Caractérisation et classification moléculaire des pathologies thyroïdiennes par approche transcriptomique

Jean-Fred Fontaine

UNIVERSITE D'ANGERS **2007**
N° d'ordre 854

CARACTERISATION ET CLASSIFICATION MOLECULAIRE DES PATHOLOGIES THYROIDIENNES PAR APPROCHE TRANSCRIPTOMIQUE

THESE DE DOCTORAT

Spécialité : Biologie Cellulaire

ECOLE DOCTORALE D'ANGERS

Présentée et soutenue publiquement

le : 7 Décembre 2007

à : Angers

par : Jean-Fred FONTAINE

Devant le jury ci-dessous :

Mr. Philippe Dessen (rapporteur), Directeur de recherche, CNRS
Mme. Brigitte Franc (rapporteur), Professeur des Universités, Hôpital Ambroise Paré
Mr. Ralf Paschke (examinateur), Professeur des Universités, Université de Leipzig
Mme. Frédérique Savagner (examinateur), Maître de conférences, Université d'Angers
Mr. Rémi Houlgatte (examinateur), Directeur de recherche, INSERM
Mr. Yves Malthièry (directeur de thèse), Professeur des Universités, Université d'Angers

Nom et coordonnées du laboratoire : INSERM U694, Laboratoire de Biochimie et Biologie Moléculaire du C.H.U., 4 rue Larrey, 49933 Angers cedex 01

ED 363

Celui qui connaît les hommes est prudent.

Celui qui se connaît lui-même est éclairé.

Celui qui dompte les hommes est puissant.

Celui qui se dompte lui-même est fort.

Celui qui sait se suffire est assez riche.

Celui qui agit avec énergie est doué d'une ferme volonté.

Celui qui ne s'écarte point de sa nature subsiste longtemps.

Celui qui meurt et ne périt pas jouit d'une (éternelle) longévité.

Dao De Jing, chapitre 33, Lao Zi (VIe siècle av. J.-C.), traduction de Stanislas Julien (1842)

A ma famille pour son sincère et irremplaçable soutien

Elise, Sully, Eugène, Valencia, Gudrun, Thomas, Tina, Mario

A Jana sans qui tout aurait été différent

Remerciements

Je remercie le Pr. Yves Malthièry de m'avoir donné la possibilité d'effectuer ce travail de thèse. Ta vision de la biologie et de l'informatique m'a toujours passionnée. J'ai eu beaucoup de chance de trouver un mentor à ce moment de ma vie professionnelle.

Je remercie le Dr. Frédérique Savagner pour son excellent encadrement scientifique et sa vision humaine de la Recherche. Ces années ont été très enrichissantes, j'espère que nous continuerons à collaborer dans le futur.

Je remercie le Dr. Philippe Dessen d'avoir accepté d'être rapporteur de cette thèse.

Je remercie le Pr. Brigitte Franc d'avoir accepté d'être rapporteur de cette thèse. Ta participation à ce travail a été déterminante. J'ai pu profiter de ton enthousiasme ainsi que de ta grande et passionnante expertise des pathologies thyroïdiennes.

Je remercie le Dr. Rémi Houlgatte d'avoir toujours été bienveillant avec moi. J'ai beaucoup appris à ton contact. Ton dynamisme reste un modèle dont j'aime m'inspirer.

Je remercie le Pr. Ralf Paschke d'avoir accepté d'être membre de mon jury de thèse. Je vous remercie de la confiance que vous m'avez montré, ainsi que du temps que vous avez pris pour m'aider à définir mes projets futurs.

Je remercie le Dr. Caroline Jacques de sa collaboration scientifique très intéressante et de toute son aide au cours de ces années passées au laboratoire.

Je remercie le Pr. Pascal Reynier de son aide et de sa bienveillance, notamment pour l'enseignement à l'U.F.R. Sciences médicales de l'Université d'Angers.

Je remercie les étudiants et membres du laboratoire qui ont été accueillant et fort sympathiques.

Je remercie tous mes amis que j'ai connus sur Angers de leur sincère amitié et de leur soutien : Adrien, Émilie, Sébastien, Magalie, Gyasi, Clothilde, Laurent, Daphné, Hélène, Sophie, Clémentine.

Table des matières

1 ABREVIATIONS 11

2 INTRODUCTION 13

2.1 EPIDEMIOLOGIE DU CANCER THYROÏDIEN 13

2.2 CLASSIFICATION DE LA PATHOLOGIE THYROÏDIENNE 16

2.2.1 La glande normale 16

2.2.2 La pathologie thyroïdienne non maligne 17

2.2.3 Les grands types de cancers thyroïdiens 18

2.3 DIFFICULTES DE CLASSIFICATION 21

2.3.1 Variations intra- et inter-observateur 21

2.3.2 Exemples de cas difficiles 23

2.4 METABOLISME ENERGETIQUE DE LA CELLULE CANCEREUSE 24

2.5 TECHNOLOGIE A HAUT DEBIT POUR LA RECHERCHE DE MARQUEURS 25

2.5.1 Les biopuces ou puces à ADN 26

2.6 ANALYSE DE DONNEES TRANSCRIPTOMIQUES A HAUT DEBIT 27

2.6.1 Design et sources de variations 28

2.6.2 Traitement des données 33

2.6.3 Analyse des données 33

2.7 PROJET DE THESE 41

3 ARTICLES ET BREVET 43

3.1 ARTICLE 1 43

3.1.1 Introduction 43

3.1.2 Article 45

3.1.3 Résultats complémentaires 68

3.1.4 Discussion 74

3.2 BREVET EUROPEEN 76

3.2.1 Titre de l'invention 76

3.2.2 Champ de l'invention ... *76*

3.2.3 Résumé de l'invention ... *76*

3.3 ARTICLE 2 ... 77

3.3.1 Introduction ... *77*

3.3.2 Article ... *78*

3.3.3 Discussion ... *108*

3.4 ARTICLE 3 ... 109

3.4.1 Introduction ... *109*

3.4.2 Article ... *110*

3.4.3 Discussion ... *135*

4 DISCUSSION ... 136

5 MATERIELS ET METHODES ... 140

5.1 TISSUS BIOLOGIQUES ... 140

5.2 METHODES POUR DETERMINER LA TAILLE DES GROUPES : ... 140

6 REFERENCES ... 142

Figures et tables

Figure 1 : Risque relatif de cancer ... 15

Figure 2 : Taille optimale des groupes ... 69

Figure 3 : Information apportée par 3 échantillons ... 72

Figure 4 : Information apportée par 3 échantillons dans un jeu de données indépendant ... 73

Table 1 : Nombre de cas observés dans 10 registres français du cancer ... 14

Table 2 : logique de test ... 31

1 Abréviations

- ADN : Acide désoxyribonucléique
- ARN : Acide Ribonucléique
- ARNm : Acide Ribonucléique messager
- AT : Thyroïdite Auto-immune
- ChIP-chip : Immuno-Précipitation de la Chromatine sur biopuce (chip)
- FTA : Adénome Folliculaire Thyroïdien
    - FTA-a : FTA d'architecture macrofolliculaire
    - FTA-b : FTA d'architecture microfolliculaire
- FTC : Carcinome Folliculaire Thyroïdien
- GD : maladie de Basedow (Graves' Disease)
- MNG : Goitre Multi-Nodulaire
- OTA : Adénome Oncocytaire de la Thyroïde
- OTC : Carcinome Oncocytaire de la Thyroïde
- PTC : Carcinome Papillaire Thyroïdien
- T-UM : Tumeur thyroïdienne de Malignité Incertaine

2 Introduction

2.1 Epidémiologie du cancer thyroïdien

L'accident de Tchernobyl, en 1986, a mis en lumière le cancer de la thyroïde. Les cancers de la thyroïde représentent 4 000 à 5 000 nouveaux cas par an en France en 2000. Ils sont plus fréquents chez la femme que chez l'homme, avec un sex-ratio de 1/3,5. Leur augmentation est ancienne et continue depuis 1975. Elle a commencé avant 1986 et ne semble pas s'être accélérée après 1986. Cette augmentation est également constatée dans la plupart des pays d'Europe de l'Ouest mais aussi aux Etats-Unis, non touchés par cet accident, avec une augmentation très similaire en France et aux Etats-Unis. La répartition Est-Ouest sur notre territoire rend peu plausible l'hypothèse d'un effet Tchernobyl (Table 1). Cette augmentation pourrait notamment être liée à une meilleure détection de ces cancers consécutive à l'évolution des pratiques de diagnostic et de traitement des maladies de la thyroïde. Des cancers de plus en plus précoces sont découverts de façon fortuite en explorant et traitant la thyroïde pour une pathologie bénigne. Sur les pièces d'autopsie, on retrouve des structures cancéreuses sur 6 à 20% des pièces, alors que l'incidence clinique est de 2 à 10 cas pour 100 000 habitants par an. Cette différence entre incidences autopsiques et clinique pose le problème de l'évolution naturelle de ces tumeurs.

Plusieurs variétés de cancers thyroïdiens sont connues, dont une forme grave à prédisposition héréditaire. La plupart de ces cancers sont des cancers différenciés à progression lente. La gravité de cette pathologie tient à sa capacité à développer des métastases, essentiellement osseuses.

Pour les cancers différenciés sporadiques (papillaires, PTC, et folliculaires, FTC), aucune étiologie n'est retrouvée dans 95% des cas. Le facteur étiologique le moins contesté est l'irradiation de la thyroïde. La latence après une exposition est variable, en moyenne de 15 à 20 ans, mais parfois beaucoup plus courte, ce qui est le cas des cancers survenus en Ukraine et Biélorussie après l'accident de Tchernobyl. Ces cancers sont généralement de type papillaire. Les autres facteurs prédisposant sont plus rares et mal connus. Une prédisposition familiale est relevée dans 20 à 30% des cancers médullaires (7% des cancers thyroïdiens). Cette prédisposition est strictement corrélée avec une mutation de l'oncogène Ret localisé sur le chromosome 10. La transmission est de type autosomique dominant. Ces cancers rentrent le plus souvent dans le cadre des Néoplasies Endocriniennes Multiples de type 2 (NEM2A et

NEM2B). Par ailleurs, 3 à 5% des patients atteints de cancers papillaires (80% des cancers thyroïdiens) ont un parent atteint lui-même d'un cancer de la thyroïde. Pourtant aucun marqueur génétique n'a été identifié.

De rares maladies génétiques sont associées à des cancers de la thyroïde : polypose colique et syndrome de Gardner, syndrome de Carney, syndrome de Cowden. Certaines mutations sont parfois retrouvées dans les tumeurs thyroïdiennes. Le gène B-Raf est parfois retrouvé muté dans les cancers papillaires, de même que le gène Ras. Le gène Ret peut subir des réarrangements. Ces mutations ne sont pas des marqueurs de prédisposition mais des marqueurs de gravité tumorale ou de sensibilité thérapeutique.

Le traitement de ces tumeurs fait appel à la chirurgie (pour les tumeurs en place) et à l'iode radioactif (pour les métastases des tumeurs différenciées). La caractérisation des diverses variétés de cancers thyroïdiens (définissant leur agressivité) est apportée par les analyses anatomopathologiques. Au cours des années passées, les études d'expression des gènes (par biopuces) ont permis de mieux comprendre les mécanismes moléculaires impliqués dans le développement de ces tumeurs et d'affiner le diagnostic.

Table 1 : Nombre de cas observés dans 10 registres français du cancer

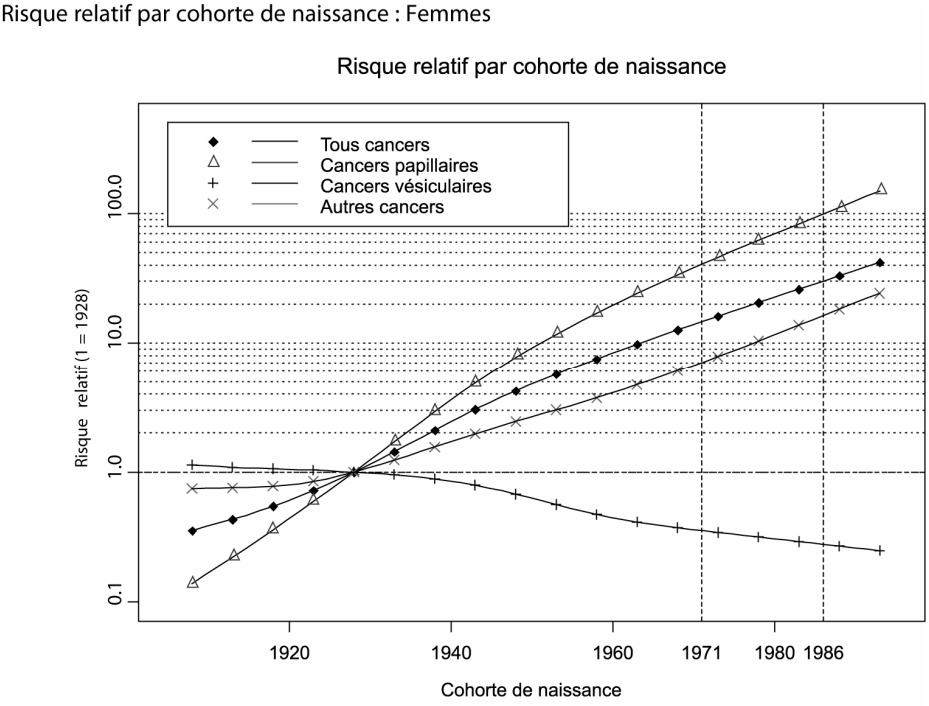
Registre	Période	Papillaires	Folliculaires	Autres*	Tous cancers
Calvados	1982-1998	379	87	65	531
Doubs	1982-2001	332	85	73	490
Hérault	1986-2000	390	94	106	590
Isère	1982-2000	681	156	143	980
Manche	1994-2001	205	21	33	259
Marne Ardennes	1982-2001	731	207	120	1058
Bas-Rhin	1982-2001	398	114	178	690
Haut-Rhin	1988-2001	200	42	83	325
Somme	1982-1998	154	58	68	280
Tarn	1982-2000	378	49	43	470
Total		**3848**	**913**	**912**	**5673**

(*) Cancers anaplasiques, oncocytaires, médullaires, sans précision et non spécifiques.

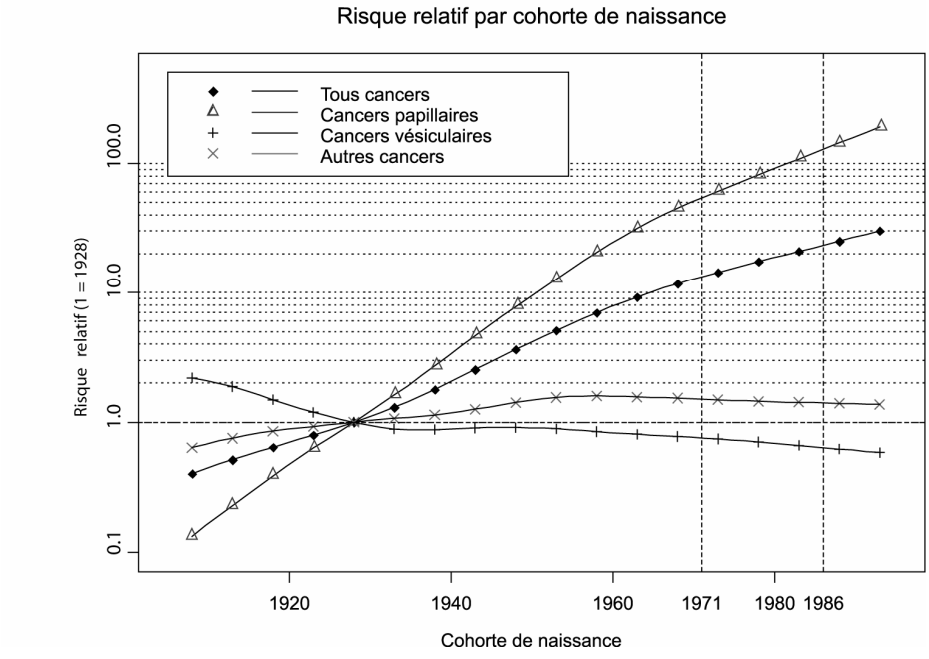
Données de l'Institut de veille sanitaire (Chérié-Challine et al., 2006).

Figure 1 : Risque relatif de cancer

Risque relatif par cohorte de naissance : Femmes



Risque relatif par cohorte de naissance : Hommes



Données de l'Institut de veille sanitaire (Chérié-Challine et al., 2006)

2.2 Classification de la pathologie thyroïdienne

(Franc 2007 ; Niccoli-Sire et Conte-Devolx, 2007 ; DeLellis *et al.*, 2004 ; Hedinger *et al.*, 1988)

La pathologie thyroïdienne non maligne est représentée principalement par les lésions bénignes nodulaires ou diffuses (thyroïdites et maladie de Basedow). Les tumeurs des cellules folliculaires sont soit bénignes (adénomes folliculaires : FTA) représentant 95% des tumeurs, soit malignes (carcinomes différenciés PTC et FTC, peu différenciés, indifférenciés ou anaplasiques) représentant 5% des tumeurs de la thyroïde. Etant donné que les tumeurs thyroïdiennes différenciées folliculaires bénignes ou malignes sont souvent encapsulées, la limite entre le bénin et le malin peut être difficile à évaluer. Les tumeurs des cellules C sont les cancers médullaires. Toute pathologie thyroïdienne qui entraîne une augmentation de volume du corps thyroïde s'appelle un goitre, celui-ci peut être uni ou bilatéral, diffus ou nodulaire, uni ou multinodulaire.

2.2.1 La glande normale

Entouré d'une capsule fibreuse continue, le tissu thyroïdien est organisé en lobules de 20 à 40 follicules (synonyme vésicules). Une seule couche de cellules épithéliales les bordent, ce sont les thyrocytes (synonymes : cellules folliculaires, cellules vésiculaires). La lumière du follicule contient de la colloïde, riche en thyroglobuline (TG). L'aspect des follicules, des thyrocytes, de la colloïde, varie en fonction de l'activité de la glande. Un stroma vascularisé entoure chaque follicule.

Quand la cellule thyroïdienne est active, le noyau s'élargit, la chromatine se disperse, le cytoplasme s'agrandit vers le pole apical de la cellule, devient plus pale, tandis que le noyau se trouve en position basale. On distingue 3 phases cellulaires : les cellules au repos aplaties, les cellules cubiques secrétantes, les cellules cylindriques qui « excrètent » les hormones thyroïdiennes dans les capillaires qui les bordent.

Les cellules C, ou cellules parafolliculaires, constituent un deuxième contingent cellulaire qui sécrète de la calcitonine (CT). Ces cellules peu nombreuses, moins de 1% des cellules thyroïdiennes, sont difficiles à identifier sur les colorations standards. Situées en général à la jonction 1/3 supérieur et moyens des deux lobes, elles peuvent être dans la zone des vestiges des corps ultimobranchiaux ou « amas cellulaires compacts» (solid cell nests) .

2.2.2 La pathologie thyroïdienne non maligne

La maladie de Basedow est une hyperplasie diffuse de la glande. Elle peut parfois prêter à confusion avec une lésion maligne au regard de certains paramètres (végétations papillaires, des noyaux très atypiques, une très mauvaise limitation périphérique de la glande). L'importance des infiltrats inflammatoires lymphocytaires est variable d'un cas à l'autre, et souvent le fait de formes récidivantes. On peut trouver au sein d'une maladie de Basedow un cancer, le plus souvent papillaire, dont le diagnostic repose sur les règles usuelles.

On distingue 3 grands types de thyroïdites: la thyroïdite lymphocytaire auto-immune de Hashimoto, la thyroïdite fibreuse de Riedel et la thyroïdite granulomateuse de Quervain.

La thyroïdite de Hashimoto est en général une lésion diffuse de la glande qui peut poser des problèmes diagnostiques avec un cancer, ou un lymphome en raison de l'importance des infiltrats lymphocytaires, des modifications épithéliales, surtout dans les phases hyperplasiques avec des noyaux clarifiés. Il peut exister un cancer associé, le plus souvent papillaire.

La thyroïdite granulomateuse de Quervain, consiste en une destruction très focale des vésicules thyroïdiennes avec foyers de nécrose et de polynucléaires et production de très volumineuses cellules géantes, colloïdophagie, histiocytes dans les lumières folliculaires, suivie d'une phase de réparation.

La thyroïdite de Riedel, ou thyroïdite fibreuse se présente comme un cancer au moment de l'intervention, correspondant à une destruction fibreuse unilatérale thyroïdienne et péri thyroïdienne. Un collagène dense, pseudo-chéloïde associé à un infiltrat lympho plasmocytaire et polynucléaire éosinophile modéré remplace le tissu thyroïdien. Une thrombose veineuse est souvent associée dont la paroi présente des infiltrats inflammatoires.

Les adénomes sont des tumeurs folliculaires bénignes très fréquentes, encapsulées, dépourvues de signes d'invasion. Ces tumeurs se présentent comme des nodules de taille variable, remaniés ou non en leur centre, de texture colloïde ou plus ferme, en fonction des architectures dominantes qui les constituent. Lorsque l'adénome est « toxique », les cellules tumorales pourront présenter des signes d'activité. On peut rencontrer dans tous les adénomes qu'ils soient ou non fonctionnels, de même que dans les autres pathologies non malignes des formations papillaires endofolliculaires.

Il est parfois difficile d'être certain de la bénignité d'un adénome. Il est également difficile lorsqu'une thyroïde renferme de nombreux nodules de faire la distinction entre un adénome et une hyperplasie nodulaire, souvent dite « adénomateuse ».

2.2.3 Les grands types de cancers thyroïdiens

La classification des tumeurs thyroïdiennes de l'OMS 2004 précise beaucoup de variantes aux grandes catégories de cancers. Pour la plupart, les variantes sont trop rares et il n'est donc pas possible d'en connaître l'évolutivité.

2.2.3.1 Le carcinome papillaire

Le carcinome papillaire représente 4% des tumeurs de la thyroïde, soit environ 4000 cas par an en France. Son diagnostic repose sur un noyau particulier : augmenté de volume, chevauché, avec des incisures multiples, un aspect clair et dépoli, et pouvant comporter de pseudos inclusions cytoplasmiques intranucléaires. La présence de papilles (axes conjonctivo-vasculaires bordés de cellules tumorales épithéliales), n'est pas nécessaire au diagnostic.

En cas de tumeur encapsulée, d'architecture folliculaire, sans formations papillaires, sans signe d'invasion, le diagnostic de PTC ne repose que sur l'aspect du noyau. Le niveau de reproductibilité du diagnostic devient plus faible.

La forme usuelle, d'architecture papillaire dominante représente moins de 60% des PTC. Les autres formes constituent les variantes. La prévalence en est mal connue en dehors du variant folliculaire (FVPTC, 20% dont 1/3 est encapsulé), des PTC à cellules hautes (4-5%), des microcarcinomes, tous variants confondus 30%. Les cancers à cellules hautes seraient des tumeurs des sujets plus âgés souvent de sexe masculin, avec une plus grande agressivité tumorale, un risque plus important de métastase à distance lors du diagnostic initial.

Une place à part doit être faite aux formes de l'enfant : solides chez le petit enfant, ou sclérosante diffuse. Celle-ci, est plus fréquente chez l'enfant ou l'adolescent avec ou sans contexte d'exposition aux radiations ionisantes. Décrite aussi dans <1% des PTC de l'adulte, elle se présente souvent d'emblée avec une atteinte pulmonaire (25% des cas décrits), sans évolution plus péjorative manifeste. Le caractère diffus de la lésion, le contexte de thyroïdite lymphocytaire avec anticorps positifs, en retarde souvent le diagnostic.

