	4,87	619	92%	78%	1	2p23.3	442-1060	codant-3'	non	model, supported by mRNA alignments
XM_210394	20	56,5	1130	89%	78%	0,78	2p25.3	38-1167	codant	20aa: ACGAPGVPAEPPACLRSPLR	model, ab initio, with EST support
XM_292899	24	32,41	778	94%	72%	0,77	2p25.3	522-1299 (atrophin-1 region)	codant	8aa: DPTPVPSA	model, ab initio, with EST support
XM_293077	39	14,6	568	99%	70%	1	2q37.3	582-1149	codant	13aa: RSSYTLGRGPVAP	model, ab initio, with EST support
XM_293082	38	28,2	1071	97%	78%	0,85	2q37.3	21-1091	codant	38aa: IAAPASLSAAGSPSPPRRLSPLRGAHRRPG VSLRCGEP	model, ab initio, with EST support
NM_138385	42	5	212	95%	78%	1	4p16.3	225-436	5'	/	**gene** with protein product, function unknown
XM_209692	41	13,02	534	61%	78%	0,78	5p13.1	439-972	codant-3'	non	model, supported by mRNA alignments
XM_302163	43	41,3	1789	95%	73%	0,74	5p15.33	3778-5566	codant	43aa: SGDGDDTGPRRGVQLRGRRRHRPPQRG SAQGTATTQAPAEGFS	model, ab initio, with EST support
XM_209786	37	10,48	388	52%	78%	0,81	6p21.31	362-749	codant-3'	non	model, supported by mRNA alignments
XM_294174	38	30,9	1184	90%	70%	0,73	7q36.3	1033-2216	codant	38aa (et blocs de 13): DATPGAARKSPPQTLRQAPTGRAHHRRYV RRRPEEPTT)	model, ab initio, with EST support
XM_300953	49	5,2	255	91%	82%	1	7q36.2	434-688	codant-3'	non	model, supported by mRNA alignments
XM_302324	38	19,9	758	96%	71%	0,83	7q36.3	868-1625	codant	38aa: ALPVSSGRGRPPARCLLPQEEDGLPRAAG FLRKRTASR	model, ab initio, with EST support
XM_302334	56	7,5	420	92%	70%	0,8	7p22.3	191-610	codant	56aa: SGPRTASELPHATASSRGTQDRGRPPSSP TPRHPAEGLRTEDGLRALPRRGIRQRG	model, ab initio
XM_291268	15	16,2	243	59%	70%	0,78	8q24.3	253-495	codant	5aa: YPQGP	**gene** with protein product, function known or inferred: **GRINA**