2.2.3.2 Le carcinome folliculaire

Le diagnostic du FTC repose sur une différenciation folliculaire, des critères de comportement tumoral (invasion capsulaire, invasion vasculaire), et l'absence anomalies nucléaires caractéristiques du cancer papillaire. Par rapport à sa contrepartie bénigne, les adénomes folliculaires, les critères architecturaux ou cellulaires ne sont pas discriminants. Le diagnostic de FTC encapsulé à invasion minime de la précédente classification de l'OMS (Hedinger *et al.*, 1988), peut être porté sur le seul fait d'une invasion capsulaire. La valeur diagnostique en est controversée tant cette forme est indolente. Les études de reproductibilité ont démontré la faible reproductibilité du diagnostic de FTC encapsulé à invasion minime, ainsi que des invasions vasculaires quand elles sont inférieurs à 4. Les formes encapsulées à tendance angio -invasive présentent plusieurs invasions vasculaires indéniables. Son pronostic est moins bon que celui du carcinome à invasion minime.

2.2.3.3 Les tumeurs de potentiel de malignité incertain

Lorsqu'un doute sur le diagnostic existe entre un adénome et un carcinome, la notion de tumeur de potentiel de malignité incertain est utilisée (UMP, uncertain malignant potential). La classification OMS de 2004 admet dans les cas difficiles les termes de « tumeur différenciée UMP » pour un doute avec un variant folliculaire de PTC, et de « tumeur folliculaire UMP » pour un doute avec un FTC. Les critères proposés pour le diagnostic sont la présence d'images douteuses d'invasion vasculaire ou capsulaire, l'absence de noyau de type PTC pour les « tumeurs folliculaires UMP », la présence de noyaux dont la morphologie est à doute diagnostic pour un PTC dans le cas des « tumeurs différenciées UMP».

2.2.3.4 Les tumeurs oncocytaires

Les tumeurs oncocytaires, ou tumeurs à cellules de Hurtle, ne sont pas considérés comme une catégorie spécifique dans la classification internationale. Les bases de données épidémiologiques nord américaines permettent de savoir qu'elles représentent 3% des cancers thyroïdiens. Le phénotype oncocytaire est lié à la richesse mitochondriale des cellules, n'est pas synonyme de tumeur, est présent dans la thyroïdite de Hashimoto. Ce type de tumeur peut affecter de nombreux organes tels que le rein, la trachée, le pharynx, les glandes salivaires et la glande thyroïde (Asa 2004 pour revue). Dans la plupart des tissus, le développement d'un oncocytome est généralement bénin. Dans le tissu thyroïdien, où il est retrouvé plus fréquemment, son caractère oncocytaire est péjoratif.

On parle de tumeur oncocytaire lorsque plus de 75% des cellules de la tumeur sont oncocytaires. Si la tumeur est encapsulée, sans invasion vasculaire, d'architecture folliculaire, avec ou sans papille, sans noyau de type papillaire, elle est classée parmi les adénomes oncocytaires. Si la tumeur présente des invasions vasculaires, elle est classée en FTC variante oncocytaire. Si la tumeur quelque soit son architecture, possède des noyaux de type PTC elle est classée parmi les PTC à cellules oncocytaires.

Plusieurs études ont évoqué une classe supplémentaire, le cancer papillaire oxyphile (Herrera *et al.*, 1992, Beckner *et al.,* 1995, Berho and Suster 1997). Salvatore *et al*. (2004) proposent une recherche systématique des anomalies et réarrangements des oncogènes B-Raf et RET/PTC, spécifiques des PTC, pour identifier les variants oxyphiles de ces tumeurs. Actuellement plusieurs marqueurs de la tumorigenèse des PTC (mutations de la protéine B-Raf, réarrangements des oncogènes RET/PTC et TRK) et des cancers folliculaires (mutations du récepteur à la TSH, de la protéine RAS, augmentation de la quantité de Reactive Oxygen Species [ROS]) sont connus. Toutefois, ces anomalies moléculaires sont retrouvées très rarement dans les oncocytomes thyroïdiens (Gozu *et al.,* 2005, Collins *et al.,* 2003, Salvatore *et al.,* 2004). Il a été proposé précédemment qu'ils constituent une classe à part mais la présence du réarrangement RET/PTC dans les cancers papillaires oxyphiles conduit à un comportement et une évolution classiques des cas papillaires.

2.2.3.5 Les carcinomes peu différenciés

Les carcinomes peu différenciés sont définis comme des tumeurs d'origine folliculaire, dont la différenciation folliculaire architecturale demeure limitée et dont les caractéristiques tant morphologiques que comportementales sont de position intermédiaire entre les carcinomes différenciés et les carcinomes indifférenciés. Les caractéristiques sont : -architecturales (insulaires, trabéculaires, ou solides dominantes), et -comportementales (invasions vasculaires manifestes, un caractère infiltrant, foyers de nécrose).

2.2.3.6 Les cancers anaplasiques

Le cancer anaplasique est une tumeur rare dont une partie ou la totalité des cellules est indifférenciée. De caractère malin, très agressif et extrêmement invasif, le pronostic de cette tumeur est très mauvais. Développée de novo ou sur une tumeur plus différenciée pré existante, seuls les marqueurs immunohistochimiques ou ultrastructuraux révèlent sa nature épithéliale. Ces tumeurs sont composés d'une proportion variable de cellules fusiformes, géantes

pléomorphes et plus «épithélioïdes». Elles sont plus épithéliales, malpighiennes, dans 20 à 30% des cas. On y décrit des ostéoclastes. Nécrose, mitoses et extension dans les murs vasculaires sont la règle. Les infiltrations à polynucléaires sont fréquentes, avec parfois une polynucléose sanguine.

2.2.3.7 Les tumeurs médullaires

Le cancer médullaire représente 0,25 à 0,5% des tumeurs de la thyroïde. Les cellules C sont à l'origine de ce type de tumeur, et peuvent présenter une hyperplasie (Niccoli-Sire et Conte-Devolx, 2007). La cytoponction en fait le diagnostic par immunomarquage anti calcitonine (CT) positif mais elle est mise en défaut dans plus de la moitié des cas. Il existe une bonne corrélation entre la concentration de CT et la masse tumorale. L'antigène carcino-embryonnaire est un marqueur moins spécifique que la CT mais un témoin de dédifférenciation donc un indicateur de mauvais pronostic. Les formes familiales représentent plus d'un tiers des cancers médullaires de la thyroïde. Elles s'intègrent dans la néoplasie endocrinienne multiple de type 2, affection héréditaire, autosomique dominante à pénétrance complète, rattachée à l'existence d'une mutation germinale de l'oncogène RET.

2.3 Difficultés de classification

2.3.1 Variations intra- et inter-observateur

Il existe 3 études qui ont spécifiquement analysé les différences intra- et inter-observateur. La première de Hirokawa *et al*. (2002) a comparé le diagnostic de 8 experts pour 21 nodules thyroïdiens. Ces experts ont unanimement diagnostiqué 10% des cas. En considérant seulement 7 experts le pourcentage monte à 29%, alors qu'il atteint 76% avec 6 experts. Le taux d'agrément entre experts du même pays est plus élevé que le taux global. Il est de 33% chez les américains et de 52% chez les japonais. Certaines distinctions sont plus simples que d'autres. En effet, la séparation des échantillons bénins et malins réunit les experts dans 62% des cas.

Par contre, certains critères sont moins évidents. Le critère principal permettant de séparer les FTAs et FTCs est la présence d'invasion vasculaire ou capsulaire plus ou moins prononcée. Cette définition implique l'existence de FTCs non invasifs (intra capsulaires) qui sont automatiquement considérés comme des FTAs. Les FTCs extensifs ne posent pas de problème (invasion extensive des vaisseaux sanguins, invasion et pénétration de la capsule et parfois extensions extra-thyroïdales). D'autre part, la reconnaissance des FTC peu invasifs souffre de

la difficulté et de la subjectivité à évaluer l'invasion. La présence de noyaux en verre dépoli est un des critères de reconnaissance des PTC et FVPTC mais il n'en existe pas de définition quantitative précise. Bien que les définitions histologiques des lésions folliculaires soient les mêmes à travers le monde, leur utilisation en critères diagnostiques est différente et subjective, ce qui entraine des appréciations différentes selon les spécialistes. Des paramètres supplémentaires (IHC, génétiques) peuvent aider les pathologistes, sans atteindre une séparation parfaite des groupes. Des définitions plus précises qualitatives et quantitatives sont nécessaires pour réduire la subjectivité de ces évaluations.

La seconde étude de Franc *et al*. (2003) sur 41 lésions folliculaires a mis en jeu le diagnostic de 5 experts. Un diagnostic de référence a été établi lors d'une réunion consensuelle. Les critères de décision étaient fixés à l'avance pour *différenci*er les carcinomes peu ou très invasifs. Dans le premier cas il s'agissait de trouver une invasion vasculaire non équivoque et / ou une invasion sur toute l'épaisseur de la capsule. Dans le second cas il s'agissait de trouver une large infiltration de vaisseaux sanguins et / ou de tissu thyroïdien adjacent avec manque d'une capsule complète, ou bien plus de 4 intrusions de vaisseaux sanguins dans une tumeur encapsulée. Le diagnostic consensuel a identifié 30 tumeurs malignes dont 24 FTC sur un total de 41 cas. Il montre clairement qu'un des experts a un leadership prononcé dans la décision finale. Un diagnostic unanime fut trouvé dans 13 cas sur 24 FTC. Certains cas ne posent pas de problème : la malignité des FTCs très invasifs, l'identification des tumeurs à cellules claires, et le caractère oxyphile de certaines tumeurs. Des problèmes de reproductibilité du diagnostic sont observés avec les FTC peu invasifs et les tumeurs atypiques, avec un accord inter- et intra-observateur égale à 0.23 et 0.68, respectivement. En effet, l'évaluation de la présence d'invasion vasculaire et capsulaire est subjective. Un cas particulièrement difficile celui du FTC à invasion minime qui mène à une faible reproductibilité inter-observateur, mais meilleure intra-observateur. Les auteurs mettent en question la validité d'une classification internationale basée sur des critères non ou peu reproductibles.

La troisième étude de Lloyd *et al*. (2004) sur 87 variants folliculaires de PTC reporte un diagnostic concordant de 10 experts à une fréquence cumulée de 39% (66.7% pour les cas métastatiques). L'unanimité est atteinte avec 7 experts. Les auteurs proposent une hiérarchie de critères pour reconnaître les FVPTC afin d'améliorer la reproductibilité du diagnostic. Le premier critère favorisé par les spécialistes est la présence d'invaginations cytoplasmiques dans le noyau. Elles s'observent dans 25 à 50% des PTC, dans aucun FTA et rarement dans les FTC. Le deuxième critère est l'abondance de « cannelures » nucléaires (nuclear grooves). Tous les

échantillons papillaires en comportent mais peu de FTA (10,8%) et pas de FTC. Le troisième critère important est l'aspect en verre dépoli des noyaux. Tous les PTC en comportent. Les FTA peuvent en avoir mais de façon plus diffuse.

2.3.2 Exemples de cas difficiles

La complexité des phénotypes associée à l'hétérogénéité cellulaire des tumeurs de la thyroïde ne permettent pas de définir des critères exclusifs pour classer toutes les tumeurs. Les cas les plus difficiles mènent à un classement intermédiaire ou incertain. Dans le cas des tumeurs folliculaires adénomateuses ou carcinomateuses, l'invasion vasculaire et capsulaire qui les différencie reste difficile à évaluer. Par définition, un adénome peut être considéré comme un carcinome non invasif. On peut aussi distinguer les FTCs à invasion minime des FTCs extensifs. Ces derniers sont facile à détecter car ils présentent une invasion extensive de vaisseaux sanguins, une invasion et une pénétration de la capsule, et parfois une extension extra thyroïdienne. Au contraire, les FTC peu invasifs sont moins évidents puisque l'invasion des cellules tumorales dans le vaisseau reste circonscrite localement.

Les tumeurs bénignes sont parfois difficiles à distinguer entre elles. Les adénomes folliculaires se différencient peu des nodules hyperplasiques Ces derniers sont fréquemment multiples et tendent à avoir une capsule incomplète. Leurs follicules sont caractérisés et de tailles variables mais ils présentent une configuration papillaire grossière. On peut aussi noter l'infiltration de cellules inflammatoires chroniques ainsi que des changements dégénératifs. Il existe une hypothèse de continuité entre ces cas. Leur distinction relative n'est pas possible sans critères plus fins et représente un faible intérêt clinique car le traitement et le pronostic sont identiques.

Le cas des variants folliculaires de carcinomes papillaire divise aussi les experts. Effectivement, il s'agit d'une tumeur aux motif de croissance folliculaire mais qui présente des noyaux en verre dépoli (que l'on trouve parfois dans les nodules et adénomes folliculaires), des invaginations cytoplasmiques intranucléaires (pseudo inclusions), des sillons et des chevauchements (« molding ») nucléaires. Le critère classant principal des tumeurs papillaires étant la présence de noyaux en verre dépoli, il est donc nécessaire d'avoir des mesures qualitatives et quantitatives plus précises.

Ainsi, il est apparu par nécessité des classes intermédiaires comme celles des tumeurs de potentiel de malignité incertain (décrites ci-dessus). On peut aussi trouver des commentaires concernant certaines tumeurs bénignes ne précisant qu’une malignité exclue. Il serait

intéressant de trouver des critères ou mesures plus précises et complémentaires pour assister la décision du spécialiste. Certaines mesures protéiques sont déjà utilisées (galectin-3, HBME-1, cytokeratin-19, TPO), ainsi que certains statuts mutationnels (PAX8/PPARG, RET/PTC, BRAF, RAS), mais ceux-ci n'ont pas une spécificité ou un taux de pénétrance parfaits. L'approche transcriptomique et la définition de signatures spécifiques est un exemple précis de technique complémentaire à haut débit.

2.4 Métabolisme énergétique de la cellule cancéreuse

(Malthièry et Savagner, 2006)

Les tumeurs cancéreuses développent majoritairement un métabolisme plus glycolytique qu'oxydatif. La cellule tumorale augmente son métabolisme en fonction de ses besoins spécifiques. L'augmentation de la consommation énergétique, de la synthèse des protéines et des acides nucléiques est le témoin de la prolifération cellulaire et pour partie de l'agressivité tumorale. Même si les modifications métaboliques ne sont pas liées à des modifications géniques, les métabolismes sont influencés par des oncogènes exprimés spécifiquement dans ce contexte. L'exemple de deux tumeurs riches en mitochondries, les paragangliomes et les oncocytomes thyroïdiens, montre que des déficits spécifiques induisent des réponses différentes.

La réorientation métabolique de la cellule cancéreuse vers la voie glycolytique n'est pas liée à une anomalie des mitochondries mais à une adaptation cellulaire sous-tendue par une augmentation de la captation et du transport du glucose, ainsi qu'à des modifications enzymatiques liées aux nécessités de la cellule cancéreuse. Cette réorientation métabolique confère un avantage sélectif à la cellule cancéreuse qui se définit, en pathologie, par une plus grande agressivité. Elle est d'ailleurs utilisée par le tissu sain lors des grands besoins de prolifération (renouvellement cellulaire et cicatrisation).

La voie oxydative de la dégradation du glucose est essentiellement dévolue à la production énergétique. A travers cette voie mitochondriale, une molécule de glucose génère 36 à 38 molécules d'ATP, là où la voie glycolytique n'en fourni que 2 à 4. La sélection de cette voie au cours de l'évolution de la cellule eucaryote peut s'expliquer par ses grands besoins énergétiques (mobilité, communications intercellulaires, adaptation à l'environnement, homéostasie thermique…). La voie glycolytique plus ancienne devait permettre de répondre aux besoins vitaux primaires (production énergétique de base et production des précurseurs

métaboliques impliqués dans la synthèse des protéines et des acides nucléiques). Cette voie est toujours indispensable à la cellule eucaryote pour ces mêmes raisons et ne peut en rien être remplacée par la voie oxydative

Le métabolisme des tumeurs endocrines riches en mitochondries, les paragangliomes et les phéochromocytomes familiaux (pgIF) ainsi que les oncocytomes thyroïdiens, est peu connu. Pourtant, il est la cible des outils d'imagerie diagnostique ainsi que des agents thérapeutiques. Le métabolisme mitochondrial des ces tumeurs présente un défaut de synthèse d'ATP au niveau des complexes de la chaine respiratoire. Ce défaut est le plus souvent associé à une prolifération importante des mitochondries. La mesure de l'activité des complexes de la chaine respiratoire des paragangliomes montre une grande différence avec celle des oncocytomes thyroïdiens : l'activité des complexes I et II est significativement abaissée dans les pgIF (comparés aux paragangliomes sporadiques) alors que l'activité des tous les complexes est significativement augmentée dans le cas des oncocytomes (comparés aux tissus sains appariés). Le niveau de production d'ATP mitochondrial est donc différent dans ces deux types de tumeurs.

Dans les pgIF il existe un mécanisme pseudo-hypoxique consécutif à l'inactivation du complexe II de la chaine respiratoire, conduisant à l'induction de l'angiogenèse et de la tumorigenèse. L'accumulation des mitochondries serait un phénomène secondaire au déficit de fonction énergétique. Le métabolisme des tumeurs pgIF est essentiellement glycolytique car les mitochondries accumulées sont peu efficaces en termes de conversion énergétique mais assurent les autres métabolismes nécessaires à la vie de la cellule (biosynthèse et dégradation des molécules, apoptose). Contrairement aux pgIF, dans le cas des oncocytomes, le signal pseudo-hypoxique est associé à une fonction énergétique mitochondriale efficace et à un statut redox particulier.

2.5 Technologie à haut débit pour la recherche de marqueurs

De nombreuses techniques permettent de mesurer l'expression d'un gène. Elles sont basées sur la propriété d'hybridation spécifique des acides nucléiques en simple brin à leur brin complémentaire. On peut ainsi citer entre autres le northern blot (détection spécifique d'un ARN par une sonde marquée) (Alwine *et al*, 1977) ; le differential display (identification des transcrits différentiels entre deux populations d'ARNm) (Liang & Pardee, 1992) ; l'analyse en série d'expression des gènes (Serial Analysis of Gene Expression, SAGE, mesure de la quantité de tous les ARNm d'une population) (Velculescu *et al*., 1995). La technologie des réseaux (ou

matrices) ordonnées d'ADN (DNA Arrays) développée au cours des années 90 s'est rapidement imposée pour étudier l'expression des gènes à grande échelle. En effet, il est possible avec cette technique d'analyser en parallèle plusieurs échantillons et de comparer l'expression de nombreux gènes simultanément dans ces différentes conditions (Schena *et al*., 1995; Lockhart *et al*., 1996; Schena *et al*., 1996; Duggan *et al*., 1999). Ainsi, l'étude de l'expression des gènes ne s'applique plus à quelques transcrits mais à l'ensemble des transcrits d'une cellule à un instant donné, dans une condition donnée (transcriptome).

2.5.1 Les biopuces ou puces à ADN

La technologie des puces à ADN, ou biopuces (microarrays), consiste en un support solide (lame de verre ou membrane de nylon) sur lequel des fragments d'ADN sont déposés à l'aide d'une micropipette robotisée (Schena *et al*., 1995). Chaque fragment d'ADN est représenté par un point (spot) sur le support (ou puce). Ils servent de sondes pour fixer spécifiquement les fragments de gènes complémentaires (cibles), présents dans les échantillons biologiques à tester. Ce phénomène d'hybridation est observable par des techniques optiques sous éclairage fluorescent ou par détection de radioactivité. Il est alors nécessaire de marquer l'échantillon par traceurs fluorescents ou radioactifs. La quantification des signaux et l'identification des fragments de gènes hybridés sont possible grâce à des logiciels informatiques spécialisés.

On peut distinguer 2 technologies pour les biopuces : les biopuces avec un dépôt direct de molécules d'ADN sur leur support et les puces à oligonucléotides avec la synthèse in situ des sondes oligonucléotidiques sur la surface solide. Dans le premier cas, les dépôts (sondes) sont des clones d'ADNc ou des produits de PCR fixés à haute densité le support solide. Elles sont déposées à une densité de plus de 1 000 sondes/cm2 par un robot sur du nylon ou des lames de verre au préalablement traitées chimiquement. Les sondes sont généralement des ADN double brin d'une longueur de 200 à 2000 bp amplifiés par la technique de PCR. Les cibles utilisées sont réalisées par transcription inverse, à partir d'ARN total ou messager. Elles sont marquées grâce à de la radioactivité (P33) ou à des fluorochromes (Cy3 et Cy5). Les technologies mono canal peuvent utiliser la radioactivité ou les fluorochromes pour marquer les cibles. Les technologies à double canal utilisent deux fluorochromes différents (Cy3 et Cy5) pour hybrider simultanément 2 cibles sur une même sonde. Les signaux d'hybridation sont analysés grâce à un lecteur capable de mesurer la radioactivité ou de discriminer les 2 fluorochromes et de générer des images représentant l'intensité du signal. Dans ce cas, les signaux de chaque fluorochrome peuvent être remplacés par une couleur différente (rouge et vert) dont la superposition renvoie

une couleur représentative du différentiel mesuré. Par exemple, un signal rouge sera considéré comme la surexpression d'un gène alors que le vert sera une sous-expression et le jaune une iso-expression.

Les puces à oligonucléotides dérivent à l'origine d'un projet de séquençage par hybridation. Les sondes sont des oligonucléotides synthétisés in situ par photolithographie. Cette technique permet d'atteindre une grande densité : des dizaines de milliers de gènes mesurés sur une surface d'environ 1 cm^2. Chaque puce sert à mesurer l'abondance relative de chacun des ARNm présent dans l'échantillon biologique étudié.

La technologie des biopuces est utilisée pour d'autres types de mesures. Par exemple, des biopuces peuvent mesurer le niveau d'expression de micro ARNs. Il s'agit d'une classe d'ARNs endogènes non codants, mais fonctionnels de 19-23 nucléotides dérivés d'un précurseur de plus grande taille (Bartel 2004). Ces biopuces ciblent simultanément quelques centaines de séquences de micro ARNs matures ou de leurs précurseurs. Les biopuces classiques (chip) peuvent être associées à des méthodes d'immuno-précipitation pour identifier les interactions d'un facteur de transcription avec la chromatine (ChIP). Cette technologie appelée « ChIP on chip » est une approche à l'échelle génomique qui cible les gènes interagissant directement avec le facteur de transcription étudié. Les biopuces classiques ne peuvent pas distinguer les régulations directes et indirectes (pour revue : Wu J *et al.*, 2006).

2.6 Analyse de données transcriptomiques à haut débit

Les expériences sur biopuces produisent de grandes quantités de données brutes. Par exemple, Une étude de 166 échantillons biologiques pour 8862 gènes mesure plus de 1,4 millions de niveaux d'expression (Fontaine *et al, sous presse*). L'information biologique potentiellement contenue dans ces données sera exploitable à plusieurs conditions. Le design expérimental doit être adapté à la question biologique posée ainsi qu'à la gestion des sources systématiques de variations. Des méthodes de normalisation adaptées à la technologie et aux données doivent rendre comparables les mesures des différentes puces. Des outils statistiques puissants et cohérents permettront au final de répondre à la question posée. Les principales applications des biopuces sont la classification, la découverte de classe, la recherche de marqueurs et de signatures différentielles, la prédiction automatique de diagnostic ou de pronostic, la corrélation des données d'expression avec les données de survie.

Dans le cas des tumeurs de la thyroïde où il existe essentiellement un problème de classification, les mesures transcriptomiques à haut débit peuvent apporter des indications complémentaires pour améliorer la sensibilité et la spécificité d'un diagnostic. En considérant l'évolution lente et le bon pronostic de ces tumeurs différenciées, il n'est pas envisagé de corréler les données d'expression à celles de survie.

2.6.1 Design et sources de variations

(Kreil et Russel 2005; Yang et Speed 2002; Churchill 2002)

Les expérimentations sur puces à ADN doivent être organisées avec attention pour que l'analyse des données et l'interprétation des résultats soient aussi simples et puissantes que possible. Pour cela, il faut prendre en compte le but de l'expérience et les contraintes matérielles. Ces contraintes se manifestent par des variations systématiques lors de la manipulation du matériel biologique et technique. Elles vont modifier ou parasiter les signaux et donc empêcher la détection de l'effet recherché. Un bon design expérimental associé à des techniques de normalisation permet de minimiser ces variations systématiques dans la mesure des niveaux d'expression des gènes. Les différences biologiques deviennent alors détectables, et les niveaux d'expressions comparables d'une expérience à une autre.