ARNm	Taille du motif	Nombre d'unités	Longueur totale	% conservation	%GC	HistoryR	Chromosome	Position de la répétition dans l'ARN	5', 3', Codant ?	Répétition d'acides aminés ?	Catégorie de locus
NM_001807	33	15,9	526	92%	80%	0,92	9q34.3	1693-2218	codant	11aa:GAPPVPPTGDS	**gene** with protein product, function known or inferred: **CEL**
NM_024718	44	16,9	744	95%	71%	0,78	9q34.3	1140-1883	3'	/	**gene** with protein product, function unknown
XM_209730	51	4,58	234	97%	83%	1	9q32	422-655	codant-3'	non	model, supported by mRNA alignments
XM_294605	24	115	2759	95%	70%	0,91	9q34.3	360-3118	codant	8aa: DTPRPRDC	model, ab initio, with EST support
XM_305804	45	6,02	271	96%	70%	1	9q34.3	2100-2370	codant	15aa: ATLIPVPPVLSPLST	model, ab initio, with EST support
XM_172821	33	6,75	223	94%	74%	1	10q26.3	62-284	codant	11aa: TPASSSSAAPP	model, supported by mRNA alignments
NM_000797	48	8,81	423	91%	84%	0,8	11p15.5	740-1162	codant	16aa: PPAPGLPRGPCGPDCA	**gene** with protein product, function known or inferred: **DRD4**
XM_296010	32	10,34	331	92%	70%	0,84	11q13.2	2-332	codant	33aa: DLPQPTLPEEEWTYPSPSSQRGGPTPAQAPRGV	model, ab initio, with EST support
XM_303443	29	8,1	235	97%	74%	1	11q25	277-511	codant	29aa: PALGGRRVGDLPSVDGVWATCPLWTASGR	model, ab initio, with EST support
XM_208611	49	4,4	216	93%	84%	1	12q21.2	677-892	3'	/	model, supported by mRNA alignments
XM_062938	39	39,3	1530	92%	73%	0,71	13q34	197-1726	codant	13aa:LGGDRTHGDAPSR	model, ab initio, with EST support
XM_292156	79	4,4	352	89%	72%	0,85	13q14.11	434-785	codant	26aa: APIAGGEAPPTRQGLRASPSSPLALK (répétés 2,2 fois seulement)	model, ab initio, with EST support
XM_208767	40	10,35	414	74%	74%	0,72	14q32.33	371-784	codant-3'	non	model, supported by mRNA alignments
XM_301533	30	7,3	220	96%	74%	1	15q22.32	82-301	codant	10aa: EPGDGGKLPP	model, ab initio
XM_303976	33	7,1	234	96%	71%	1	16q24.3	1316-1549	codant	11aa: GYRNSLFPPGE	model, ab initio, with EST support
NM_000421	15	18	275	62%	79%	0,63	17q21-q23	1352-1626	codant	5aa: SSGGG	**gene** with protein product, function known or inferred: **KRT10**
NM_024510	48	4,8	230	98%	70%	1	17q25.3	871-1100	codant	16aa: VPEPVHRPQDPWHIPG	**gene** with protein product, function unknown
XM_302704	59	4,1	242	94%	72%	0,66	17	844-1085	codant	59aa: AGTSRASSRPRLHLLWDWPPQAPPGHPPAPGCSFCGTEPRRHLPGILPPQAAPSVGLSP	model, supported by mRNA and EST alignments
XM_046434	59	4,1	242	94%	72%	0,66	17q25.3	848-1089	codant	59aa: AGTSRASSRPRLHLLWDWPPQAPPGHPPAPGCSFCGTEPRRHLPGILPPQAAPSVGLSP	**gene** with protein product, function unknown
XM_210355	39	5,3	207	95%	76%	1	18q23	378-584	codant	13aa: GSRSGEPRAPTDR	model, ab initio
XM_301656	31	24,87	771	97%	74%	1	18p11.32	342-1112	codant	31aa: PPRTSARPPDPSSDLGQTSGPLLGPRPDLRT (et 10:ALLGPRPDLRT)	model, ab initio

ARNm	Taille du motif	Nombre d'unités	Longueur totale	% conservation	%GC	HistoryR	Chromosome	Position de la répétition dans l'ARN	5', 3', Codant ?	Répétition d'acides aminés ?	Catégorie de locus
NM_130761	24	8,6	206	96%	71%	1	19p13.3	692-897	codant	8aa: TSPESPDT	**gene** with protein product, function known or inferred: **MADCAM1**
NM_007346	60	7	421	94%	73%	0,85	20q13.3	1576-1996	codant	20aa: SPSETPGPRPAGPAGDEPAE	**gene** with protein product, function known or inferred: **OGFR**
NM_080425	36	8,13	293	92%	79%	1	20q13.2-q13.3	805-1097	codant	12aa: PDAPADPDSGAA	**gene** with protein product, function known or inferred: **GNAS**
XM_097702	65	5,38	350	95%	74%	1	20q13.33	2054-2403	3'	/	model, supported by mRNA alignments
XM_212468	56	14,28	800	90%	70%	1	20q13.33	19-818	codant	56aa: GRGRTRLCAHVPRARDSATEEDAPGCVHMCPGPGTAPRKRTHPAVWTCAQGPGQRH	model, ab initio, with EST support
NM_053277	30	13,2	396	82%	74%	0,7	21q22.12	465-860	codant	20aa: GPAGDSVDAEGRVGDSVDAE	**gene** with protein product, function known or inferred: **CLIC6**
XM_304616	56	5,53	310	87%	74%	0,66	21q22.3	792-1101	codant	56aa: TPHGWPPRTGPAPHSASHPPHLAGLSTPGRHPTAPVTHPTRLVSPHRAGTLQRQSP	model, ab initio, with EST support
XM_066408	33	32,6	1076	97%	71%	0,88	22q13.33	785-1860	codant	11aa: GEAQSFPVEES	model, ab initio, with EST support
XM_209396	127	4,83	614	90%	77%	1	22q13.2	443-1056	codant-3'	non	model, supported by mRNA alignments
XM_301914	33	8,1	267	97%	73%	1	22q13.33	139-405	codant	11aa: HTAPDTPAQPF	model, ab initio
XM_304642	18	27,4	494	89%	70%	0,95	22q13.1	103-596	5'	/	model, ab initio, with EST support
NM_153448	27	14,6	394	90%	75%	0,86	Xq22.1	784-1177	codant	9aa: VPPGPPMAP	**gene** with protein product, function known or inferred: **ESX1L**