2.6.1.1 Sources de variations dans les expérimentations de puces à ADN

Les variations biologiques sont intrinsèques à tous les organismes. Elles peuvent être influencées par des facteurs génétiques et environnementaux, ainsi que par le traitement individuel ou collectif (« pooling ») des échantillons. Ces variations sont principalement celles recherchées dans les études biologiques avec comme exemple la caractérisation des niveaux d'expression des gènes de plusieurs tissus. Les variations techniques sont introduites durant l'extraction, le marquage et l'hybridation des échantillons. Des imprécisions de mesures sont associées à la lecture des signaux qui peuvent être influencées par d'autres facteurs comme la poussière ou la température ambiante. Un design habituel consiste à répliquer systématiquement les mesures pour compenser les variations techniques.

2.6.1.2 Stratégies de réplication

Les réplications techniques demandent plusieurs hybridations d'ARNm cibles provenant de la même extraction. Cela permet de diminuer les variations des mesures en étudiant la moyenne. Des mêmes points de mesure dupliqués sur une même puce présentent en général une très

bonne corrélation (r>95%) alors que celle-ci est plus faible si les mesures sont répétées sur d'autres puces (r>60 à 80%) ou dans d'autres laboratoires (r>30%). Des mesures dupliquées sur le même type de puce partagent de mêmes variations (caractéristiques d'extraction répétables, hybridation, conditions de scanning). Elles ne sont donc pas indépendantes. La proximité physique des spots sur le support solide amplifie cette dépendance puisque les effets locaux sont identiques. Cela explique les fortes corrélations observées et indique un intérêt assez faible pour cette stratégie de réplication. L'utilisation d'autres supports ou d'autres matériels et locaux pour dupliquer les mesures assure une indépendance plus marquée, ce qui rend cette stratégie plus intéressante que la précédente. Cependant, les faibles corrélations observées entre les mesures dupliquées mettent en doute la reproductibilité de la technologie. En fait, il faut attendre l'étape nécessaire de normalisation pour pouvoir comparer les données.

Les réplications biologiques consistent en l'hybridation des ARNm de différentes extractions ou de différents individus (avec une plus grande variabilité). Cela permet d'étudier les variations interindividuelles en offrant des critères d'indépendance maximaux entre les mesures. Elles sont donc à préférer puisqu'elles permettent la généralisation des résultats expérimentaux lorsque beaucoup d'individus sont utilisés, alors que cela est impossible avec 1 seul individu même s'il existe plusieurs réplications. Les réplications techniques augmentent la précision et le contrôle de qualité. Elles sont utiles pour mettre au point une plate-forme technologique mais moins utiles pour les expériences biologiques sur des groupes d'individus.

2.6.1.3 Randomisation

Une façon supplémentaire d'éviter les variations systématiques consiste à randomiser le traitement des échantillons. D'un point de vue biologique, il faut choisir des individus aléatoirement dans une population pour leur appliquer un traitement. Un mauvais exemple serait d'appliquer le traitement dont on veut observer les effets sur un groupe d'individus de même sexe. Dans ce cas, les différences associées au sexe, comme l'expression différentielle des récepteurs aux androgènes, pourront être détectées et non discernées de l'effet initialement observé. D'un point de vue technique on peut jouer sur le marquage fluorescent Cy5 et Cy3 des échantillons, sur les échantillons associés aux campagnes de spottage, ainsi que sur le personnel technique qui effectuera les manipulations.

2.6.1.4 Taille des groupes

(Hwang et al., 2002 ; Yang et Speed, 2002; Tibshirani, 2006; Yang et al., 2003; Pavlidis et al., 2003)

Pourquoi la taille des groupes peut-elle poser des problèmes ?

Lorsqu'une expérience sur biopuces est envisagée elle doit être construite de façon à pouvoir répondre à une ou plusieurs questions biologiques. Un objectif courant des études transcriptomiques est la recherche de signatures différentielles entre plusieurs groupes d'échantillons biologiques. Cela permet par exemple d'observer l'effet d'un traitement ou de caractériser un phénotype. La recherche des marqueurs d'une maladie s'inscrit dans cette description. Des tests statistiques sont utilisés pour déterminer les gènes différentiels présents sur les puces. Ces tests demandent le respect de certaines conditions. Le nombre d'échantillons par groupes est un de ces paramètres essentiels. D'un coté, il ne doit pas être trop petit sinon aucune conclusion ne pourra être faite à partir des données. D'un autre coté, il ne pourra pas être trop grand pour des raisons de coût financier et temporel. La disponibilité de certains tissus peut aussi limiter le design des expériences.

Comment estimer une taille minimale ?

Avant toute expérience il est intéressant de connaître le nombre d'échantillons biologiques nécessaires à intégrer dans chaque groupe pour pouvoir détecter les effets recherchés. Quelques études se sont intéressées à déterminer la taille minimale ou optimale des groupes. Pour estimer cette taille, il faut faire appel à la logique de test (Table 2). En partant d'une hypothèse nulle H0 qu'il n'y a pas de différences entre les groupes, le test nous permettra de prendre une décision en rejetant ou en acceptant H0 pour un risque donné (en général 5%). En effectuant un test par gène on peut sélectionner tous ceux qui indiquent une différence significative en espérant se tromper le moins possible (reflet de la sécurité du test) tout en détectant le plus possible de différences (reflet de la puissance du test). Une analyse de puissance permet d'étudier ce compromis. Cependant, le calcul de puissance est difficile car il nécessite la connaissance *a priori* de certains paramètres (Pan *et al.*2001 ; Yang et Speed, 2002). Il faut connaître la variance pour tous les gènes, alors qu'elle peut différer grandement, ainsi que la magnitude de l'effet à détecter alors que cela est impossible à définir. Pourtant, plusieurs études ont utilisé des variantes de cette analyse pour estimer la taille minimale ou optimale des groupes à comparer.

Table 2 : logique de test

Hypothèse H0 **pas de différences entre les moyennes**		***Décision***	
		H0 vraie ***Accepter l'hypothèse nulle***	***H0 fausse*** ***Rejeter l'hypothèse nulle***
En réalité	*H0 vraie :* *pas de différences*	Probabilité = 1 – β (puissance du test)	risque α d'avoir un faux positif (Erreur de type I)
	H0 fausse : *il existe une différence*	risque β d'avoir un faux négatif (Erreur de type II)	Probabilité = 1 - α (seuil de confiance)

Dans l'étude de Hwang *et al.* (2002), une analyse simple de puissance est proposée pour déterminer la taille des groupes. Un jeu de données de départ est réduit de certains échantillons et les distributions de H0 et H1 sont obtenues à partir de données permutées. Si la puissance et la sécurité des tests sont supérieures à un seuil fixé (ex. 95%) alors le nombre d'échantillons est valable. La comparaison de deux groupes (Leucémie aiguë lymphoblastique, Leucémie aiguë myéloïde) montre qu'au moins 8 échantillons doivent être utilisés par groupe et la comparaison de 3 groupes (Leucémie aiguë lymphoblastique B, Leucémie aiguë lymphoblastique, Leucémie aiguë myéloïde) en montre au moins 7. Cette méthode se heurte à l'assomption d'un échantillonnage homogène qui doit être représentatif de la population alors qu'un faible nombre d'échantillons implique de s'en écarter. De plus le nombre proposé d'échantillons par groupe est identique alors que cela n'est pas assuré dans une expérimentation.

L'étude de Tibshirani (2006) propose de s'affranchir des hypothèses d'égalité des variances et de d'indépendance des profils de gènes en utilisant des tests permutés. Ils permettent de conserver les variances et les dépendances présentent dans le jeu de données de départ. L'objectif est de minimiser le taux de fausses découvertes et de maximiser la puissance des tests. L'étude n'étant réalisée qu'à partir de données simulées et uniquement sur 2 groupes, elle ne peut pas être généralisée aux cas réels et à plus de groupes (>2). De plus, les tailles de groupes

recommandées de 60 à 100 échantillons sont quasiment irréalistes dans le contexte actuel et reflètent l'écart du modèle à la réalité, ou une trop grande précision attendue.

Une étude plus réaliste a été proposée par Pavlidis *et al*. (2003). Les vrais gènes différentiels d'un jeu de données réelles ne sont pas connus. De plus, la variance estimée est dépendante du modèle et de la qualité des ses paramètres qui diffèrent pour chaque gènes. On ne peut donc pas connaître la puissance des tests et une analyse stricte serait complexe à mettre en place. D'ailleurs les études précédentes (Hwang et *al.*, 2002 ; Pan et *al.*, 2002) utilisent une variante plus simple de ce genre d'analyse. Cette étude estime la puissance apparente (nombre de gènes détectés) et la stabilité des détections (listes de gènes et rangs) en fonction du seuil choisi. Il est montré pour plusieurs jeux de données réelles de 2 groupes qu'il n'y a pas de stabilité avec moins de 5 échantillons par groupe. La stabilité optimale est obtenue entre 10 et 15 échantillons.

Adapter les méthodes d'analyses

L'utilisation d'un jeu de données simulées permet de connaître les paramètres qui font défaut pour le calcul de la puissance. Cependant, ces modèles ne peuvent pas être représentatifs de données réelles. L'utilisation d'un jeu de données de même nature est un meilleur choix mais il n'est pas forcément disponible ou reste difficile à évaluer. L'idéal étant un jeu de données utilisant la même technologie et les mêmes types d'échantillons. Le problème reste entier lorsqu'un jeu de données inédit est utilisé avec des groupes de petite taille.

De plus, la disponibilité des échantillons varie en fonction de leur type avec par exemple des pathologies rares et d'autres abondantes. Ceci mène à la réalisation d'expériences comprenant des groupes de tailles non équilibrées. Les seuils utilisés dans les statistiques de tests dépendent des degrés de liberté calculés à partir des tailles de groupes. Pour le test de Student les degrés de liberté considérés sont égaux à la somme des tailles des 2 groupes moins 2. Deux groupes de 10 échantillons donne 18 degrés de liberté mais il en va de même pour un groupe de 15 et un groupe de 5 échantillons. Ce dernier effet reste problématique puisqu'il modifie les propriétés des groupes (variance, taille de l'effet détectable). Il peut avoir un fort impact lors de la sélection de gènes différentiels en fonction des méthodes statistiques employées (Yang *et al.*, 2006).

Toute étude comportant des biais évidents dans son design devrait les prendre en considération. Ils doivent être détectés et quantifiés par analyse simple de puissance ou optimisation de

paramètres qualitatifs du jeu de données ou d'un sous-ensemble sélectionné. Dans ce cas, les méthodes d'analyses ne peuvent être décidées *a priori* pour pouvoir s'adapter aux données.

2.6.2 Traitement des données

Le traitement et la normalisation des données a pour but de rendre les mesures comparables sur un même type de puce ou entre puces d'origines différentes. Les méthodes utilisées doivent être adaptées à la plate forme utilisée ainsi qu'aux données analysées (Kreil et Russel, 2005). Il existe des traitements à appliquer pour prendre en compte l'effet du bruit de fond. Cet effet serait maximal sur les signaux de faibles intensités mais négligeable sur les signaux de fortes intensités. Il a été montré que la modification des mesures par une quantité proportionnelle au bruit de fond augmente la variance observée pour les faibles signaux (Qin *et al.*, 2004). Cela a pour effet de limiter la capacité à déterminer les gènes différentiels. Le bruit de fond pourra par contre servir à filtrer les mesures trop faibles pour les considérer comme des valeurs manquantes dans le but d'éliminer les gènes trop faiblement exprimés de l'analyse. Dans le cas des puces à ADNc sur support nylon, la normalisation par la médiane est utilisée. Elle permet de relativiser les mesures à la médiane d'une puce, représentative d'une iso expression. La méthode Lowess (« locally weighted least squares régression ») de lissage non linéaire permet de corriger un biais de mesure qui entraîne une dépendance selon l'intensité de la mesure.

2.6.3 Analyse des données

(Mount and Pandey, 2005; Kim et al., 2004; Armstrong et van de Wiel, 2004; Krajewski et Bocianowski, 2002)

Les données d'expression se présentent dans une grande matrice dont les lignes représentent les gènes et les colonnes les échantillons. Il y a des milliers de gènes pour des dizaines d'échantillons. Les données sont complexes et multivariées, elles incluent des dépendances et des corrélations. Plusieurs approchent sont possibles pour analyser ces données. L'approche univariée s'intéresse individuellement aux gènes pour détecter par exemple ceux au potentiel discriminatif le plus fort, pouvant être de potentiels marqueurs. L'approche multivariée s'intéresse à tous les gènes pour extraire des sous ensembles informatifs. Elle peut résumer l'information en 2 ou 3 dimensions représentatives de la plus grande variabilité dans les données.

2.6.3.1 Méthodes univariées

Les tests statistiques

Pour détecter une différence de moyenne entre des groupes d'échantillons, nous pouvons utiliser des tests statistiques paramétriques (Student, Anova) ou non paramétriques (Wilcoxon, Kruskal-Wallis). Certaines versions modifiées de ces tests ont été développées pour les adapter aux contraintes spécifiques des puces à ADN (t-test et Anova modifiés). Pour pouvoir utiliser un test paramétrique les données doivent respecter certaines conditions. Principalement, elles doivent suivre une loi de distribution normale, les gènes doivent être indépendants et les groupes à comparer doivent avoir des variances égales. Elles ne sont pas forcement respectées avec les données d'expression. Il est difficile de fixer une limite au delà de laquelle une méthode serait préférable à une autre, sauf dans le cas des très petits groupes où les tests non paramétriques ont l'avantage. Les tests paramétriques montrent une certaine robustesse en ce qui concerne le non respect de la loi de distribution mais il est souvent conseillé de se tourner vers les tests non paramétriques. Ces tests non paramétriques, libérés des problèmes de loi de distribution et de variance, ont aussi des contraintes, dont celle de l'indépendance des gènes. Pour prendre un autre exemple, le test de Kruskal-Wallis non paramétrique de comparaison des moyennes de plus de 2 groupes nécessite l'utilisation d'au moins 5 échantillons par groupe.

Dans le cas où aucun test ne voit ses conditions satisfaites, lequel se montrera le plus robuste ? Une comparaison des différents tests peut-être envisagée mais elle n'est pas évidente. Les mêmes problématiques de tests pour les puces à ADN abordées ci-dessus sont à prendre en considération.

Les tests basés sur des permutations ne font pas d'hypothèse sur la distribution des données, ni sur la taille des groupes. Le profil d'expression d'un gène sert de modèle pour représenter toutes les valeurs d'expression possibles pour ce gène. Un échantillon virtuel peut être créé en sélectionnant aléatoirement pour chaque gène une valeur de leur profil. Pour une matrice de données d'expression, on peut permuter les labels des colonnes pour chaque ligne et ainsi obtenir une nouvelle matrice simulée ayant gardé des groupes de même taille. Il est courant de générer au moins 100 à 1000 matrices aléatoires pour la procédure. En calculant les statistiques de tests pour toutes les matrices aléatoires, et en les comparant avec celles de la matrice originale, on peut déduire des p-values. Pour chaque gène, combien a-t-on trouvé de statistiques de tests plus élevées dans les jeux aléatoires ? Cette fréquence sera considérée comme la p-value. Soit un gène A ayant obtenu une statistique égale à 2 dans le jeu de données original. Si

l'on trouve 30 jeux aléatoires sur 1000 où la statistique associée à ce gène est supérieure à 2, alors sa P-value sera 30 / 1000 = 0.03. Il faut générer assez de matrices aléatoires pour que les p-values déduites soient stables. Les méthodes basées sur les permutations sont très robustes pour le jeu de données considéré. Il faut donc s'assurer que le jeu de données soit représentatif des populations étudiées.

Multiplicité des tests

Le fait de pouvoir mesurer simultanément l'expression de milliers de gènes doit être pris en compte dans les analyses univariées. En simulant un jeu aléatoire de 1 000 gènes, de distribution homogène ou normale, pour 2 groupes de 5 individus, environ 50 gènes seront automatiquement considérés significativement différentiels à 5% par le test de Student. Cette erreur est due à la multiplicité des tests, ce qui rend obligatoire des procédures de correction. La plus connue et la plus simple est la correction de Bonferroni. Elle contrôle l'erreur de type 1 : la probabilité qu'au moins un gène soit déclaré significatif alors qu'il ne l'est pas. Cela se traduit dans l'exemple précédant par le changement du seuil de significativité de 5% à 0,05 / 1000 = 5E-05. Ce choix est très conservateur et réduit grandement la puissance de l'analyse. Des variantes moins conservatrices existent pour le contrôle de l'erreur de type 1 (Holm, 1979 ; Westfall et Young, 1993).

Un contrôle plus adapté aux données de puces est le taux de fausses découvertes (Gordon *et al*., 2007) introduit par Benjamini et Hochberg (1995). Il permet de contrôler le nombre de faux positifs dans la sélection finale. Etant donné que l'on peut s'attendre à la sélection de plusieurs dizaines ou centaines de gènes, intégrer quelques faux positifs sera peu nuisible et acceptable par un biologiste. Un faible taux de faux positifs a un effet négligeable dans les analyses d'enrichissement des groupes de gènes (recherche de motifs ou ontologies). D'autres versions existent pour tenir compte des corrélations entre les gènes (Benjamini et Yekutieli 2001).

2.6.3.2 Approche multivariée

Bien que les méthodes de base considèrent les gènes comme des traits individuels, ce qui est conforme aux règles générales de la conception expérimentale, plusieurs approches ont été développées pour voir, dans le jeu de données de ratios d'expression, les gènes comme des cas et les puces comme des variables. L'algèbre correspondante à cette approche a été décrite par Kuruvilla *et al.* (2002). La plupart des méthodes bien connues basées sur la décomposition en valeurs singulières ont été employés : analyse en composantes principales (Wall *et al.* 2001 ;

Yeung et Ruzzo, 2001), analyse de correspondance (Fellenberg *et al*. 2001) et biplots (Chapman *et al*., 2002) Une méthode basée sur la distance de Mahalanobis pour détecter des gènes différentiels a été décrite par Chilingaryan *et al.* (2002).

Les algorithmes de regroupement existants ont été appliqués aux données d'expression de différentes manières. Par exemple, Eisen *et al.* (1998) a employé le regroupement hiérarchique des gènes basé sur la distance de Pearson. Getz *et al.* (2000) ont étudiés différents aspects des méthodes de regroupement bidirectionnelles (c'est-à-dire, un regroupement appliqué à la fois aux gènes et aux échantillons) selon plusieurs paramètres et plusieurs mesures de distance. Les méthodes de regroupement basées sur des modèles (modèles de mélanges) ont été considérées par Yeung *et al.* (2001), Ghosh et Chinnaiyan (2002) et Mclachlan *et al.* (2002). Quelques propositions intéressantes au sujet des méthodes de regroupement ont été faites par Hastie *et al.* (2000, 2001). Ils ont considéré la situation dans de la laquelle des observations supplémentaires sont faites pour certaines conditions. Cela donne la possibilité de lier les groupes obtenus par l'algorithme à des données externes de type différent, quantitatif ou qualitatif.

2.6.3.3 Sélection de gènes différentiels

De nombreuses méthodes de sélection ont été employées pour détecter les gènes différentiels sur puces à ADN. Différentes méthodes appliquées à un même jeu de données produisent des listes de gènes étonnement différentes (Pan 2002). Etant donné le manque de jeux de données réelles contenant un nombre suffisant de vrais et faux positifs connus, peu d'études existent pour comparer ces méthodes. Plusieurs études récentes se sont intéressées au problème avec des approches différentes.

Dans l'étude de Jeffery *et al.* (2006), des méthodes de sélection sont comparées sur 9 jeux binaires de données publiques. Ces méthodes sont premièrement comparées par rapport aux listes de gènes sélectionnés puis par leur pouvoir prédictif en suivant 4 algorithmes. La méthode ROC se montre la meilleure à condition d'avoir au moins 15 échantillons par groupes. Cette dernière condition rend ce test moins intéressant puisqu'il est difficile de trouver des études ayant regroupé autant d'échantillons pour chaque groupe. Les tests de type t sont sensibles au bruit dans les données ainsi qu'aux faibles tailles de groupes, ils ont de faibles performances. La méthode SAM se révèle moyenne ou faible surtout avec de petits groupes. Lorsque peu d'échantillons sont présents, les méthodes qui ne modélisent pas la variance sont meilleures (rangs, « fold change »). La méthode de comparaison de cette étude assume que l'association des meilleures sélections individuelles de gènes fera le meilleur prédicteur.

Comme le comportement des algorithmes de prédiction est parfois instable selon le seuil choisi dans la sélection des gènes (Yukinawa *et al.*, 2006), il serait préférable de dissocier la sélection de marqueurs de la sélection de prédicteurs. Pour une méthode donnée ce ne sont pas forcément les meilleurs gènes différentiels qui feront les meilleurs prédicteurs. Ajoutée à une dépendance au jeu de données et au comportement imprévisible des algorithmes, il est difficile de comparer directement les méthodes de sélection. Une dernière remarque sur cette analyse : elle ne suit pas la procédure de test mais prend en compte uniquement les valeurs des statistiques. Cela prive la sélection d'un contrôle par seuil de significativité qui pourrait filtrer des gènes incohérents ayant pourtant une statistique élevée.

Dans l'étude de Jafari et Azuaje (2006), une revue de la littérature nous montre les méthodes les plus utilisées. L'ANOVA est utilisée dans un tiers des 141 articles étudiés, le test t dans 14.89 % des cas, l'analyse en composantes principales dans 7.96%, les tests non paramétriques dans 7.80 % des cas, les autres méthodes de sélection de gènes sont marginales. La théorie soutenue par quelques études montre que les tests t et l'ANOVA sont robustes aux digressions de leurs assomptions sauf pour les très petits groupes et les groupes de tailles inégales (Carlin et Doyle, 2001 ; Seldrup, 1997 ; Moher *et al.*, 1994, Williams *et al.*, 1997). Il est souvent conseillé d'utiliser les tests non paramétriques pour résoudre ces problèmes mais ils ont aussi leurs limites. Par exemple, le test de Kruskal-Wallis de comparaisons multiples demande au moins 5 échantillons par groupes puisque une loi du Chi2 est utilisée pour déduire des probabilités. L'utilisation plus régulière des tests paramétriques, alors que les données ne respectent pas le plus souvent leurs prérequis, pourrait s'expliquer par le fait que la théorie générale ne s'applique pas toujours au cas particuliers. Etant donné les sources nombreuses de variations systématiques, les données de puces pourraient toutes être considérées comme particulières.

Dans l'étude de Qin *et al.* (2004) la comparaison des méthodes s'effectue sur les valeurs de statistiques pour des échantillons identiques comportant 6 gènes « spike-in » sur 17 000. Le test t apparait faible et les meilleures performances sont atteintes par la moyenne, la médiane et le fold change. Ce dernier est reconnu par tous comme étant historiquement la première méthode utilisée, au vue de sa simplicité, mais la moins efficace et la moins stable (Jeffery *et al.*, 2006). La détection de 6 gènes seulement sur 17 000 pose aussi des problèmes. Les résultats de cette étude ne peuvent donc pas se généraliser.

Dans l'étude de Chen *et al.* (2005) les méthodes appliquées à plus de 2 classes sont comparées. 5 jeux publics sont utilisés pour comparer 6 méthodes. Les 50, 100 et 500 meilleurs gènes de chacune d'elles sont comparés selon leur potentiel prédictif pour 5 algorithmes. Là aussi seules

les statistiques de tests sont utilisées et pas la procédure de test complète. La comparaison est faite de façon indirecte par des algorithmes instables selon les seuils et les listes de gènes.

2.6.3.4 Classification multiple

La méthodologie pour construire une règle de prédiction à partir de données de puces à ADN implique plusieurs étapes. Elles correspondent à la sélection d'un jeu de gènes de qualité, au choix d'un algorithme, à la sélection des gènes rapporteurs dans le profil final, ainsi qu'à une règle de validation indépendante. Plusieurs variations autour de cette procédure basique ont été proposées, impliquant régulièrement différentes combinaisons de stratégies de sélection de rapporteurs et de prédicteurs. Dudoit *et al*. (2002a) ont comparés plusieurs prédicteurs et en ont conclu que des prédicteurs simples, comme l'analyse diagonale linéaire discriminante, atteignent de bonnes performances sur les données d'expression par rapport à des méthodes plus compliquées. Li *et al*. (2003) ont comparé des prédicteurs multi-classes en s'intéressant aux problèmes relatifs aux combinaisons de prédicteurs binaires dans un but de classification multiple. Inza *et al*. (2004) ont comparés le filtrage ou le « wrapping » pour la sélection de gènes rapporteurs sur 2 jeux de données.

Les algorithmes utilisés dans la classification sont de complexité variable. Parmi les plus simples nous pouvons citer le centroid le plus proche (Van't Veer *et al.*, 2002), l'analyse discriminante linéaire diagonale (Dudoit *et al.*, 2002a), les k meilleurs voisins (Barnard, 1935). Les plus complexes incluent les machines à vecteur de support (Vapnik 1999), et les méthodes Gaussiennes et Bayesiennes (Domingo et Pazzani. 1997 ; Dudoit *et al.*, 2002b).

Les méthodes de validation indépendantes doivent faire appel à un jeu de données différent que celui qui a servi à choisir le classificateur (gènes et algorithme). Une seconde expérimentation indépendante de la première peut être envisagée. Il est aussi possible d'utiliser une partie des données pour déterminer le classificateur et une autre partie pour le valider. Les méthodes de validation croisées les plus utilisées sont le « Leave one out » et le « K-fold ». Le « Leave one out » exclut un individu avant de construire le classificateur, il sert ensuite à valider les performances. Chaque individu est exclu une fois et la moyenne des performances est retournée comme validation croisée. Le « K-fold » procède d'une manière similaire mais au lieu d'exclure 1 seul individu, il en exclut une fraction K du nombre total. Ainsi, le 10-fold divise le jeu de données en 10 fractions, utilise l'une d'elle pour la validation des classificateurs construits par les 9 autres fractions. Ambroise et McLachlan (2002) ont montré que la méthode « Leave one out » était non biaisé, même s'il reste variable (Efron, 1983). Le 10-fold est plus biaisé mais

moins variable. Dans un cas comme dans l'autre il est possible de sur- ou sous-estimer les performances. En effet, si les gènes rapporteurs ne sont sélectionnés qu'une fois en début de procédure alors la validation ne sera pas indépendante, les échantillons utilisés pour la validation ayant servi à déterminer les gènes (Ambroise et McLachlan, 2002 ; Simon *et al.,* 2003)

Bien que le cadre théorique puisse définir des méthodes *a priori* plus robustes, l'application aux données réelles peut favoriser certaines méthodes à d'autres. Dudoit *et al*. (2002) montrent que les algorithmes simples sont très performants. Wessels *et al*. (2005) montrent que la sélection de gènes est meilleure avec des méthodes simples (filtrage, shrunken centroids) plutôt qu'avec la méthode des moindres carrés partiels. Comme pour la sélection de gènes différentiels, la structure des données (variabilité, amplitude, taille de l'échantillon) a un grand rôle sur l'efficacité des algorithmes.

2.6.3.5 Recherche de sites de fixation de facteurs de transcription

(Xie *et al.*, 2005 ; Blanchette *et al.*, 2006)

La transcription des gènes est dépendante des interactions entre des facteurs de transcription qui se fixent sur des éléments cis-régulateurs de l'ADN, de la présence de cofacteurs et surtout de l'influence de la structure de la chromatine (GuhaThakurta 2006 ; pour revue Wasserman et Sandelin, 2004). La compréhension des mécanismes de contrôle de la régulation pourrait aider dans l'interprétation des données complexes générées par les technologies transcriptionnelles à haut débit. Les méthodes informatiques d'analyse des séquences régulatrices s'intéressent principalement à l'initiation de la transcription. D'autres mécanismes de contrôle de l'expression des gènes ne doivent pas être négligés car la régulation des gènes peut s'effectuer à toutes les étapes de transformation d'un transcrit en une protéine fonctionnelle. Néanmoins, la transcription sélective des gènes par l'ARN polymérase-II sous des conditions spécifiques est essentielle pour beaucoup, sinon presque tous les gènes qui seront traduits.

La première étape à considérer est la détermination des régions génomiques qui contiennent des éléments régulateurs. Les outils existants utilisent la conservation entre gènes orthologues, la composition nucléotidique et les données transcriptomiques. Ces régions se situent généralement près des sites d'initiation de la transcription. La position précise de ces sites est pourtant difficile à établir. De plus, les gènes peuvent avoir des sites alternatifs d'initiation de la transcription. La caractéristique dominante des régions promotrices du génome humain est

l'abondance des dinucléotides CpG. Dans ces régions, les cytosines de ces dinucléotides ne sont pas méthylées alors qu'elles le sont à 80% ailleurs. La méthylation d'une région promotrice impliquera la non transcription des gènes dépendants, ce qui constitue un mode de régulation épigénétique de la transcription. Les cytosines méthylées sont mutées à un fort taux, ce qui résulte en une réduction de 20% de la fréquence des CpG dans les séquences sans rôle régulateur. Le déséquilibre en CpG est donc utilisé par les méthodes bioinformatiques puisque ces régions contiennent probablement des promoteurs. Seulement 60% des promoteurs de gènes humains sont situés à proximité d'îlots CpG. La découverte des promoteurs sans régions riches en CpG est possible par des méthodes alternatives. Les données transcriptomiques peuvent être utilisées par alignement d'ESTs et de cDNA pour localiser les régions promotrices.

Pour prédire les sites de fixation de facteurs de transcription, il est possible d'utiliser les propriétés de conservation des séquences, les empreintes phylogénétiques. L'assomption suivante est faite : les mutations dans les régions génomiques fonctionnelles accumulent moins de mutations que les régions sans fonction spécifique. Un autre point important est l'hypothèse implicite que la régulation de gènes orthologues est soumise aux mêmes mécanismes. Ceci présente une limite relative à la distance évolutive entre les espèces comparées, les plus proches ayant des mécanismes plus comparables.

A partir d'un jeu de sites de fixation connus, il faut choisir une modélisation des données expérimentales et une méthode de recherche dans les promoteurs adaptées. La construction des modèles est limitée par la faible quantité d'éléments cis-régulateurs connus. Les méthodes les plus utilisées se basent sur une hypothèse qui ne sera pas forcement respectée : chaque facteur se fixe indépendamment sur sa cible. Plusieurs limitations en découlent puisque les interactions combinatoires de multiples facteurs qui se fixent sur plusieurs sites sont essentielles pour la régulation spécifique de la transcription (Palstra *et al.*, 2003). Même si le modèle statistique est confirmé in vitro pour la majorité des sites (Tronche *et al.*1997), il est impossible de distinguer les sites fonctionnels *in vivo* des sites non fonctionnels. La présence d'une activation n'active pas forcément la fonction, ce qui induit de forts taux de faux positifs dans les recherches (1/500 à 1/5000 paires de bases).

Les séquences consensus peuvent être utilisées pour modéliser les sites de fixation. Elles n'intègrent pas les caractéristiques quantitatives des sites contrairement aux matrices poids-position. Ces dernières sont les plus utilisée puisqu'elles sont aussi efficaces que des méthodes plus complexes et plus gourmandes en temps de calcul. Après alignement de plusieurs séquences de sites de fixation connus d'un facteur de transcription, le nombre total

d'observations de chaque nucléotide à chaque position est reporté dans une matrice fréquence-position (le motif). Cette matrice peut être normalisée pour contenir des probabilités d'occurrence des nucléotides à chaque position en s'assurant que le total de chaque colonne soit égal à 1. Pour une position dans une séquence génomique, la probabilité de correspondre au modèle sera le produit des fréquences de chaque site (colonne de la matrice).

Les assomptions suivantes doivent aussi être considérée lors de l'interprétation des résultats : (1) un nucléotide à une position donnée n'a pas d'effet sur la vraisemblance de ses voisins à être observés, (2) les sites de fixations ont des tailles fixes.

2.7 Projet de thèse

Selon la classification internationale des tumeurs (DeLellis, *et al.,* 2004), les tumeurs thyroïdiennes sont classées selon leur origine cellulaire, épithéliale ou médullaire. La malignité des tumeurs est essentiellement basée sur l'importance de l'effraction capsulaire et de l'envahissement vasculaire du tissu tumoral. Le niveau d'invasion au-delà duquel le pronostic devient défavorable n'est toutefois pas clairement défini. L'examen histologique, notamment dans les cas peu invasifs, ne permet pas toujours de classer la tumeur avec certitude, ce qui a une implication directe sur le suivi et le traitement des patients. L'identification de marqueurs moléculaires spécifiques est essentielle pour augmenter la précision d'un diagnostic qui prend en compte la sous classification des lésions de la thyroïde de l'OMS.

L'étude des profils d'expression des gènes par puces à ADN est utilisée depuis plusieurs années pour identifier les marqueurs moléculaires du cancer. Les précédentes études thyroïdiennes se sont principalement portées sur la distinction des cas bénins et malins en comparant les cancers papillaires qui représentent 80% des cancers (Huang *et al.*, 2001), ou les cancers folliculaires à un ensemble de contrôles (Barden *et al.*, 2003). Ces 2 types de cancers ont aussi été comparés dans les mêmes études ou par méta-analyse pour identifier les gènes différentiels (Aldred *et al* ., 2004). Des données d'expression existent pour plusieurs types de tumeurs de la thyroïde mais elles ont été générées par des équipes et des technologies différentes. Leur comparaison est difficile ou impossible pour 3 principales raisons. (1) Certaines études utilisent des comparaisons intra-individuelles alors que d'autres font des comparaisons interindividuelles, (2) les tissus de référence sont parfois le tissu sain avoisinant la tumeur ou bien un ensemble de tissus sains et de lésions bénignes, (3) Les différentes plateformes technologiques de puces à ADN ne sont que partiellement compatibles ou pas du tout (Eszlinger *et al.*, 2007). Considérer simultanément toutes les lésions connues de la thyroïde permettrait d'atteindre une meilleure

spécificité biologique en prenant en compte l'infiltration lymphocytaire et les lésions atypiques. Et surtout, il ne nous semble pas possible de définir les marqueurs spécifiques de chaque type tumoral si nous ne savons définir une signature moléculaire spécifique de chacun des principaux types de tumeurs thyroïdiennes folliculaires différenciées.

Nous avons regroupé dans une même étude 166 échantillons thyroïdiens représentant 11 classes histologiques, soit 90% des lésions folliculaires. L'analyse des signatures moléculaires des tissus permet de retrouver et de compléter en partie la classification histologique des tumeurs (Article 1). Les tumeurs oncocytaires et les tumeurs de malignité incertaine sont mal définies dans la classification anatomopathologique classique. Les études moléculaires permettent d'individualiser et de caractériser tout particulièrement ces variétés tumorales. De plus, l'étude différentielle permet la sélection de marqueurs spécifiques permettant d'optimiser une prédiction automatique du diagnostic de toutes les classes de tissu étudiés (Brevet européen). Nous avons identifié des fonctions dérégulées dans les tumeurs par l'intermédiaire de grands groupes de gènes d'expression corrélée.

Par la suite, nous avons exploré plus précisément les profils des T-UMs qui ont montré une forte corrélation avec les carcinomes (Article 2). Les données d'expression placent la majorité ces tumeurs dans les classes malignes. Avec une analyse complémentaire des mutations caractéristiques de certaines tumeurs, nous pouvons identifier les cas atypiques comme situés à un stade intermédiaire vers les carcinomes papillaires.

Une 3^{e} étude (Article 3) sur l'analyse de l'expression du gène dap3 et de sa protéine (97 lésions appariées au tissu sain avoisinant étudiés en immuno-histochimie) dans les tumeurs thyroïdiennes montre l'apport d'une méta-analyse des données d'expression (à partir de 100 jeux de données publiques) dans la recherche des régulateurs moléculaires potentiels associés aux lésions.

3 Articles et brevet

3.1 Article 1

3.1.1 Introduction

Titre : Gene expression analysis distinguishes 11 classes of follicular thyroid lesions

L'examen histologique ne peut pas toujours précisément classer les tumeurs. Particulièrement, la distinction des FTC et des FTA ont un faible taux de reproductibilité entre spécialistes (Franc *et al.*, 2003). La cytoponction est l'outil le plus précis et sensible pour le diagnostic des pathologies thyroïdiennes. Dans 20% des cytoponctions montrant une prolifération folliculaire, la chirurgie est la seule méthode pour différencier les FTAs, des FTCs et des FVPTCs (Gharib 1997). Par conséquent, pour un meilleur diagnostic, des marqueurs moléculaires pourraient être utile pour discerner les tumeurs bénignes et malignes. Jusqu'à présent, plusieurs marqueurs ont été évalués dans différentes conditions (LGALS3, HBME1, CK19 et aussi TPO), ainsi que des gènes liés à l'initiation de la transformation, comme les réarrangements PAX8/PPARgamma et H-Ras, et les mutations BRAF (Bartolazzi, 2001; de Micco, 1991; Fagin, 2002; Gimm, 2001). Pourtant, aucun n'est actuellement applicable à un diagnostic de routine. Les difficultés rencontrées ont un impact sur le suivi et le traitement des patients.

Des analyses de données de biopuces sur les tumeurs différenciées de la thyroïde ont été menées sur des lésions bénignes et malignes. Elles ont étudié un nombre limité de classes, moins de 6 simultanées, et elles on principalement comparé les signatures des carcinomes folliculaires et papillaires par rapport aux tissus sains ou aux adénomes folliculaires, sans distinction de leurs sous classes. Les signatures observées corrèlent bien avec le diagnostic, elles définissent des profils de progression cellulaire que l'on peut retrouver dans d'autres types de cancers. Ne concernant pas toutes les pathologies de la thyroïde, la spécificité des signatures ne peut être que partielle. De plus, elles ne permettent pas de caractériser les cas atypiques qui n'ont pas été considérés.

Nous avons étudié simultanément les profils d'expression de 11 catégories de lésions folliculaires de la thyroïde. Nous avons inclus 166 échantillons représentant des adénomes macro et micro-folliculaires, des adénomes et carcinomes oncocytaires, des goitres multi-nodulaires, des carcinomes folliculaires et papillaires, des thyroïdites auto-immunes, des

maladies de Basedow ainsi que des contrôles non tumoraux de tissu thyroïdien (tissu sain). Nous avons considéré les cas des T-UM ou oncocytaires séparément des classes précédentes. Ces adénomes ne présentent pas d'invasion vasculaire ou capsulaire. Ils montrent une prolifération prononcée et une histologie remodelée. Ils ne présentent pas les noyaux caractéristiques des carcinomes papillaires.

Les données d'expression de chaque tissu pour 9000 gènes ont été générées par une plate-forme "Transcriptome" (plateforme RIO, laboratoire TAGC, Inserm ERM206, Marseille). Nous avons étudié les modules fonctionnels régulés dans les pathologies par regroupement hiérarchique et analyse ontologique des gènes. Nous avons observé une grande hétérogénéité des tumeurs qui est en relation avec le chevauchement des critères histologiques qui définissent les classes. Les gènes différentiels ont été définis grâce à des tests permutés et un contrôle du taux de fausses découvertes à 5%. L'hétérogénéité des signatures nous a conduits à établir une classification générale des tissus en considérant les signatures moyennes des classes. Les tests de prédiction de classe ont été réalisés par étude comparative de plusieurs algorithmes avec optimisation des seuils de sélection des gènes. Finalement, nous avons effectué des mesures sur 40 tissus par Western Dot Blot pour étudier la corrélation des niveaux d'expression géniques et protéiques.

3.1.2 Article

Title: Gene expression analysis distinguishes 11 classes of follicular thyroid lesions

Authors:

Jean-Fred Fontaine[1-2], Frédérique Savagner[1-3], Délphine Mirebeau[1-3], Mahatsangy Raharijaona [4-5], Stéphane Triau[6], Patrice Rodien[7], Olivier Goëau-Brissonniére[8], Lucie Karayan[9], Brigitte Franc[10], Rémi Houlgatte[4-5], Yves Malthiery[1-3].

1: INSERM, U694, Angers, F-49033 France

2: Université d'Angers, Faculté de Médecine, Angers, F-49033 France

3: CHU Angers, Laboratoire de Biochimie, Angers, F-49033 France

4: INSERM, U533, Nantes, F-44035 France.

5: Université de Nantes, Faculté de Médecine, Institut du Thorax, Nantes, F-44035 France.

6: CHU Angers, Fédération de Pathologie Cellulaire et Tissulaire, Angers, F-49033 France

7: CHU Angers, Département d'Endocrinologie et de Médecine Interne, Angers, F-49033 France

8: Service de Chirurgie Vasculaire, Hôpital A Paré, 92104 Boulogne, France

9: CHU Poitiers, EA 3805, F-86021 France

10: Laboratoire d'Anatomie Pathologique, Hôpital A Paré, 92104 Boulogne, France

Corresponding author:

Frédérique Savagner, Inserm U 694, Laboratoire de Biochimie, CHU, 4 rue Larrey, 49033 Angers, France, tel : +33 241 35 33 14, Fax : +33 241 35 40 17, frsavagner@chu-angers.fr.

Running title: Molecular classification of follicular thyroid tumours

Keywords: Thyroid tumours, differentiated follicular lesions, classification, subclasses

Abstract

The histological distinction between the various types of follicular thyroid tumour is sometimes difficult to make. The identification of specific gene markers would therefore help to implement the thyroid tumour sub-classification proposed by the World Health Organization. Microarray analysis offers a promising approach on condition that all the subclasses are covered, taking into account lymphocyte infiltration and atypical lesions. Our microarray analysis of 166 thyroid tissue samples included 90% of the different types of follicular thyroid pathologies. All the well-defined histological classes were represented, including follicular adenomas and carcinomas, papillary carcinomas, oncocytic adenomas and carcinomas, Graves' disease, and autoimmune thyroiditis. The samples also included some atypical follicular tumours. Despite the heterogeneity of some follicular thyroid tumours, the microarray analysis allowed the identification of accurate classifiers. We are currently testing the panel of selected genes in order to refine the diagnosis and treatment of follicular thyroid tumours.

Introduction

Most thyroid nodules arising from follicular cells develop into hyperplastic nodules, adenomas or malignant tumours (Hedinger *et al*., 1989). The prevalence of carcinomas is fairly low, representing less than 10% of the thyroid nodules. Papillary carcinomas represent 80% of the thyroid carcinomas and nearly half are smaller than a centimetre in diameter. Since the frequency of thyroid nodules is rather high, the detection of a papillary thyroid carcinoma is as difficult as finding the proverbial needle in a haystack. Currently, fine-needle aspiration is considered to be the best procedure for investigating malignancy in thyroid nodule management. However, the cytological examination may miss the malignant zones and lead to delays in prescribing appropriate treatment. Diagnostic difficulties in the so-called "grey zone" are due to the histological overlap between some benign lesions and carcinomas that may be classified as either atypical thyroid adenomas or well-differentiated follicular tumours of uncertain malignant potential (Williams *et al.*, 2000; Lubitz *et al.*, 2006). Furthermore, in the absence of distinct vascular or capsular invasion in an encapsulated follicular thyroid tumour, or when the typical papillary carcinoma nuclei are not visible, morphological criteria may fail to detect the malignancy. Thus, there is clearly a need for more precise methods of detection. Indeed, new techniques based on RET/PTC rearrangements and B-RAF gene mutations have already been developed for the classical and follicular variants of papillary carcinoma. However, although these techniques have allowed the re-qualification of some types of thyroid tumour as malignant (Chiapetta *et al*., 2002), the full set of papillary carcinoma histotypes has

not been covered. Questions have arisen concerning follicular thyroid carcinomas and variants, such as those with PAX8/PPARγ rearrangements, which therefore do not belong to the same category (Marques *et al*., 2002). The frequent association of lymphocytic thyroiditis with papillary thyroid carcinomas and Hurtle cell tumours involves a series of cellular modifications including atrophy, hyperplastic foci, and oncocytic or clear cell metaplasia, all of which render the diagnosis hazardous (Tamimi, 2002).

The modification of genes crucially involved in cancer is not associated with immediate structural changes at the cellular level. However, microarray techniques offer the possibility of investigating the relationship between gene expression patterns and the phenotypic variations that occur during certain steps of cancer development. Thus, gene profiling studies based on microarray analysis have compared various types of differentiated follicular thyroid tumour (Barden *et al.*, 2003; Finley *et al.*, 2004a; Yano *et al.*, 2004; Jarzab *et al.*, 2005; Weber *et al.*, 2005). Although these studies examined the gene expression signature of some follicular or papillary thyroid carcinomas compared to follicular adenomas and normal thyroid tissues, the subcategories to which the tumours belonged were not specified. The gene expression patterns were shown to be correlated with the pathologic diagnosis. Nevertheless, the thyroid cancer signature identified by these studies did not differ from the signatures of other cancer tissues represented in the Oncomine database (Rhodes *et al*., 2004; http:/www.oncomine.org/). The cancer cell profile described may help to identify neoplastic transformation but it fails to take into account the thyroid tumours situated in the grey zone. The predictive accuracy of the gene clusters identified in these studies is therefore rather limited (Aldred *et al.*, 2004; Weber *et al*., 2005). However, other studies have greatly increased this accuracy by including several histotypes of follicular thyroid carcinomas, such as the minimally invasive and oncocytic carcinomas (Finley *et al.*, 2004b; Lubitz *et al.*, 2005); and benign follicular tumours have also been subcategorized by molecular profiling (Finley *et al.*, 2005). These results show that microarray technology is capable of distinguishing potentially malignant from benign follicular thyroid nodules.

The molecular classification of thyroid tumours has been hampered by the small number of cases corresponding to some of the histologically defined classes and subclasses. However, some recent work has revived interest in the subclassification of follicular thyroid adenomas since it was found possible to detect gene profile similarities between one subclass of adenoma and follicular carcinoma (Finley *et al.*, 2004c). Our study, based on the simultaneous analysis of the gene expression signatures of 11 types of histologically defined follicular thyroid lesions,

proposes an extensive molecular classification of these pathologies. This classification should lead to better diagnosis and treatment of follicular thyroid tumours.

Results

Total RNA, isolated from 132 human thyroid lesion samples and 10 benign lesions and 24 controls, as detailed in Table 1, was used to generate radio-labelled cDNA that was hybridized to microarrays. We analysed the expression profiles of 5,549 expressed genes in order to identify the main signatures of the thyroid tissues. Finally, we compared the gene expression profiles of the 11 types of histologically identified follicular thyroid lesions to set up a molecular classification.

Regulated functions in thyroid tumours

From the hierarchical clustering of the genes (Figure 1), we selected 11 main clusters of highly correlated genes ($r>0.5$) representing 2,419 genes. All the samples were heterogeneously classified by the algorithm except for some cases of thyroiditis, oncocytic adenoma and carcinoma, papillary and follicular carcinoma and microfollicular adenoma (Supplementary Figure 1). Examination of the gene ontologies showed that some biological functions are significantly enriched in thyroid pathologies (Figure 1 and Supplementary Table 1). There were four large clusters of more than 400 genes (Clusters 1, 3, 4 and 8). They were representative of most samples (WT, FTA, MNG, and OT) or of most differentially expressed signals (AT). Gene ontologies from Cluster 1 (510 genes) and Cluster 8 (587 genes) were enriched in cell cycle activity, metabolism, and response to stimuli. These two clusters were over-expressed in benign tumours and normal thyroid tissues. Genes from Cluster 3 (418 genes), over-expressed in oncocytic tumours, were involved in cell communication, calcium ion binding, and the immune response. Genes from Cluster 4 (513 genes) were over-expressed in all the AT samples, and in some benign lesions. They were related to the immune response and apoptosis. There were also seven clusters of 26 to 110 genes. These genes were involved in diverse functions such as lipid catabolism and cell differentiation (Cluster 2), translation (Cluster 5), steroid hormone receptor activity and the transcription factor complex (Cluster 6), protein binding and tissue development (Cluster 7), receptor binding (Cluster 9), regulation of transcription and cell differentiation (Cluster 10), membrane fraction and proteolysis (Cluster 11). Several clusters were specific to some types of thyroid lesion: Clusters 2 and 7 were specific to oncocytic tumours as well as AT (either over- or under-expressed); Clusters 5 and 6

to AT (under-expressed); Cluster 9 to oncocytic tumours (under-expressed); and Clusters 10 and 11 to PTC (over-expressed).