Résumé

Les répétitions en tandem sont constituées de successions de motifs d'ADN. Ces structures sont présentes dans tous les organismes, procaryotes comme eucaryotes et, même si leur rôle biologique est encore peu compris, elles ont des applications dans de nombreux domaines. Tout d'abord, chez les bactéries, les répétitions en tandem polymorphes, dont le nombre d'unités varie, se révèlent un outil puissant pour l'identification de souches à des fins épidémiologiques. Par ailleurs, certaines répétitions en tandem humaines ont la propriété de muter à des fréquences élevées : les minisatellites hypermutables sont les éléments les plus instables du génome humain. Ils peuvent être utilisés comme biomarqueurs d'exposition à des agents potentiellement mutagènes tels que les radiations ionisantes. D'un point de vue plus fondamental, ils sont également un modèle d'étude des mécanismes d'instabilité des génomes. Dans cette thèse, nous mettons à profit les données issues du séquençage afin d'identifier des répétitions en tandem polymorphes. Nous avons tout d'abord élaboré une base de données des répétitions en tandem accessible sur le web (http://minisatellites.u-psud.fr), qui fournit un accès aux répétitions en tandem de génomes entiers. Ensuite, dans le but de sélectionner les répétitions en tandem polymorphes, plusieurs stratégies ont été mises en œuvre. D'une part, chez les bactéries pour lesquelles les séquences de plusieurs souches étaient disponibles, nous avons créé un utilitaire de comparaison de souches, afin d'identifier des marqueurs polymorphes utilisables en épidémiologie. D'autre part, une étude menée sur les minisatellites humains a permis de définir des critères prédictifs du polymorphisme à partir de la séquence d'un seul allèle de minisatellite, et a en outre mis en évidence un nouveau minisatellite hypermutable situé dans une séquence codante putative. Les critères prédictifs ont également été appliqués à l'identification de minisatellites codants potentiellement polymorphes dans le génome humain.

Abstract

Tandem repeats are consecutive occurrences of a DNA unit. Such structures are found in all organisms, prokaryotes as well as eukaryotes. Although their biological function is not fully understood, they have diverse practical applications. In bacteria, polymorphic tandem repeats (with varying copy numbers), are powerful tools for strain identification in bacterial epidemiology. In humans, some tandem repeats mutate at a very high rate: hypermutable minisatellites are the most unstable elements in the human genome. They can be used as biomarkers for the monitoring of genotoxic agents such as ionizing radiations. More fundamentally, they are a model to study genome instability. In this thesis, we take advantage of the release of genome sequence data in order to identify polymorphic tandem repeats. First, we developed a tandem repeats database freely accessible on the web (http://minisatellites.u-psud.fr), which provides tandem repeats from complete genomes. Then, we used several approaches to identify polymorphic tandem repeats. The first one consisted in developing a strain comparison page for bacteria with several sequenced strains, which directly selects polymorphic markers for use in epidemiology. The second approach, applied to human minisatellites, sought to define polymorphism predictive criteria from the sequence of one single allele. This study also identified a hypermutable minisatellite located in a putative coding sequence. The predictive criteria were subsequently used to search for potentially polymorphic coding minisatellites in the human genome.