Molecular classification of thyroid tissues

In order to compare the 12 histopathological tissue classes studied, we computed the 66 pairwise comparisons on the class centroids (Supplementary Table 2) and applied an unsupervised clustering method. This allowed a global classification of the thyroid tumours, and the visualization of all the pairwise comparisons (Figure 2). We used multi-dimensional scaling to study the homogeneity or heterogeneity of the samples (Figure 3).

The global classification revealed three well-defined clusters of thyroid tumours grouping the benign/normal, malign and oncocytic classes by intra-group similarity and inter-group dissimilarity. The benign/normal cluster grouped AT, GD, FTA-a, WT and MNG. We found significant similarities ($p<0.05$) between AT, GD and WT, between FTA-a and MNG, and between MNG and WT. AT was also found significantly similar to OTA-aty and PTC. Figure 3a shows the relative distribution of the samples. The samples, particularly those of the GD class, were rather heterogeneous. The FTA-a and MNG samples were close together, suggesting their similarity. In the malignant cluster (FTC, FTA-aty and PTC), FTA-aty appeared to be significantly similar to FTC as well as to PTC. Figure 3b shows the intermediate position of FTA-aty samples between the two carcinomas. In the oncocytic cluster (OTA, OTA-aty and OTC), the three classes were significantly similar to each other. With multi-dimensional scaling on two main axes, the oncocytic samples were not separated (Figure 3c). However, when three main axes were used, OTC samples were separated from OTA (data not shown), whereas the OTA-aty samples were heterogeneous and could not be distinguished from OTA. FTA-b was not found significantly similar to any other class as a whole, but its samples were close to some OTA samples.

Determination of optimal classifiers

We evaluated the predictive power of the gene-expression data for the classification of each sample. It has been shown that the performance of a classifier depends on the cut-off used for gene selection (Yukinawa *et al*., 2006). We used the scoring function for four algorithms to search for the best subset of genes by varying the gene selection (Figure 4a).

During the leave-one-out cross-validation, the nearest centroid (NC) classifier and the linear diagonal discriminant analysis (LDDA) behaved similarly. These algorithms were powerful

enough to distinguish the greatest number of classes with at least 349 genes (p<1E-07). When fewer genes were involved, the 1-nearest-neighbour (1-NN) classifier was better, reaching its best performance with 170 genes (p<1E-09). The 3-nearest-neighbours classifier (3-NN) had the worst performance. For all the classifiers, very stringent cut-offs (<1E-13) impaired performance.

The cross-validated results of the best global classifier (1-NN, p<1E-9) are shown in Figure 4b. These results were supported by 1000 bootstrapped datasets (p<1E-03). We examined two intuitive parameters i.e. the sensitivity (SEN) and the positive predictive value (PPV). We also examined two criteria that summed up the performances, i.e. the classical accuracy coefficient (ACC) and the geometric accuracy (G). With the SEN and PPV criteria, the performances varied greatly from very good detection for FTC and WT to null detection for OTA-aty. Though the cases were very different, the ACC did not reflect this and was consistently greater than 0.81. For GD, the global classifier performed poorly (SEN=0.4, PPV=0.25, G=0.32) but the ACC was high (0.95), greater than that of the OTA class (SEN=0.67; PPV=0.63; G=0.65; ACC=0.87). For OTA-aty, no true positives were predicted (SEN=0; PPV=0), but the ACC was 0.95, whereas G was null. The ACC led to over-estimation but G, the geometric mean as used in machine-learning techniques (Kubat, 1998), proved more intuitive to use.

The algorithm worked well for FTC, WT, PTC and OTA (G values between 0.67 and 0.82), whereas GD, MNG and FTA-b were poorly detected (G values between 0.32 to 0.42), and OTA-aty had the worst result (G=0). Other classes were detected with medium performances (G values between 0.5 to 0.63). Remarkably, the classifier selected only true positives for FTC, FTA-aty and OTC (PPV=1).

By considering each class separately, we found 12 individual classifiers which maximized the class prediction (Figure 4c). We found a good individual classifier for each class except for FTC and FTA-aty for which the global classifier was the best. For six classes, 1-NN was the best individual classifier. Cut-offs varied from 1E-03 (811 genes) to 1E-15 (12 genes). GD and OTA-aty were also poorly classified. The WT, AT, PTC and FTC classes showed the best results (G > 0.82).

Protein expression of selected candidate marker genes

Nine protein markers were chosen from the list of differential genes. Antibodies were selected for their commercial availability. The homogeneity of protein levels was checked against α-

tubulin expression. The genic ontology of these markers showed that they were representative of four cellular functions: lipid and mitochondrial metabolisms, cell proliferation and metastatic potential.

For the protein study, we grouped the follicular adenomas (FTA-a, FTA-b, MNG), excluding the atypical adenomas (OTA-aty and FTA-aty). We compared the mean expression of the proteins and the related genes in seven types of thyroid tumour and the wild type tissue (Figure 5a). Genes, proteins and samples were ordered according to the hierarchical clustering of gene expression levels. According to the thyroid tumours and the genes considered, the expression levels were either similar or dissimilar. For each data set, we considered levels close to the global median as iso-expressed (10 percentiles around the median value). There were 37 (51%) coherent gene and protein expression levels, 8 (11%) combinations of iso-expression and under/over-expression, and 27 (37.5%) incoherent expression levels.

The comparison of the gene and protein expression profiles (Figure 5b) revealed five markers (SDHA, CTNNA1, CRABP1, TIMP1 and PDK1) with high coefficients of correlation ($R>0.63$); two markers (APOD and VDAC1) had relatively high coefficients (respectively 0.37 and 0.34), and two markers (LPL and SNCB) had low coefficients ($R<0.15$). The correlations of the gene and protein expression signatures (Figure5c) showed that only FTC and PTC did not have high coefficients ($R<0.4$).

Discussion

Gene expression profiling is a powerful technique for finding new molecular markers for tumour growth as well as for providing insights into the molecular mechanisms involved in tumorigenesis. However, current microarray studies on thyroid tumours fail to propose truly specific markers since they are based on incomplete sets of thyroid tumour classes.

Thyroid tumours are heterogeneous and often subject to diagnostic misclassification. The classical methods of analysis are affected by weak intra-class homogeneity and by unbalanced classes since very few samples are generally available for the rarer pathologies. Although the molecular signatures of thyroid tumours can be very heterogeneous, they allowed a classification (Figure 2), dividing the tumours into three distinct groups: benign/normal, malignant and oncocytic. Pairwise distinctions of benign/normal and malign tumours were consistent with published reports, as in the case of PTC and FTA. Moreover, thanks to the almost complete set of thyroid tumour classes included in our study, the results may be

generalized to the whole range of thyroid tumours, from the benign/normal to the malignant and oncocytic tumours. Microfollicular thyroid adenoma presented a signature different from that of macrofollicular adenoma or follicular carcinoma. The malignant evolution of the microfollicular adenoma is still under discussion (Schmid *et al.,* 2006). However, we may consider microfollicular thyroid adenoma as a distinct subtype because its gene expression reflects its pathogenesis even though we do not know the impact of this specific profile on the tumoral outcome. Moreover, variations in the gene expression signatures of macrofollicular adenomas (FTA-a and MNG) do not reflect the apparition of multiple nodules in the thyroid (Figure 3a), indicating that the automated classifiers used performed poorly in these cases (Figure 4). The two types of thyroid tumour should therefore be considered as belonging to the same class on the basis of global gene expression analysis. The reported incidence of thyroid carcinoma in MNG (Gandolfi *et al.,* 2004) suggests that it would be worth investigating differences in macro- and microfollicular adenomas to better understand microfollicular mechanisms and their potential link to oncocytomas.

Oncocytomas appear to be unrelated to PTCs or FTCs in the classification of thyroid tumours based on gene expression (Figure 3a); they should therefore be analysed independently. This finding is in accordance with our previous study on oncocytic tumours that showed no functional relation between oncocytic tumours and mitochondrial rich PTCs (Baris *et al.,* 2005). The FTA-aty class of tumours was significantly similar to the PTC as well as the FTC classes (Figure 2). The correlation with the PTC class was higher but this may be explained by the small size of the FTC sample; larger samples should produce more conclusive results. The close relationship between the PTC and FTA-aty classes argues in favour of the proposed terminology referring to these tumours as differentiated tumours of uncertain malignant potential.

The identification of pathologic tissue by its morphological features is not always precise. In the case of thyroid tumours, significant inter- and intra-observer variations have been reported (Hirokawa *et al*., 2002; Franc *et al*., 2003; Llyod *et al*., 2004). Gene mutations and rearrangements can define a class and its evolution more precisely than the morphology, as has been demonstrated in the case of papillary and follicular thyroid carcinomas (Giordano *et al.,* 2005; Giordano *et al.,* 2006). However, some histotypes, such as oncocytic carcinomas, escape this molecular classification (Nikiforova *et al.,* 2003). Several authors have evaluated the usefulness of biomarker combinations for the diagnosis of suspicious thyroid tumours (Ito *et al*., 2005; Prasad *et al.,* 2005; Cerutti *et al*., 2006). However, biomarker combinations have

failed to make a significant contribution in clinical practice. We postulate that the weak specificity of the markers may be compensated by the definition of the functional profiles of thyroid tumours. In our study, we tested protein markers belonging to four functional classes, i.e. cell proliferation (PDK1, CRABP1, SNCB and CTNNA1), metastasis (TIMP1), lipid metabolism (LPL and APOD) and mitochondrial metabolism (SDHA and VDAC1). Gene and protein expression levels differed with only two of the nine markers, confirming the relevance of the genes selected. The development of new antibodies against better protein markers may lead to more accurate diagnoses of suspicious lesions. However, the advantages would be rapidly limited by the number of antibodies usable in practice. Our study shows that the classification of thyroid lesions based on gene expression patterns is a useful complement for the pathologic classification (Figure 4).

With microarray data obtained from a large number of samples and unbalanced class sizes, the classical accuracy coefficient (ACC) could not be used as it would have tended to over-estimate the performances (Figure 4b). We therefore preferred to use the geometric accuracy G, which is unbiased and more intuitive (Kubat, 1998). Based on forward selection and univariate tests, the identification of the 12 molecular classes of follicular thyroid tumour demonstrated good predictive accuracy. However, one class, the OTA-Aty class, presented a low G value, suggesting that this class, in contrast to the FTA-Aty class, cannot be distinguished at the molecular level. We have previously suggested the existence of a continuum in the development of oncocytic tumours since OTAs and OTCs were found to be frequently close together on microarray analysis (Baris *et al*., 2004). The meta-analysis of thyroid tumour microarray data may be compromised by partial gene chip compatibility and differences in data processing techniques (Eszlinger *et al.,* 2007). Our data set circumvents these difficulties and it would be worth using to test classifier algorithms based on multivariate gene selection methods.

Finally, we have developed a highly sensitive method for the molecular classification of histologically identified types of thyroid tumour. Our study shows that dedicated microarray analysis with specific subsets of differential genes allows the identification of the majority of tumours originating from follicular thyroid cells. In particular, the method can be conveniently applied to fine-needle aspiration biopsies since it requires very small samples compared to cytological studies. As some thyroid lesions are morphologically heterogeneous, it may be necessary to sample several fine-needle aspirates so as not to miss the atypical and malignant tumours that may be present. The gene expression profiles could then be used as a complement

to the histological findings for the improved diagnosis and treatment of follicular thyroid tumours.

Materials and Methods

Thyroid tissue samples

Tissue samples were obtained from 132 human thyroid tumours, 10 benign lesions and 24 normal tissues. The thyroid tumours included 26 macrofollicular, 17 microfollicular and 10 atypical adenomas. Atypical adenomas were defined by their nuclear features, and vascular or cellular modifications; they showed no capsular or vascular invasion. In addition, there were 24 thyroid adenomas with the largest nodule originating from multinodular goitres; 23 of these had macrofollicular features and one had microfollicular features. We also examined 30 oncocytic follicular adenomas and 5 atypical oncocytic adenomas defined on the same criteria as the non-oncocytic adenomas. Our study included three types of thyroid carcinoma: follicular (3 cases), oncocytic (4 cases), and papillary carcinoma (13 cases). Benign lesions were 5 samples of autoimmune thyroiditis, and 5 samples of Graves' disease. Control samples consisted of 24 normal counterpart thyroid tissues. Table 1 summarizes the characteristics of the samples.

The mitochondrial quantities were evaluated by immunohistochemistry, using monoclonal anti-cytochrome c oxydase antibody (Clone 113-1, Biogenex Laboratories, Inc., San Ramon, CA, USA). The diagnoses were made according to the WHO classification of thyroid tumours (DeLellis *et al*., 2004). Eighty-seven anonymous samples were obtained from the Ambroise Paré Hospital (APHP, Boulogne s/Seine, France), and 79 anonymous samples from the University Hospital of Angers, France.

cDNA arrays

RNA extraction, cDNA preparation and hybridization, scanning, and image analysis of the arrays were done according to the manufacturers' protocols (http://tagc.univ-mrs.fr/plateforme/protocoles/), and as previously described (Baris *et al.*, 2004). All data were subjected to print-tip Lowess normalization (Yang *et al.*, 2002). The microarray data collected during the analysis have been deposited in NCBI Gene Expression Omnibus (Edgar *et al.*, 2002) and are accessible through GEO Series N° GSE6339.

Data analysis and statistical methods

Statistical tests and clustering

Differential genes were selected by permuted non-parametric tests (Ge *et al.,* 2003). A false discovery rate of 0.05 computed from 1000 bootstrapped datasets allowed the selection of 818 genes. Hierarchical clustering of the genes and the samples was done using average linkage and uncentred correlation distances. Computation and visualization were done with Cluster and Java TreeView software (Eisen *et al*., 1998). Gene ontology enrichments into gene clusters were computed by means of the GOToolBox web server (Martin *et al.,* 2004). Outlier values were filtered for protein-expression data by removing measurements that diverged from the class mean by more than 1.5 times the standard deviation.

Similarity analysis

We defined a centroid for each class as its mean gene-expression signature. The similarity S was given by the correlation coefficient of two centroids. The significance of similarity was assessed by two-tailed permuted tests (5 % risk with the Bonferroni correction) from 20 000 bootstrapped data sets. For clustering, pairwise similarity parameters were normalized by their expected value, clustered by Cluster software (Kendall's Tau metric, average linkage method), and visualized by TreeView Software.

Automated classifiers

Automated classification was performed using BRB ArrayTools developed by Dr. Richard Simon and Amy Peng (http://linus.nci.nih.gov/BRB-ArrayTools.html). We tested four algorithms: LDDA, 1-NN, 3-NN and NC. We computed the prediction for several gene subsets selected by cut-offs (ranging from 1E-2 to 1E-15) determined by univariate tests. All the performances were assessed from a leave-one-out procedure unbiased for feature selection (Ambroise *et al.,* 2002). To evaluate classifier performances for class A we used the sensitivity (true positives / total number of A samples), the positive predictive value (true positives / samples detected as A). We used the classical accuracy coefficient ((true positives + true negatives) / total number of samples) and the geometric accuracy (*G*) defined as the geometric mean of the sensitivity and the positive predictive value.

Dot blot analysis

Two micrograms of protein corresponding to the TRIzol fraction of several tissues used for microarrays analysis were spotted onto nitrocellulose membranes at room temperature, using a

dot-blot apparatus following the manufacturer's recommendations (Biorad, Hercules, CA). Seven groups of thyroid pathologies were defined on non-ambiguous cytological criteria: AT, GD, FTA-a, FTC, PTC, OTA, and OTC. Five samples were tested for each group. For two groups (FTC, OTC), samples previously untested on microarrays were added to bring the total number of samples to five. Ten sera were used at specific dilutions: 1/750 for PDK1 and TIMP1, 1/1000 for APOD, SDHA, CTNNA1, VDAC1, CRABP1, LPL, SNCB, 1/5000 for α-tubulin. All of the sera were obtained from Abcam (Cambridge, UK). The spots revealed by ECL methodology (ECL Plus reagent kit, Amersham, Chalfont, UK) were quantified on a GelDoc XRS apparatus using the Quantity One software (Biorad). All the assays were made in duplicate.

Acknowledgments

We thank Marielle Mello, Benoit Ballester, Dominique Couturier and Anne Coutoleau for technical help and data treatment. We thank Kanaya Malkani for the critical reading of this paper. This work was supported by grants from the French Ministry of Research, the *Institut National de la Recherche Médicale* (INSERM), the University Hospital of Angers and the University of Angers (PHRC 03-10).

References

- Aldred MA, Huang Y, Liyanarachchi S, Pellegata NS, Gimm O, Jhiang S, et al. (2004). Papillary and follicular thyroid carcinomas show distinctly different microarray expression profiles and can be distinguished by a minimum of five genes. J Clin Oncol, 22: 3531-3539.
- Ambroise C , McLachlan GJ. (2002). Selection bias in gene extraction on the basis of microarray gene-expression data. Proc Natl Acad Sci U S A, 99: 6562-6566.
- Barden CB, Shister KW, Zhu B, Guiter G, Greenblatt DY, Zeiger MA, et al. (2003). Classification of follicular thyroid tumors by molecular signature: results of gene profiling. Clin Cancer Res, 9: 1792-1800.
- Baris O, Mirebeau-Prunier D, Savagner F, Rodien P, Ballester B, Loriod B, et al. (2005). Gene profiling reveals specific oncogenic mechanisms and signaling pathways in oncocytic and papillary thyroid carcinoma. Oncogene, 24: 4155-4161.
- Baris O, Savagner F, Nasser V, Loriod B, Granjeaud S, Guyetant S, et al. (2004). Transcriptional profiling reveals coordinated up-regulation of oxidative metabolism genes in thyroid oncocytic tumors. J Clin Endocrinol Metab, 89: 994-1005.
- Cerutti JM, Latini FR, Nakabashi C, Delcelo R, Andrade VP, Amadei MJ, et al. (2006). Diagnosis of suspicious thyroid nodules using four protein biomarkers. Clin Cancer Res, 12: 3311-3318.
- Chiappetta G, Toti P, Cetta F, Giuliano A, Pentimalli F, Amendola I, et al. (2002). The RET/PTC oncogene is frequently activated in oncocytic thyroid tumors (Hurthle cell adenomas and carcinomas), but not in oncocytic hyperplastic lesions. J Clin Endocrinol Metab, 87: 364-369.
- DeLellis R, Lloyd R, Heitz P , Heng C. (2004). 320.
- Edgar R, Domrachev M , Lash AE. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res, 30: 207-210.
- Eisen MB, Spellman PT, Brown PO , Botstein D. (1998). Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A, 95: 14863-14868.
- Eszlinger M, Krohn K, Kukulska A, Jarzab B , Paschke R. (2007). Perspectives and Limitations of Microarray-Based Gene Expression Profiling of Thyroid Tumors. Endocr Rev. Apr 12; [Epub ahead of print]
- Finley DJ, Zhu B, Barden CB , Fahey TJ, 3rd. (2004a). Discrimination of benign and malignant thyroid nodules by molecular profiling. Ann Surg, 240: 425-436.
- Finley DJ, Arora N, Zhu B, Gallagher L , Fahey TJ, 3rd. (2004b). Molecular profiling distinguishes papillary carcinoma from benign thyroid nodules. J Clin Endocrinol Metab, 89: 3214-3223.
- Finley DJ, Zhu B , Fahey TJ, 3rd. (2004c). Molecular analysis of Hurthle cell neoplasms by gene profiling. Surgery, 136: 1160-1168.
- Finley DJ, Lubitz CC, Wei C, Zhu B , Fahey TJ, 3rd. (2005). Advancing the molecular diagnosis of thyroid nodules: defining benign lesions by molecular profiling. Thyroid: 15, 562-568.
- Franc B, de la Salmoniere P, Lange F, Hoang C, Louvel A, de Roquancourt A, et al. (2003). Interobserver and intraobserver reproducibility in the histopathology of follicular thyroid carcinoma. Hum Pathol, 34: 1092-1100.
- Gandolfi PP, Frisina A, Raffa M, Renda F, Rocchetti O, Ruggeri C, et al. (2004). The incidence of thyroid carcinoma in multinodular goiter: retrospective analysis. Acta Biomed, 75: 114-117.
- Ge Y, Dudoit S, Speed TP (2003). Resampling-based multiple testing for microarray data hypothesis. Test 12: 1-44.
- Giordano TJ, Au AY, Kuick R, Thomas DG, Rhodes DR, Wilhelm KG, Jr., et al. (2006). Delineation, functional validation, and bioinformatic evaluation of

gene expression in thyroid follicular carcinomas with the PAX8-PPARG translocation. Clin Cancer Res, 12: 1983-1993.

- Giordano TJ, Kuick R, Thomas DG, Misek DE, Vinco M, Sanders D, et al. (2005). Molecular classification of papillary thyroid carcinoma: distinct BRAF, RAS, and RET/PTC mutation-specific gene expression profiles discovered by DNA microarray analysis. Oncogene, 24: 6646-6656.

- Hedinger C, Williams ED , Sobin LH. (1989). The WHO histological classification of thyroid tumors: a commentary on the second edition. Cancer, 63: 908-911.

- Hirokawa M, Carney JA, Goellner JR, DeLellis RA, Heffess CS, Katoh R, et al. (2002). Observer variation of encapsulated follicular lesions of the thyroid gland. Am J Surg Pathol, 26: 1508-1514.

- Ito Y, Yoshida H, Tomoda C, Miya A, Kobayashi K, Matsuzuka F, et al. (2005). Galectin-3 expression in follicular tumours: an immunohistochemical study of its use as a marker of follicular carcinoma. Pathology, 37, 296-298.

- Jarzab B, Wiench M, Fujarewicz K, Simek K, Jarzab M, Oczko-Wojciechowska M, et al. (2005). Gene expression profile of papillary thyroid cancer: sources of variability and diagnostic implications. Cancer Res, 65: 1587-1597.

- Kubat M. (1998). Machine Learning for the Detection of Oil Spills in Satellite Radar Images. Machine Learning, 30: 195-215.

- Lloyd RV, Erickson LA, Casey MB, Lam KY, Lohse CM, Asa SL, et al. (2004). Observer variation in the diagnosis of follicular variant of papillary thyroid carcinoma. Am J Surg Pathol, 28: 1336-1340.

- Lubitz CC, Gallagher LA, Finley DJ, Zhu B , Fahey TJ, 3rd. (2005). Molecular analysis of minimally invasive follicular carcinomas by gene profiling. Surgery, 138: 1042-1048.

- Lubitz CC, Ugras SK, Kazam JJ, Zhu B, Scognamiglio T, Chen YT, et al. (2006). Microarray analysis of thyroid nodule fine-needle aspirates accurately classifies benign and malignant lesions. J Mol Diagn, 8: 490-498.

- Marques AR, Espadinha C, Catarino AL, Moniz S, Pereira T, Sobrinho LG, et al. (2002). Expression of PAX8-PPAR gamma 1 rearrangements in both follicular thyroid carcinomas and adenomas. J Clin Endocrinol Metab, 87: 3947-3952.

- Martin D, Brun C, Remy E, Mouren P, Thieffry D , Jacq B. (2004). GOToolBox: functional analysis of gene datasets based on Gene Ontology. Genome Biol, 5: R101.

- Nikiforova MN, Lynch RA, Biddinger PW, Alexander EK, Dorn GW, 2nd, Tallini G, et al. (2003). RAS point mutations and PAX8-PPAR gamma rearrangement in thyroid tumors: evidence for distinct molecular pathways in thyroid follicular carcinoma. J Clin Endocrinol Metab, 88: 2318-2326.

- Prasad ML, Pellegata NS, Huang Y, Nagaraja HN, de la Chapelle A , Kloos RT. (2005). Galectin-3, fibronectin-1, CITED-1, HBME1 and cytokeratin-19 immunohistochemistry is useful for the differential diagnosis of thyroid tumors. Mod Pathol, 18: 48-57.

- Rhodes DR, Yu J, Shanker K, Deshpande N, Varambally R, Ghosh D, et al. (2004). ONCOMINE: a cancer microarray database and integrated data-mining platform. Neoplasia, 6: 1-6.

- Schmid KW , Farid NR. (2006). How to define follicular thyroid carcinoma? Virchows Arch, 448: 385-393.

- Tamimi DM. (2002). The association between chronic lymphocytic thyroiditis and thyroid tumors. Int J Surg Pathol, 10, 141-146.

- Weber F, Shen L, Aldred MA, Morrison CD, Frilling A, Saji M, et al. (2005). Genetic classification of benign and malignant thyroid follicular neoplasia based on a three-gene combination. J Clin Endocrinol Metab, 90: 2512-2521.

- Williams ED. (2000). Guest Editorial: Two Proposals Regarding the Terminology of Thyroid Tumors. Int J Surg Pathol, 8: 181-183.

- Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, et al. (2002). Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res, 30: e15.

- Yano Y, Uematsu N, Yashiro T, Hara H, Ueno E, Miwa M, et al. (2004). Gene expression profiling identifies platelet-derived growth factor as a diagnostic molecular marker for papillary thyroid carcinoma. Clin Cancer Res, 10: 2035-2043.

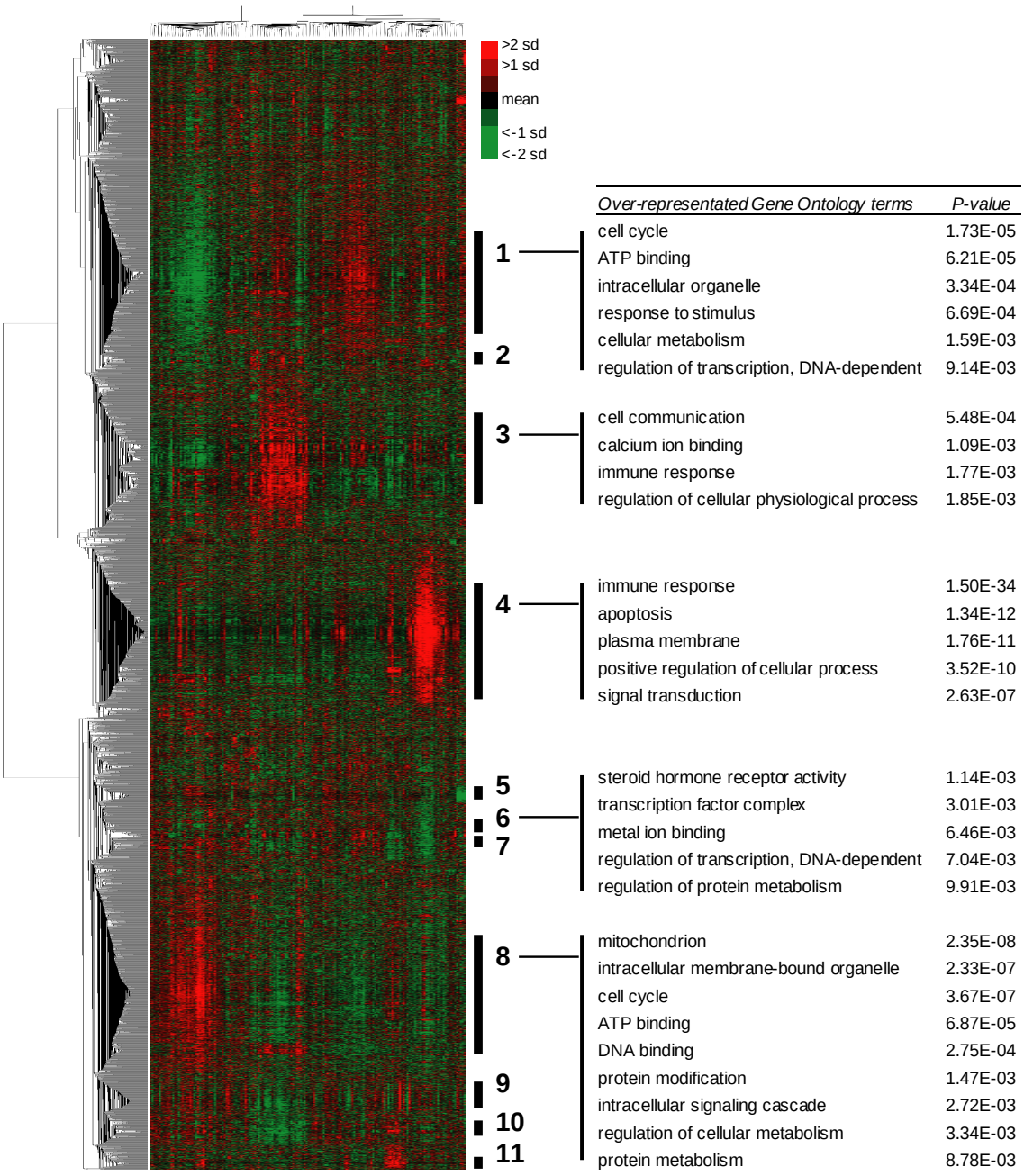
- Yukinawa N, Oba S, Kato K, Taniguchi K, Iwao-Koizumi K, Tamaki Y, et al. (2006). A multi-class predictor based on a probabilistic model: application to gene expression profiling-based diagnosis of thyroid tumors. BMC Genomics, 7: 190.

Tables et figures

Table 1: Thyroid tumours and control tissues examined for gene expression patterns.

Thyroid tumours and controls	***N***	***Histology***	***Dominant microfollicular architecture***
Controls			
Wild type (WT)	24	Normal	
Non-tumoural lesions			
Graves' disease (GD)	5	Diffuse goitre	
Autoimmune thyroiditis (AT)	5	Diffuse lymphocyte thyroiditis	
Benign tumours			
Oncocytic adenomas (OTA)	30	Over 95% oncocytic cells and mitochondrial antibody ++	17 cases
Follicular thyroid adenomas (FTA) *with complete fibrous capsule*			
Macrofollicular (FTA-a)	26	Unique nodule	
Microfollicular (FTA-b)	17	Unique nodule	17 cases
Multinodular goitre (MNG)	24	Largest nodule of a multinodular goitre	1 case
Atypical thyroid adenomas with complete fibrous capsule and moderate vascular invasion			
Follicular (FTA-aty)	10	Unique nodule and atypical nucleus	1 case
Oncocytic (OTA-aty)	5	Unique nodule and atypical nucleus	2 cases
Differentiated malignant thyroid tumours with capsular and vascular invasion			
Follicular carcinomas (FTC)	3	Unique	3 cases
Oncocytic carcinomas (OTC)	4	Over 95% oncocytic cells and mitochondrial antibody ++, Trabecular architecture	
Papillary carcinomas (PTC)	13	Typical papillary nuclear features, 8 follicular variants	

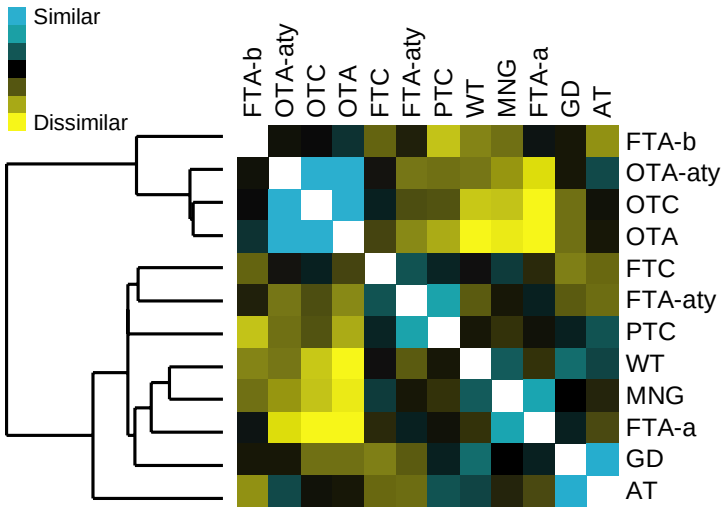
Figure 1: Hierarchical clustering of the genes.



Over-representated Gene Ontology terms	*P-value*
1	
cell cycle	1.73E-05
ATP binding	6.21E-05
intracellular organelle	3.34E-04
response to stimulus	6.69E-04
cellular metabolism	1.59E-03
regulation of transcription, DNA-dependent	9.14E-03
2	
3	
cell communication	5.48E-04
calcium ion binding	1.09E-03
immune response	1.77E-03
regulation of cellular physiological process	1.85E-03
4	
immune response	1.50E-34
apoptosis	1.34E-12
plasma membrane	1.76E-11
positive regulation of cellular process	3.52E-10
signal transduction	2.63E-07
5	
6	
7	
steroid hormone receptor activity	1.14E-03
transcription factor complex	3.01E-03
metal ion binding	6.46E-03
regulation of transcription, DNA-dependent	7.04E-03
regulation of protein metabolism	9.91E-03
8	
mitochondrion	2.35E-08
intracellular membrane-bound organelle	2.33E-07
cell cycle	3.67E-07
ATP binding	6.87E-05
DNA binding	2.75E-04
9	
protein modification	1.47E-03
intracellular signaling cascade	2.72E-03
10	
regulation of cellular metabolism	3.34E-03
11	
protein metabolism	8.78E-03

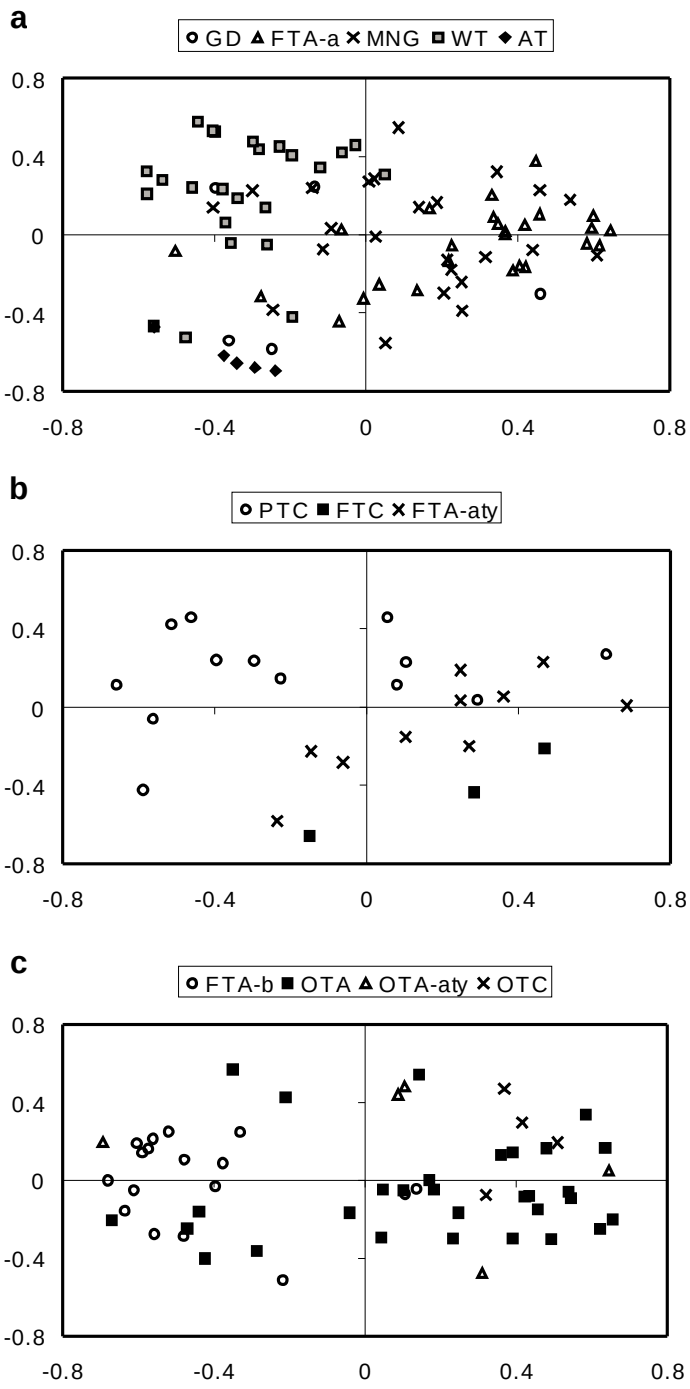
This matrix depicts the expression signatures of 166 thyroid tumour tissues (columns) through 5,549 selected genes (rows). A colour gradient indicates gene expression levels, from green (low levels), to red (high levels), with black representing the median value. Clusters of highly correlated genes (r>0.5) are flagged and numbered. Five clusters are annotated with the best enriched Gene Ontology terms.

Figure 2: Classification of pathologies according to gene expression similarities.



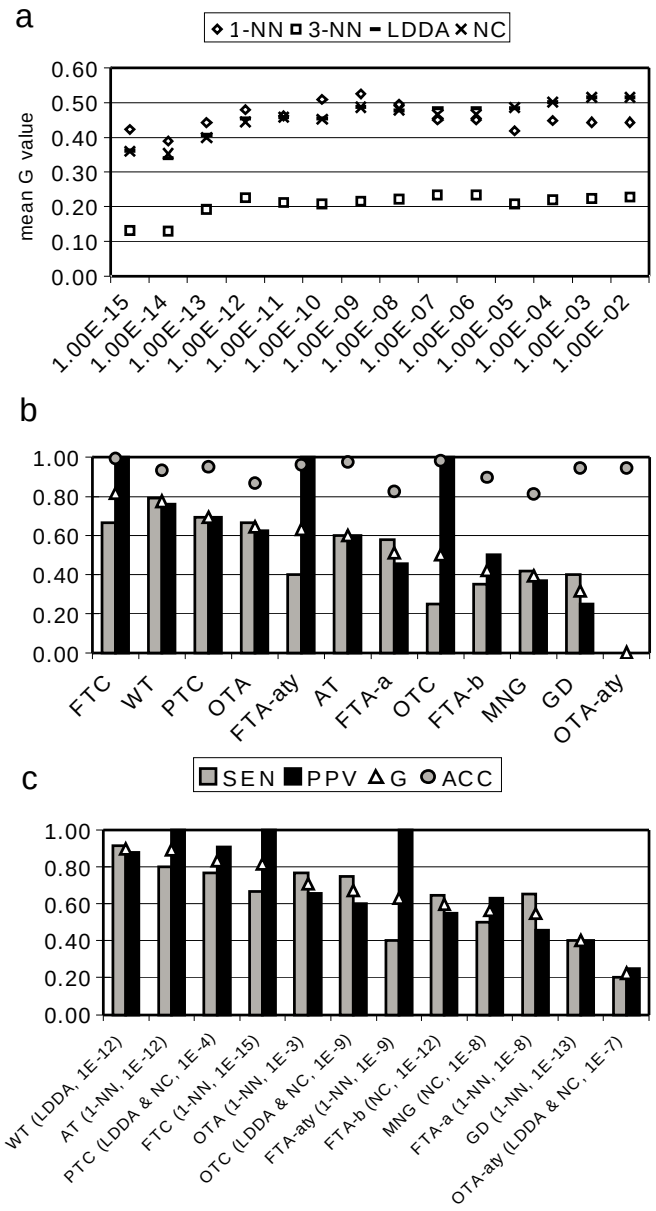
This symmetrical colour-coded matrix shows similarity coefficients between pairs of thyroid tumour classes. The values were centred on their expected mean from 20,000 bootstrapped datasets. The dendogram at the top was generated by a hierarchical clustering algorithm (Kendall's Tau metric, average linkage method).

Figure 3: Multi-dimensional scaling of the samples.



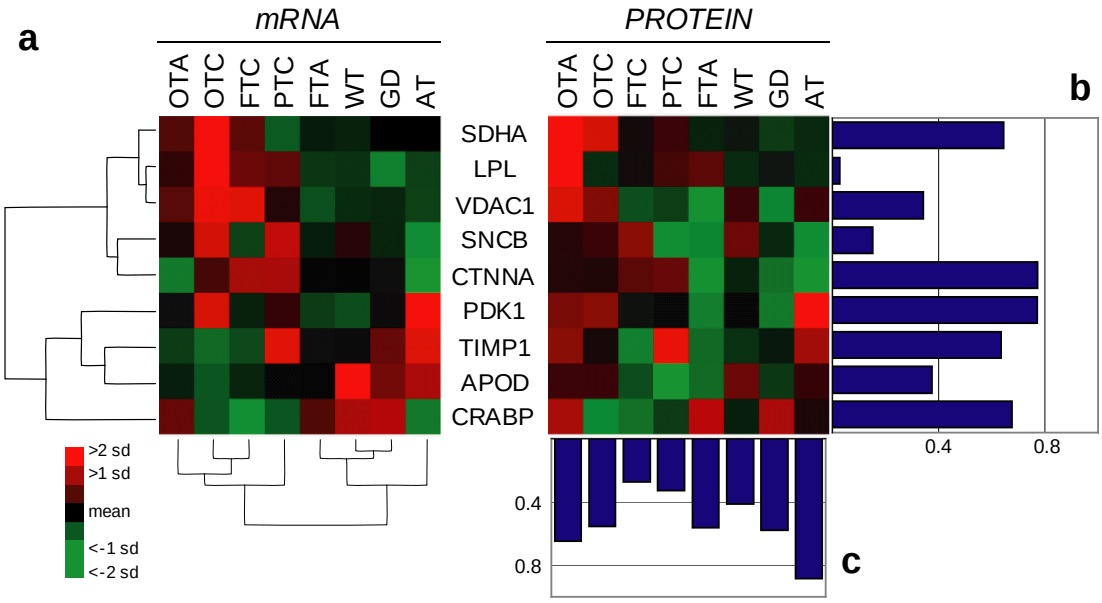
Samples are represented along the two main axes of a multi-dimensional scaling scheme for the differential genes. Panel a shows the projection of benign thyroid tumours and normal thyroid tissue; Panel b shows FTA-aty and the two carcinomas; and Panel c shows FTA-b and the oncocytic classes of thyroid tumour.

Figure 4: Automated classifier performance.



From the 818 differential genes, we selected 14 subsets according to cut-offs determined by univariate tests (from 1E-2 to 1E-15). Panel a shows the comparison of the four algorithms used (1-NN, 3-NN, LDDA and NC) to select the best global classifier. Panel b shows the best global algorithm (1-NN, cut-off of 1E-9). Class performances are summarized by either the classical accuracy coefficient ACC (grey squares) or the geometric accuracy G (black squares). Discrepancies between ACC and G can be compared with the intuitive parameters, i.e. the sensitivity SEN (white bars) and the positive predictive value PPV (grey bars). Panel c shows for each class the best performance that can be attained by the classifiers.

Figure 5: Protein expression levels of selected marker genes



Gene and protein expression levels are compared for 8 groups of samples (panel a). Correlation coefficients of gene and protein profiles across samples are reported on panel b. Correlation coefficients of tissue groups across gene and protein expression levels are reported on panel c.

Supplementary Table 1: Enriched functions in highly correlated gene clusters.

Cluster	*Genes*	*Gene Onotology term*	*Fref*	*Fset*	*P-value*
1	510	cell cycle	0.04	0.10	1.73E-05
		ATP binding	0.10	0.17	6.21E-05
		intracellular organelle	0.47	0.59	3.34E-04
		response to stimulus	0.13	0.20	6.69E-04
		cellular metabolism	0.54	0.63	1.59E-03
		regulation of transcription, DNA-dependent	0.13	0.18	9.14E-03
2	26	lymphocyte activation	0.00	0.15	1.24E-03
		lipid catabolism	0.00	0.15	1.56E-03
		plasma membrane	0.10	0.50	1.80E-03
		positive regulation of cell proliferation	0.01	0.15	2.30E-03
		cell differentiation	0.02	0.23	2.56E-03
3	418	cell communication	0.24	0.34	5.48E-04
		calcium ion binding	0.05	0.11	1.09E-03
		immune response	0.05	0.11	1.77E-03
		regulation of cellular physiological process	0.19	0.28	1.85E-03
4	513	immune response	0.05	0.31	1.50E-34
		apoptosis	0.03	0.13	1.34E-12
		plasma membrane	0.10	0.26	1.76E-11
		positive regulation of cellular process	0.02	0.11	3.52E-10
		signal transduction	0.19	0.33	2.63E-07
5	39	translation	0.01	0.14	2.45E-03
		cytoplasm organization and biogenesis	0.00	0.10	4.37E-03
		protein metabolism	0.25	0.52	5.30E-03
6	32	steroid hormone receptor activity	0.00	0.13	1.14E-03
		transcription factor complex	0.01	0.15	3.01E-03
		metal ion binding	0.21	0.50	6.46E-03
		regulation of transcription, DNA-dependent	0.13	0.40	7.04E-03
		regulation of protein metabolism	0.01	0.13	9.91E-03
7	67	protein binding	0.19	0.48	1.78E-04
		tissue development	0.01	0.10	2.01E-03
8	587	mitochondrion	0.04	0.13	2.35E-08
		intracellular membrane-bound organelle	0.38	0.55	2.33E-07
		cell cycle	0.04	0.11	3.67E-07
		ATP binding	0.10	0.17	6.87E-05
		DNA binding	0.11	0.19	2.75E-04
		protein modification	0.12	0.19	1.47E-03
		intracellular signaling cascade	0.07	0.11	2.72E-03
		regulation of cellular metabolism	0.15	0.21	3.34E-03
		protein metabolism	0.25	0.31	8.78E-03
9	110	receptor binding	0.03	0.21	5.08E-05
10	87	protein binding	0.19	0.36	5.59E-03
		nervous system development	0.02	0.11	6.29E-03
		regulation of transcription	0.14	0.30	7.15E-03
		cell differentiation	0.02	0.11	7.87E-03
11	30	membrane fraction	0.03	0.31	5.53E-04
		proteolysis	0.05	0.27	4.91E-03
		integral to plasma membrane	0.07	0.31	9.40E-03

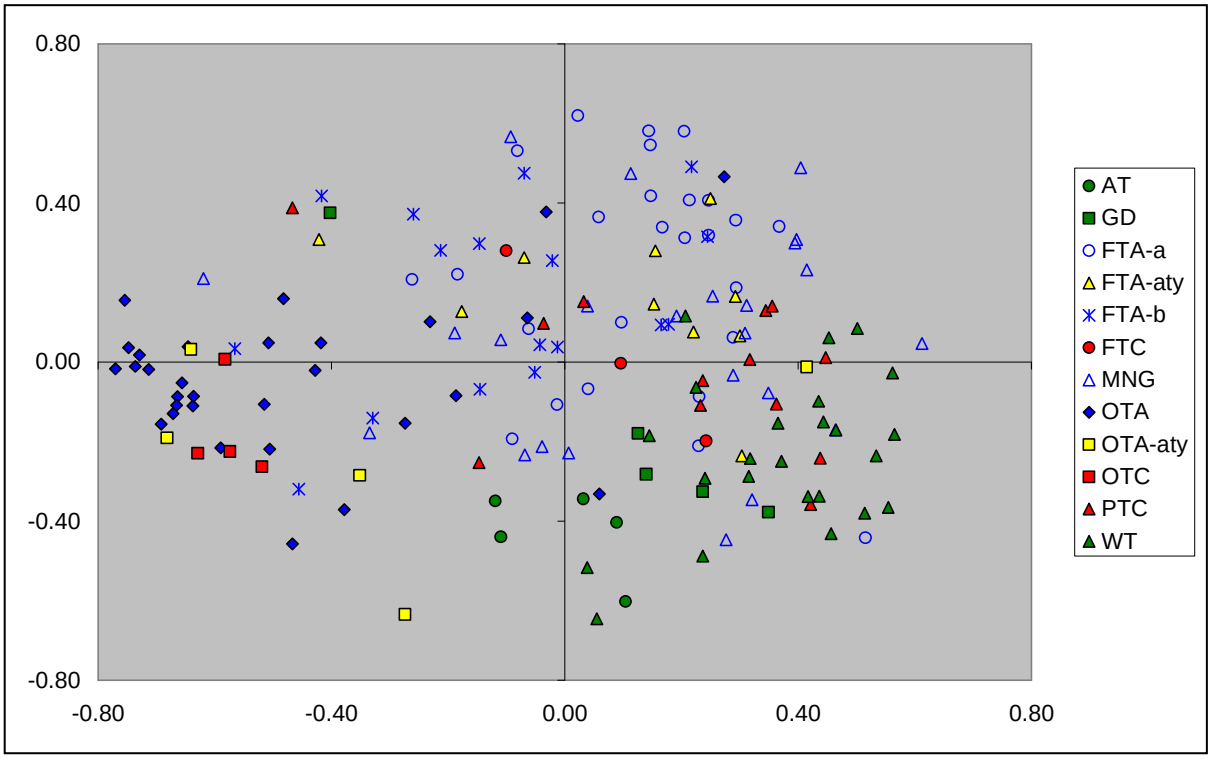
Fref: GO term frequency into the human genome, Fset: GO term frequency into the cluster.

Supplementary Table 2: Pairwise similarity of the thyroid classes.

Similarity	*AT*	*GD*	*FTA-a*	*FTA-aty*	*FTA-b*	*FTC*	*MNG*	*OTA*	*OTA-aty*	*OTC*	*PTC*	*WT*
AT		**0.56**	**-0.13**	**-0.22**	**-0.29**	**-0.22**	-0.05	-0.02	**0.17**	-0.03	**0.2**	**0.18**
GD	**0.56**		0.1	**-0.19**	-0.02	**-0.27**	0.03	**-0.21**	-0.04	**-0.24**	0.09	**0.27**
FTA-a	**-0.13**	0.1		0.12	0.1	-0.07	**0.48**	**-0.48**	**-0.46**	**-0.55**	0.01	-0.04
FTA-aty	**-0.22**	**-0.19**	0.12		-0.03	**0.2**	-0.01	**-0.25**	**-0.24**	**-0.15**	**0.43**	**-0.16**
FTA-b	**-0.29**	-0.02	0.1	-0.03		**-0.2**	**-0.19**	0.18	-0.01	0.02	**-0.39**	**-0.23**
FTC	**-0.22**	**-0.27**	-0.07	**0.2**	**-0.2**		0.15	**-0.13**	-0.02	0.08	0.1	0.04
MNG	-0.05	0.03	**0.48**	-0.01	**-0.19**	0.15		**-0.44**	**-0.3**	**-0.4**	-0.05	**0.27**
OTA	-0.02	**-0.21**	**-0.48**	**-0.25**	0.18	**-0.13**	**-0.44**		**0.71**	**0.77**	**-0.32**	**-0.52**
OTA-aty	**0.17**	-0.04	**-0.46**	**-0.24**	-0.01	-0.02	**-0.3**	**0.71**		**0.64**	**-0.22**	**-0.23**
OTC	-0.03	**-0.24**	**-0.55**	**-0.15**	0.02	0.08	**-0.4**	**0.77**	**0.64**		**-0.16**	**-0.41**
PTC	**0.2**	0.09	0.01	**0.43**	**-0.39**	0.1	-0.05	**-0.32**	**-0.22**	**-0.16**		0
WT	**0.18**	**0.27**	-0.04	**-0.16**	**-0.23**	0.04	**0.27**	**-0.52**	**-0.23**	**-0.41**	0	

Significant values ($p<0.05$) are coloured in red (similarity) or green (dissimilarity).

Supplementary Figure 1: Multi-dimensional scaling of all the samples.



Samples are represented along the two main axes of a multi-dimensional scaling scheme for the differential genes.

3.1.3 Résultats complémentaires

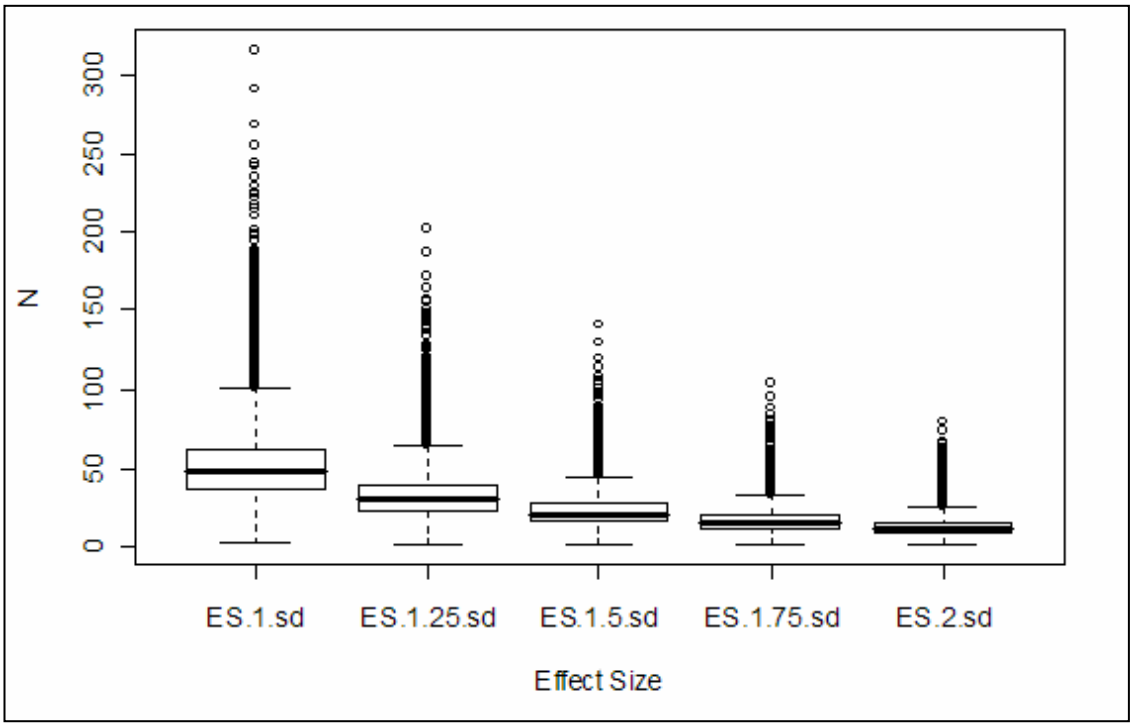
Taille optimale des groupes pour une validation à grande échelle sur 1000 échantillons :

La signature définie par le meilleur prédicteur global pourrait servir à créer un outil de diagnostic sous la forme de puce à ADN dédiée. Dans le cadre d'un projet de validation à grande échelle de cette signature prédictive nous avons besoin de connaître la taille optimale des groupes de tumeurs à tester. La détermination *a priori* des tailles de groupes dans les études sur puces à ADN n'est pas triviale. Elle implique la connaissance de certains paramètres statistiques non disponibles comme la variance de l'expression des gènes. Plusieurs études se sont intéressés à la détermination du nombre minimal d'échantillons à utiliser dans les études sur puces à ADN (pour exemple : Hwang *et al.*, 2002, Wu FX *et al.*, 2006). Il a été montré par étude comparative de plusieurs jeux de données qu'un nombre de 10 à 15 échantillons biologiques par groupes est suffisant pour atteindre un niveau de puissance élevé (Pavlidis *et al.*, 2003).

A partir de la variance observée de 258 gènes dans un échantillon de 166 individus représentant 12 groupes de tissus thyroïdiens, en s'intéressant aux 9 groupes de tumeurs mesurés, peut-on prévoir la taille des groupes pour pouvoir observer des effets significatifs lors d'une analyse indépendante de 1000 individus ? Nous disposons des données d'expression de ces 258 gènes pour 166 prélèvements thyroïdiens. Nous pouvons calculer la variance observée de chaque gène dans chacun des 9 groupes de tumeurs. L'écart type observé de la population peut aussi être calculé pour évaluer l'effet à prendre en compte. Nous utilisons les formules de calcul de la taille des groupes pour le test t définies par Seo *et al*. (2006).

Nous avons calculé la taille minimale N de chaque groupe dans le cas de comparaisons 2 à 2 qui permet d'observer un effet donné ES (exprimé en fonction de l'écart type (sd) de la population) pour des risques $\alpha = 0.05$ et $\beta = 0.10$ (puissance $1 - \beta = 0.90$). La méthode de Bonferroni a été utilisée contre l'effet de tests multiples. La Figure 2 montre les distributions des valeurs N en fonction de l'effet à observer. La taille N des groupes à considérer varie en fonction de l'effet désiré. Plus l'effet est grand et moins N est grand. De façon intuitive, les effets les plus subtils nécessitent plus d'individus pour être détectés. Pour chaque distribution, l'espace interquartile est resserré autour de la médiane, il représente environ 8% de l'amplitude. Les médianes observées varient de N=48 pour des différences subtiles (ES=1sd) à N=12 pour des différences plus nettes (ES=2sd). Les 3^{e} quartiles varient de N=62 (ES=1sd) à N=16 (ES=2sd). Ils sont inférieurs à 40 dès que l'effet dépasse 1.25 écart-type.

Figure 2 : Taille optimale des groupes



Taille désirée des groupes (N) selon la taille de l'effet à observer (Effect Size, ES). L'effet est exprimé en fonction de l'écart type observé (sd) de la population.

L'intégration de risques corrigés par la méthode de Bonferroni permet de prendre en compte la multiplicité des classes de façon très conservative. Les calculs effectués pour chaque gène sont ainsi valables pour une possible stratification de 9 classes d'individus. La détection d'un effet raisonnable, correspondant à une différence de moyenne d'1.25 écart-type observé dans la population, sera possible dans 90% des cas si les groupes comportent 40 individus. Tout effet de plus grande amplitude sera détecté avec moins d'individus par groupes pour les mêmes conditions. Ces observations sont valables pour ¾ des observations possibles et l'augmentation de la taille des groupes sera toujours favorable.

Envisager le regroupement de 1000 individus pour mesurer et tester le pouvoir discriminatif et prédictif des 258 gènes sera cohérent aux vues des résultats précédents. Dans le meilleur des cas chaque groupe aura le même nombre d'individus (i.e. 1000/9 ~ 111 individus). Puisque la prévalence des différentes pathologies est différente, il faudra au mieux essayer d'atteindre 40 individus pour chacun des groupes.

Représentativité des signatures transcriptionnelles de 3 individus :

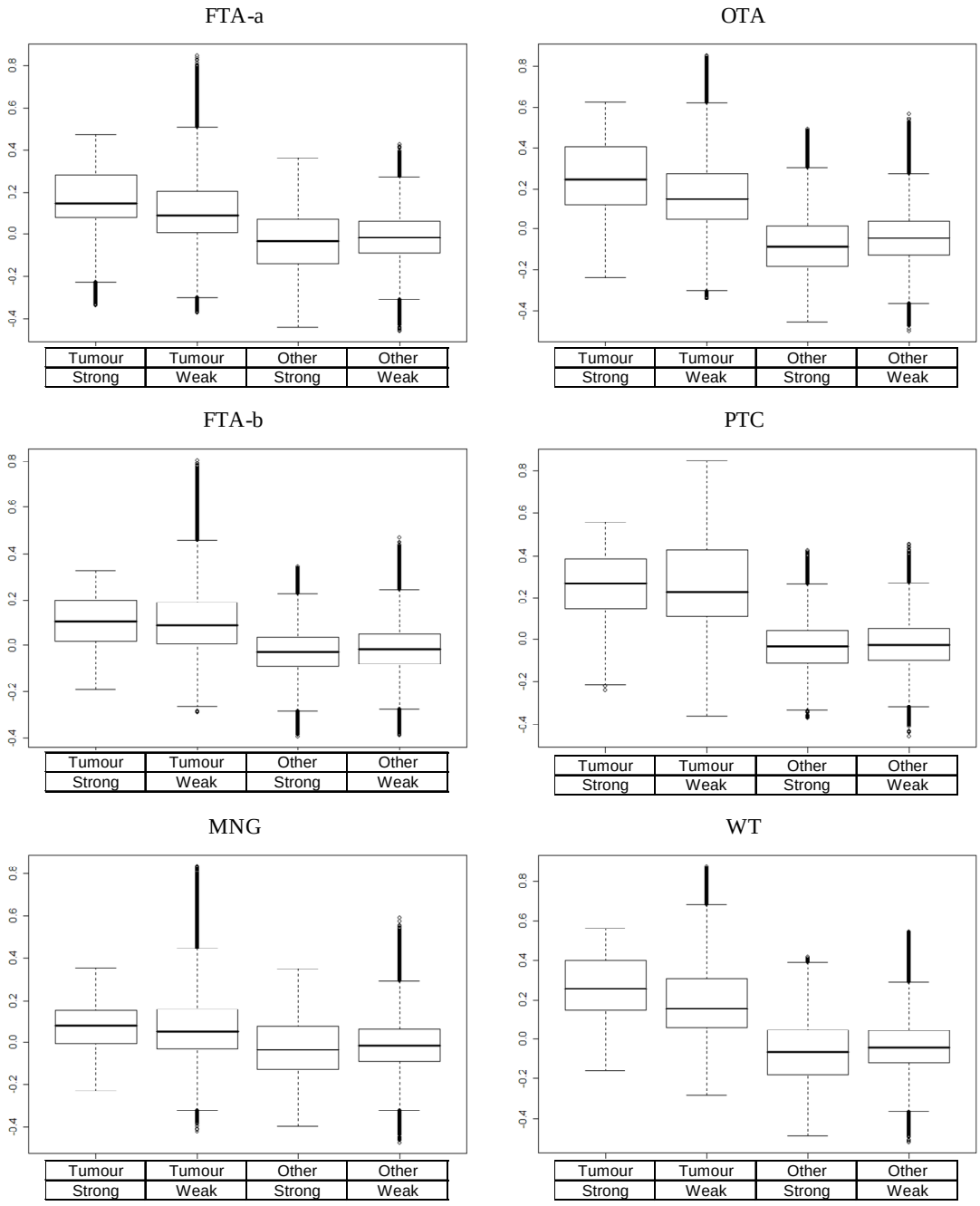
Malgré la connaissance des considérations statistiques théoriques, il est parfois impossible d'inclure le nombre souhaité d'individus dans une étude. Ceci est évident dans le cas de maladies rares, et reste souvent vrai pour des coûts financiers et temporels. Comme vu précédemment, il est possible de définir un nombre minimal théorique d'individus par classe. Ces calculs permettent de choisir la puissance d'un test statistique effectué sur le profil d'un gène. Avoir un nombre suffisant d'individus permettra par exemple de définir individuellement de bons marqueurs pour chaque groupe. Grâce aux biopuces, qui produisent des milliers de données d'expressions par échantillon biologique, une signature définie par peu d'individus est-elle représentative de leur classe ? Dans notre étude, est-il possible d'utiliser 3 FTCs et 4 OTCs dans les analyses ou doit-on les exclure dès le départ ?

Pour étudier la valeur représentative de 3 échantillons sur une classe pathologique, nous avons systématiquement étudié, pour chaque classe de plus de 10 échantillons, la similarité d'une signature définie avec peu d'échantillons (centroid faible, n=3) ou un nombre plus conséquent (centroid fort, n=taille de la classe – 3). La Figure 3 montre la corrélation des échantillons d'une classe donnée et des autres échantillons de l'étude au centroid faible et fort de cette même classe. Les centroids sont les signatures moyennes des classes calculées pour 1000 gènes montrant les plus grands coefficients de variation. Pour chacune des 6 classes testées (FTA-a, -b, MNG, OTA, PTC et WT), le centroid fort et le faible permettent d'identifier les échantillons de la même classe par rapport aux contrôles respectifs. Par exemple, quand la classe FTA-a est définie par un centroid fort (n=23), les corrélations sont supérieures pour les échantillons FTA-a (Tumour/Strong) que pour les autres (Other/Strong). Lorsque la classe FTA-a est définie par un centroid faible (n=3), les corrélations des échantillons de la classe (Tumour/Weak) sont aussi supérieures aux contrôles (Other/Weak). Les corrélations obtenues par les centroids forts sont dans tous les cas plus élevées que celles obtenues par le centroid faible. La corrélation médiane peut ne pas se différencier entre les 2 centroids (Test de Wilcoxon : P_{PTC}=0.8 et P_{FTA-b}=0.098) mais l'espace interquartile est plus élevé avec le centroid faible. Des résultats similaires (Figure 4) sont trouvés dans un jeu de données indépendant (Giordano *et al.*, 2006).

Ces résultats nous montrent que les signatures transcriptionnelles des tissus thyroïdiens permettent l'utilisation de 3 échantillons pour définir une classe. Ces résultats valident l'approche utilisée dans le premier article de cette thèse pour les classes FTC et OTC de 3 et 4 échantillons, respectivement. Même si de l'information est exploitable avec 3 échantillons, un

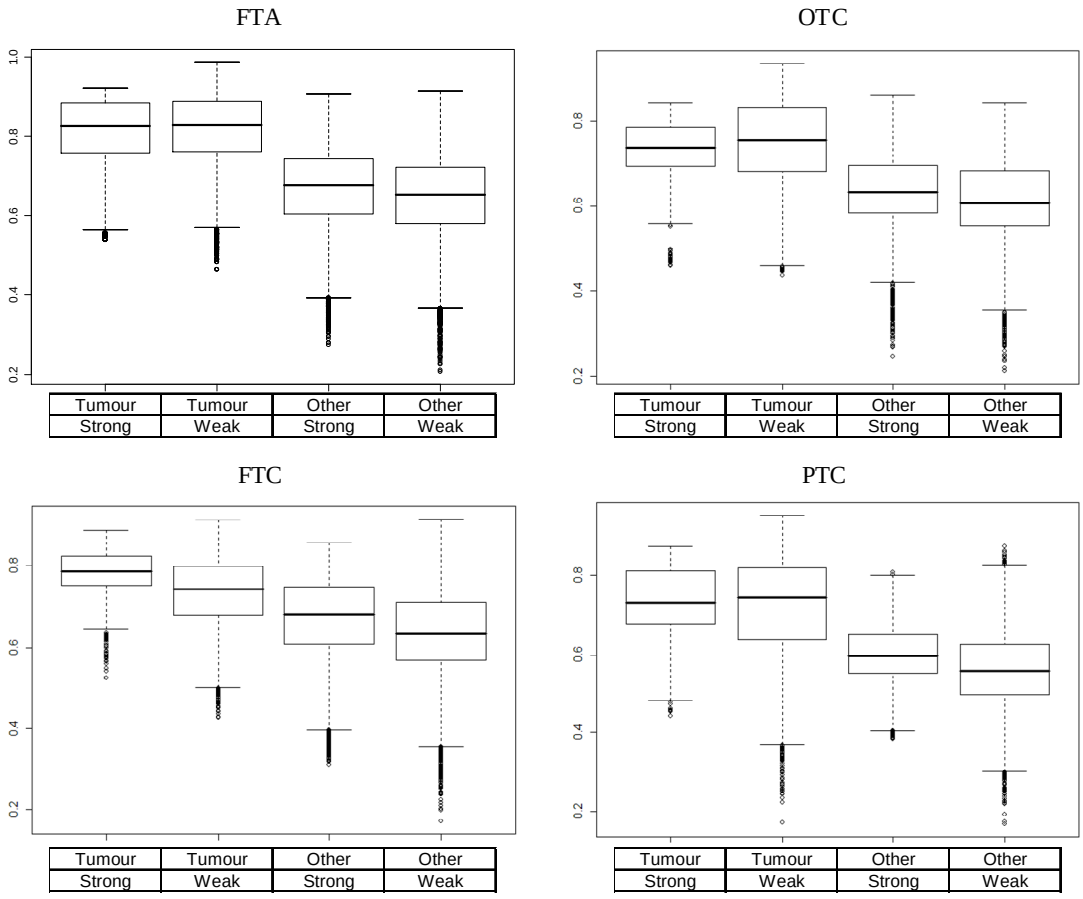
si petit nombre d'échantillons ne sera pas adapté pour une analyse précise et puissante de ces classes (définition de marqueurs, analyse de fonctions ontologiques).

Figure 3 : Information apportée par 3 échantillons



Pour chaque classe de tumeur (Tumour), un centroid faible (Weak) et fort (Strong) ont été calculés, et leur corrélation avec les échantillons de cette classe et les autres (Other) sont représentés par les boites à moustaches. Nous avons testé toutes les combinaisons de 3 éléments parmi ceux de chaque classe avec une validation croisée non chevauchante.

Figure 4 : Information apportée par 3 échantillons dans un jeu de données indépendant



Dans le jeu de données de Giordano et al. (2006) et pour 4 classes de tumeur (Tumour), un centroid faible (Weak) et fort (Strong) ont été calculés, et leur corrélation avec les échantillons de cette classe (nombre d'échantillons : 10 FTAs, 12 FTCs, 12 PTCs et 8 OTCs) et les autres (Other) sont représentés par les boites à moustaches. Nous avons testé toutes les combinaisons de 3 éléments parmi ceux de chaque classe avec une validation croisée non chevauchante.

3.1.4 Discussion

L'étude simultanée de 90% des lésions thyroïdiennes différenciées par puce à ADN nous a permis de définir des signatures spécifiques à chaque classe de tissu malgré une grande hétérogénéité observable dans les classes. Le choix des outils statistiques est crucial pour limiter les faux positifs. Les tests par permutations ont le mieux permis de modéliser le contrôle de l'erreur après prise en compte de l'effet de la multiplicité des tests.

Lorsqu'on analyse les signatures moyennes des classes, ces tissus se regroupent selon leur caractère bénin, malin ou oncocytaire. Nous pouvons ainsi voir que des classes sont similaires et ne devraient pas être opposées dans les études (adénomes macro folliculaires et goitres multi-nodulaires). Les adénomes micro folliculaires se distinguent fortement des adénomes macro folliculaires et des carcinomes. Les tumeurs papillaires sont homogènes et leurs profils concordent avec les nombreuses études d'expression connues. Les tumeurs oncocytaires ne peuvent pas être reliées aux carcinomes folliculaires ou papillaires. Elles peuvent constituer une classe à part entière en considérant que le dérèglement mitochondrial est associé à une prolifération augmentée. La prédominance de la signature métabolique pourrait masquer la similarité de ces tumeurs aux deux précédentes classes.

Les études à grande échelle définissent un cadre pour l'analyse d'un sous ensemble de gènes et de phénotypes (projet soumis à l'Agence Nationale de la Recherche). L'utilisation de technologies différentes est envisageable et permet de croiser les mesures sur différentes plateformes (RT-PCR, Western, Dot Blots). Les méta-analyses de données sont délicates puisque les valeurs mesurées ne sont pas forcement compatibles. Nous avons validé l'expression de 9 marqueurs au niveau protéique sur 40 échantillons biologiques. Une concordance existe entre les données d'expression génique et protéique. Les discordances peuvent s'expliquer par des phénomènes de régulations intervenant à différents niveaux de la transcription à la traduction des gènes, avec notamment l'action des facteurs de transcription et des petits ARN.

Cette étude dévoile une ressemblance moléculaire des T-UM au groupe des carcinomes. Un signal générique relatif aux cancers semble être présent dans ces tumeurs. Nous ne pouvons pas distinguer plus précisément leur affinité envers les carcinomes folliculaires ou papillaires. L'étude suivante, intitulée « Microarray data analysis refines classification of follicular thyroid tumours of uncertain malignancy », réalisée en parallèle à ce travail apporte des informations à ce sujet (voir ci-dessous).

Des signatures spécifiques existent pour chaque type histologique. Un choix restreint de gènes prédicteurs dans ces signatures peut classer automatiquement les tissus avec de bonnes performances. Nous pensons que l'utilisation de puces dédiées pourrait améliorer le diagnostic clinique des tumeurs sur cytoponction de façon à orienter la thèse chirurgicale, plus ou moins extensive, ou médicamenteuse. Nous avons déposé un brevet européen intitulé « Methods for the prognosis or for the diagnosis of a thyroid disease » pour une valorisation et une exploitation en diagnostic médical

3.2 Brevet européen

Jean-Fred Fontaine, Frédérique Savagner, Brigitte Franc, Rémi Houlgatte, Yves Malthièry. *European Patent Office*, EP 07 290 373.5 (déposé le 29 mars 2007)

3.2.1 Titre de l'invention

Methods for the prognosis or for the diagnosis of a thyroid disease

3.2.2 Champ de l'invention

The present invention relates to methods for predicting or diagnosing a specific thyroid disease in an individual, which methods are based on the quantification of the level of expression of thyroid disease-specific marker genes.

3.2.3 Résumé de l'invention

The present invention relates to an integrated in vitro method for predicting or diagnosing a specific thyroid disease in an individual, which method is based on the quantification of the expression level of disease-specific marker genes, selected from the group consisting of marker genes specific for (i) autoimmune thyroiditis (AT), (ii) marker genes specific for Grave's disease (GD), (iii) marker genes specific for macrofollicular adenoma (FTA-a), (iv) marker genes specific for atypical follicular adenoma (FTA-aty), (v) marker genes specific for microfollicular adenoma (FTA-b), (vi) marker genes specific for follicular carcinoma (FTC), (vii) marker genes specific for multinodular goitres (MNG), (viii) marker genes specific for oncocytic adenoma (OTA), (ix) marker genes specific for atypical oncocytic adenoma (OTA-aty), (x) marker genes specific for oncocytic carcinoma (OTC) and (xi) marker genes specific for papillary carcinoma (PTC).

3.3 Article 2

3.3.1 Introduction

Titre: Microarray data analysis refines classification of follicular thyroid tumours of uncertain malignant potential

Les tumeurs de malignité incertaine présentent des caractéristiques morphologiques proches des carcinomes. Il peut s'agir de variants folliculaires de carcinomes papillaires, de carcinomes folliculaires à invasion minime ou d'adénomes aux caractéristiques nucléaires inhabituelles (Vasko *et al*., 2004). Ces tumeurs sont de diagnostic difficile.

Ces tumeurs ont été ignorées des analyses antérieures utilisant les biopuces qui se sont focalisées sur la séparation des tumeurs bien définies. Grâce au jeu de données utilisé dans le premier travail (voir ci-dessus), nous avons étudié ces tumeurs pour affiner leur classification. Nous les avons particulièrement comparés aux carcinomes papillaires, aux carcinomes folliculaires ainsi qu'aux adénomes classiques.

Nous avons analysé le regroupement hiérarchique de tous les gènes pour trouver des profils de gènes co-régulés associés aux T-UMs. L'analyse statistique nous a permis d'établir les gènes différentiels des classes. Nous avons associé un recoupement des listes de gènes spécifiques à l'évaluation de la déplétion ou de l'enrichissement ontologique. Nous avons utilisé des tests permutés pour évaluer les corrélations des échantillons avec les profils des classes de carcinomes. L'étude des mutations caractéristiques des carcinomes nous a servi à valider les résultats. Des analyses immuno-histochimiques ont été réalisées pour étudier certaines hypothèses expliquant les cas divergents.

Cet article a été publié dans le journal scientifique international Oncogene (Nature Publishing Group) le 29 Octobre 2007 (PMID: 17968324)

3.3.2 Article

Title: Microarray analysis refines classification of non medullary thyroid tumours of uncertain malignancy

Authors

Jean-Fred Fontaine[1-2], Délphine Mirebeau-Prunier[1-3], Brigitte Franc [4], Stéphane Triau[5], Patrice Rodien[6], Rémi Houlgatte[7-8], Yves Malthièry[1-3], Frédérique Savagner[1-3].

1: INSERM, U694, Angers, F-49033 France

2: Université d'Angers, Faculté de Médecine, Angers, F-49033 France

3: CHU Angers, Laboratoire de Biochimie, Angers, F-49033 France

4: Université Versailles Saint-Quentin en Yvelines, faculté Paris Ile de France Ouest , Service de Pathologie, Hôpital A Paré (APHP), 92104 Boulogne, France

5: CHU Angers, Fédération de Pathologie Cellulaire et Tissulaire, Angers, F-49033 France

6: CHU Angers, Département d'Endocrinologie et de Médecine Interne, Angers, F-49033 France

7: INSERM, U533, Nantes, F-44035 France.

8: Université de Nantes, Faculté de Médecine, Institut du thorax, Nantes, F-44035 France.

Corresponding author

Frédérique Savagner, Inserm U 694, Laboratoire de Biochimie, CHU, 4 rue Larrey, 49033 Angers, France, tel : +33 241 35 33 14, Fax : +33 241 35 40 17, frsavagner@chu-angers.fr.

Running title: Non medullary thyroid tumours of uncertain malignancy

Keywords: thyroid tumours; microarray analysis; uncertain malignancy; classification

Abstract

Conventional histology failed to classify part of non medullary thyroid lesions as either benign or malignant. The group of tumours of uncertain malignancy (T-UM) concerns either atypical follicular adenomas and the recently called "tumours of uncertain malignant potential". To refine this classification we analysed microarray data from 93 follicular thyroid tumours: 10 T-UM, 3 follicular carcinomas, 13 papillary thyroid carcinomas and 67 follicular adenomas, compared to 73 control thyroid tissue samples. The diagnosis potential of 16 selected genes was validated by quantitative RT-PCR on 6 additional T-UM. The gene expression profiles in several groups were examined with reference to the mutational status of the RET/PTC, BRAF and RAS genes. A pathological score (histological and immunohistochemical) was estimate for each of the T-UM involved in the study. The correlation between the T-UM gene profiles and the pathological score allowed a separation of the samples in two groups of benign or malignant tumours. Our analysis confirms the heterogeneity of tumours of uncertain malignancy and highlighted the molecular similarities between some cases and true carcinomas. We demonstrated the ability of few marker genes to serve as diagnosis tools and the need of a T-UM pathological scoring.

Introduction

The World Health Organisation classification for thyroid tumours (DeLellis *et al.,* 2004) still distinguishes two types of non medullary differentiated thyroid cancer: the papillary thyroid carcinoma (PTC), diagnosed on the basis of characteristic nuclear features, and the follicular thyroid carcinoma (FTC), defined by its capsular and/or vascular invasiveness. However, for decades, conventional histology failed to classify some encapsulated follicular thyroid tumours as benign or malignant because these lesions share overlapping histological features. This already known difficulty is underscored by recent substantial interobserver variability (Franc *et al.,* 2003; Hirokawa *et al.,* 2002; Lloyd *et al.,* 2004), either in the pathological assessment of thyroid nodules and in the identification of underlying diagnosis criteria, such as papillary nuclear features, vascular and/or capsular invasion. These difficulties are especially relevant in encapsulated differentiated thyroid carcinomas for their belonging to the encapsulated follicular variant of papillary thyroid carcinoma (FVPTC) or the minimally invasive FTC group, and for their ability to harbour or not predominant oncocytic features. In order to point out to the clinicians such difficulties, several terminologies have been proposed over time (Williams 2000; DeLellis *et al.,* 2004). The term of atypical thyroid follicular adenoma is now restricted to adenomas with pronounced cellular proliferation, less regular cytoarchitectural patterns,

lacking evidence of capsular and or vascular invasion. For tumours showing questionable capsular and/or vascular invasion it is recommended to call them "follicular tumour of uncertain malignant potential" if papillary carcinoma-type nuclear changes are absent, and "well differentiated tumour of uncertain malignant potential" if those nuclear changes are imprecise.

Gene expression profiling by microarray techniques has revealed some tumour sub-classifications (Bertucci *et al.,* 2003; Takahashi *et al.,* 2003). Distinctive gene expression patterns are associated with a wide variety of morphological, biological and clinical parameters. This approach has now begun to be applied to thyroid cancer (Huang *et al*., 2001; Finley *et al.,* 2004; Giordano *et al.*, 2005; Jarzab *et al.,* 2005; Eszlinger *et al.,* 2006). Although some studies have included FVPTCs and minimally invasive FTCs, none has examined the challenging group of tumours of uncertain malignancy (T-UM). Only one microarray study has been carried out on minimally or widely invasive follicular carcinomas (Lubitz *et al.,* 2005). It showed that some minimally invasive carcinomas were subclassified whereas others could be reclassified as either follicular thyroid adenomas (FTAs) or widely invasive carcinomas. These data support the theory of a continuum of tumours from the benign to the malignant. So far, all the microarray studies on thyroid tumours have excluded samples with a non clear cut diagnosis. Since the main objective was to distinguish benign from malignant tumours; the T-UM were ignored. However, given the accuracy of microarray analysis, it may be possible to refine the classification of this group of tumours.

We therefore investigated a group of T-UM using microarrays and taking into account all their underlying counterparts (immune background, cellular atypia, density, mitotic activity, patterns) in comparison with FVPTCs, PTCs, minimally invasive FTCs s and FTAs. The results were expected to lead to an improved classification and to better predict true malignancy.

Results

Comparison of gene expression in PTC subtypes

We examined the gene-expression signatures in two variants of PTC, i.e. 5 usual PTCs and 8 FVPTCs. These two PTC subtypes were compared using the Student’s t-test; the raw P-values were corrected for multiplicity (Benjamini *et al.*, 1995) over the full set of 5,549 expressed genes analysed by the microarrays. The results showed no significant difference (FDR>0.05)

between the two subtypes (data not shown). We therefore treated the PTC tumours as a single class for the rest of the study.

Overexpressed and co-regulated gene clusters in T-UM

The expressed genes were clustered hierarchically in order to discover the correlated genes involved in T-UM. We found five clusters of overexpressed and correlated genes (Figure 1). Underexpressed genes were also present but no cluster was clearly specific to the T-UM samples. Individual sample specificity with respect to the overexpressed clusters is indicated by the mean log-value signatures (bar charts). Only 21 of the 166 tissue samples had a mean log-value greater than 0.6, chosen as an empirical cut-off value. These included 11 (85%) PTCs, 7 (70%) T-UM, one (1.5%) FTA and two (5.3%) oncocytic thyroid (OT) tumours. Some genes known to be markers of the FTC group showed coordinated upregulation with T-UM or PTC, e.g. the ASTN2 and the GALNT3 genes. The five clusters of overexpressed genes identified were not specific to the other tumours (low log-values). This unsupervised method showed that gene regulation occurred in T-UM and PTCs.

Supervised analysis of the gene-expression signatures

We determined the differential genes in each class of thyroid tumour using permutated F-tests followed by post-hoc tests ($P<0.05$). This method allows the simultaneous detection of differential genes for the various classes of tumour. We were thus able to identify the differential genes shared by T-UM and other thyroid tumours (Figure 2). We found 66 upregulated and 92 downregulated differential genes for T-UM. There were 23 (35%) upregulated and 37 (40%) downregulated shared genes with the PTCs and Fisher's exact test indicated that these genes were significantly overrepresented ($P<0.05$). For the other classes of tumour, the number of upregulated shared genes varied from 9 to 12 in FTAs, OTs and wild type (WT) groups, or from 3 to 5 in FTCs, AT (autoimmune thyroiditis)/GD (Grave's disease) tumours. Since the OT, FTA and WT classes contained from 24 to 67 samples whereas the FTC and AT/GD classes contained only 3 to 10 samples, the discrepancy observed is probably explained by a class-size effect. Nevertheless, the OT, WT and AT/GD upregulated shared genes were considered significantly underrepresented ($P<0.05$). This finding allowed us to select specific candidate marker genes for each class, or pair of classes of tumour. Supplementary Table 1 lists carcinoma and adenoma-specific genes in T-UM after filtering of the differential T-UM, FTA and PTC genes from those of the other groups. The supervised

determination of the differential genes showed that T-UM and PTCs had significant differences with other classes.

The similarity of the gene expression of T-UM with the other classes of tumour was also estimated using the Pearson's r correlation coefficient calculated from the mean class signatures as defined by the differential genes, i.e. the class centroids (Table 1). The similarity between the T-UM and the PTC class was greater (r=0.47) than between any of the other classes (P<3.33E-05); the FTC class also showed significant similarity (P=0.003) but non relevant r value (r=0.1). The FTA and WT classes were dissimilar compared to the T-UM class of tumours. The dissimilarity between the T-UM and OT or AT/GD classes was greater than any other within the simulations (P=1), indicating that these classes were significantly dissimilar from the T-UM class. Finally, the two classes of thyroid tumour, T-UM and PTC, had very similar gene-expression signatures. However, we noticed that one PTC presented a divergent signature strongly correlated with the OT class centroids (data not shown). It was the same as in Figure 1 (PTC 152). Two T-UM samples had also negative similarity (T-UM 1 and 3, same ones as in figure 1) and highly correlated with OT or FTA class centroids (data not shown).

Diagnosis potential of the T-UM gene-expression signature

In order to assess the diagnosis potential of the gene-expression signature, we compared the expression signature of a set of 16 predictive genes for each T-UM sample in addition with their pathological score (figure 3). The 16 selected genes were highly predictive of both the PTC and the T-UM group (Supplementary Table 2). The pathological score was based on histological criteria and protein expression analysis (immunohistochemistry) of 4 malignancy candidate markers (HBME-1, Galectin3, CK19 and TPO) as detailed in Supplementary Table 3 (DeLellis *et al*., 2004; De Micco *et al*., 1999). The normalised mean gene-expression signature of the T-UM samples varied from negative (-1.48E-02) to positive values (1.42E-01). The pathological score highly correlated with the similarity values (Spearman rank correlation, ρ=0.71).

The mutational status of the BRAF, RET/PTC and RAS genes was determined for seven PTC, five T-UM and five FTA samples (Table 2). We also computed the similarity of each sample signature to the PTC class (Table 2, Supplementary Table 4). We identified molecular anomalies in five of the seven PTCs; one had a BRAF mutation and four had RET/PTC rearrangements. RET/PTC1 rearrangements were also found in three of the T-UM samples that

had the poorest molecular similarity with PTC. In contrast, the two T-UM samples without molecular anomalies had higher correlation coefficients (R_{PTC}=0.45 and 0.34). No mutations were found in the five FTAs. None of the tumours had overlapping mutations or RAS gene mutations. These findings indicated that some T-UM share high similarities with PTCs and questioned whether or not they should be classified as PTCs.

The diagnosis potential of the gene-expression signatures was validated on a new set of samples by quantitative RT-PCR. This set comprised 6 T-UM, 5 FTAs and 5 PTCs. We measured the expression of 9 genes out of the 16 predictive genes. The hierarchical clustering of the data separated clearly the samples in 2 groups (Figure 4). The overall robustness for these 2 clusters over 1,000 noised datasets was equal to 0.802. The first group contained all the 5 FTAs and 2 T-UMs. The second group contained all the 5 PTCs and 4 T-UMs. The new T-UM samples were also evaluated and scored by a pathologist (Supplementary Table 3). The mean gene-expression signature of the T-UM samples to the PTC group highly correlated to the pathologist's score (ρ=0.57) in this independent experiment (Figure 5).

Immunohistochemical evaluation

Immunostaining results for HBME-1, Galectin-3, CK19 and TPO for the 10 T-UM were summarised in Supplementary Table 3 and associated to the histological features for each sample. The results of the T-UM 1 and 3 favoured a diagnosis of benignancy because of the negative HBME-1, Gal-3, CK19 and the TPO expression well preserved. On the contrary, the T-UM 16 and 188 were the only ones positive for HBME-1 staining, in relation to their high mean expression signature on Figure 3.

We explored the mitochondrial immunostaining for the 3 samples that were divergent on figures 1: the sample 152 (FVPTC), and the two T-UM 1 and 3. Immunostaining data for HBME-1, TPO and mitochondria for these 3 divergent samples are represented on Figure 6. Mitochondrial staining was 5 and 15 % positive for sample 1 and 3 respectively. With the FVPTC sample 152, all the follicular cells took up a medium but homogeneous staining with the mitochondrial antibody.

Discussion

Differentiated thyroid follicular carcinomas are usually diagnosed either on the basis of nuclear features, capsular or vascular invasion. However, there is a group of tumours for which, on the basis of the current diagnostic tools (histological criteria, immunomarkers), it is very difficult to

establish whether these tumours will eventually behave as malignant or benign disease. In this group of tumours of uncertain malignancy (T-UM), diagnostic tools are urgently needed. The aim of this study was to use microarray data to classify T-UM adenomas into one of these three classes of thyroid tumour: benign, malignant versus PTC, malignant versus FTC.

Our study of the molecular profiles of T-UM showed that some were correlated with the profiles the papillary carcinoma classes of thyroid tumour (Table 1). This suggests that some T-UM are probably true carcinomas. Microarray analysis of eight of the ten T-UM showed significant similarities with PTCs In front of the low number of FTC samples, a more exhaustive study will be necessary to conclude on sample similarity results for the FTC group. The five classical and the eight follicular variants of PTC in our study were considered to constitute a homogeneous class of papillary tumours. Though we chose one of the less stringent ways of correcting for multiplicity, the FDR controlling procedure defined by Benjamini and Hochberg (1995), our result must be balanced by the relatively small number of involved PTC samples. Two studies on PTCs and FVPTCs have investigated a large set of FVPTCs (Mazzanti *et al.,* 2004; Giordano *et al.,* 2005). The first study failed to discriminate between PTCs and FVPTCs on the basis of molecular profiling, whereas the second was able to distinguish between PTCs and FVPTCs because of the high frequency of the RAS gene mutation in the FVPTC group. Nonetheless, Finn *et al.* showed differences but strong gene-expression similarity between classical and follicular variant of PTCs (Finn *et al.,* 2007).

Differentially-expressed genes determined by our training set included CITED1, CTNNA1, DPP4 and CDH3, which are well corroborated as belonging to the PTC class of tumours (Huang *et al.*, 2001; Jarzab *et al.*, 2005). The 43% mutation frequency in PTCs was consistent with published reports (Giordano *et al.*, 2005; Nikiforova *et al.*, 2002). We also found a BRAF mutation in a classical PTC sample, as reported in the literature (Nikiforova *et al.,* 2002). We were able to define a PTC cancer signature, independent of the RET/PTC or the BRAF mutational profile of the tumours (Table 2): In regard to the PTC similarity coefficient, we showed that 2 PTC samples with either BRAF (sample 141) or Ret/PTC1 (sample 157) mutation were similar (0.69 and 0.61 respectively), whereas 2 Ret/PTC1 PTC samples (124 and 173) had dissimilar index (0.53 and 0.79 respectively).

Three of the ten T-UM were strongly correlated with PTCs, whereas five were poorly correlated (Supplementary Table 4). Tumours with RET/PTC mutations were found in the strongly correlated as well as the poorly correlated groups (Table 2). Our results may take into account the recent demonstration of Ret/PTC rearrangements in non neoplastic follicular cells

presenting extensive inflammation as in Hashimoto's thyroiditis (Rhoden *et al.*, 2007). However, high correlation between mutational and expression profiles reinforces the notion that the 3 T-UM may be carcinomas. In addition, the definition of a pathological score for each T-UM permitted to detect a good correlation between a high pathological score and a signature significantly similar to PTCs (Figure 3). Therefore we postulated that this score may be used to identify benign from malignant tumours among the T-UM class. We propose such a scoring system before assigning a tumour to the "uncertain malignant potential category".. We showed that the expression profile of 9 selected genes was able to clearly separate the T-UM into 2 groups; benign and malignant (Figure 4) and strongly suggest the belonging of some "at risk of malignancy" T-UM to the PTC class of carcinoma. We suggest that, after a selection by the pathological score, some T-UM should be classified by our set of 9 expression markers into putative PTCs.

Two T-UM (Samples 1 and 3) and one FVPTC (Sample 152) showed no correlation with either FTCs or PTCs but with the OT or the FTA classes. The FVPTC 152 presented several foci of mitochondrial-rich cells that may have modified the gene-expression signature of the tumour. Indeed, since 39 of the 166 thyroid tumours studied were oncocytic tumours, mitochondrial genes were largely represented in our microarray analysis. We have already shown that the homogeneous mitochondrial signature of oncocytic tumours may mask the signature of the putative transformation of a thyroid adenoma into a carcinoma (Baris *et al.*, 2004; Baris *et al.*, 2005). For the 2 T-UM, we verified that the mitochondrial staining was weak while the HBME-1, CK19, Galectin-3 and TPO results favoured a diagnosis of benignancy. More, we controlled that none of already known genes specifically overexpressed in the widely or minimally invasive FTC classes (Lacroix *et al*., 2005; Lubitz *et al*., 2005) correlated with the expression profile of the two lesions (data not shown). The correlation between low score and expression profile of benign tumours was confirmed on the validating set of T-UM for 2 T-UM samples (13C and 14C).

Interestingly, we were able to identify several genes that were specific to the gene-expression signature of T-UM compared to all other benign or malignant signatures (Supplementary Table 1). We identified mitogenic signatures involving the cAMP (TSHR) and the MAPK (GRB10 and NR2F2) pathways, as well as lymphangiogenesis event (VEGF-C), as described in the literature (Durick *et al.*, 1996; Du Villard *et al.*, 2000; More *et al.*, 2003; De la Torre *et al.*, 2006). This may be associated with the proliferative and invasive signature recently identified in the relatively indolent FVPTC (Finn *et al.*, 2007) ressembling to our T-UM. The

overexpression of TSHR suggests a less aggressive phenotype (Mirebeau *et al.*, 2004; Hoffman *et al.*, 2006). This results confirm that T-UM behave to a special tumour group .

Conclusion

The analysis of microarray data allowed us to characterise the functional relationships between T-UM and PTCs. The mutational and our defined pathological status of the tissues allowed us to classify some T-UM either as benign or as putative PTCs. Here we show that the molecular signature of the tumour can help to refine the classification of T-UM and, in particular, allow to the identification of tumours at risk for malignancy.

Materials and Methods

Tissue samples

The microarray analysis concern 93 samples of thyroid tumours belonging to four classes – PTC, FTC , FTA, T-UM – in accordance with the WHO 2004 (DeLellis *et al.*, 2004). The samples included 13 PTCs (five classical and eight follicular variants), three FTCs, 67 FTAs and 10 T-UM. Histological features of the 10 T-UM were summarised into table 3. The FTA group included macro- and micro-follicular adenomas originating from single nodules, and adenomas originating from the largest nodule of multinodular goitres. Three groups, comprising a total of 73 thyroid tissue samples served as controls. The WT group contained 24 normal samples; the AT/GD group (10 samples) included samples of autoimmune thyroiditis and Grave's disease; and the OT group (39 samples) included oncocytic adenomas and carcinomas. All thyroid tissue samples, obtained from the Ambroise Paré Hospital (Paris, France) and the Angers University Hospital (France), were rendered anonymous before beginning the study.

Microarrays

RNA extraction, cDNA preparation and hybridization, scanning and image analysis of the arrays were done according to protocols of the manufacturer (http://tagc.univ-mrs.fr/plateforme/protocoles/) and as previously described (Baris *et al.*, 2004), as was probe set intensity estimation and normalization. Each microarray contained 9,216 spotted probes including controls and 8,862 cDNA human clones. Genes with an expression similar to the background and genes with missing values over an entire sample class were withdrawn from the analysis. Thus 5,549 genes were expressed and included to the analysis. All data were subjected to print-tip Lowess normalisation (Yang *et al.*, 2002). The microarray data have been

deposited in NCBIs Gene Expression Omnibus (Edgar *et al.*, 2002) and are accessible through GEO Series number GSE6339.

Data analysis and statistical methods

Statistical tests were done with R Bioconductor statistical software (Gentleman *et al.*, 2004). PTC subtypes were compared by Student's t-tests and multiplicity adjustments using FDR controlling procedure (Benjamini and Hochberg, 1995).

Hierarchical clustering of the data were computed on log-transformed, median gene-centred and normalised data using average linkage and uncentred correlation distances. Computations and visualisation were done with Cluster and TreeView software (Eisen *et al.*, 1998), and with BRB ArrayTools v3.6.0b2 developed by Dr. Richard Simon and Amy Peng Lam. The biological activity of genes with known functions was determined using the web tool FatiGO with the Gene Ontology database (Al-Shahrour et *al.*, 2004).

Differential genes were defined for the 7 classes of tissue by permutated non-parametric two-sided F-tests (FDR=5% for 1000 permutations). Multiple-test correction on raw P-values was done by the single-step maxT procedure (Ge *et al.*, 2003). Up and downregulated genes were identified by post-hoc two-sided t-tests comparing one class to the others ($p<0.05$, unadjusted).

In the similarity analysis, we defined a centroid as the signature of its mean gene-expression profiles. Similarity was defined as the correlation coefficient of two centroids, or of a sample signature and a centroid. P-values were generated from 30,000 bootstrapped data sets.

The list of 16 predictive genes was generated by using the 6 prediction methods implemented in BRB ArrayTools (Compound covariate predictor, Bayesian compound covariate, Diagonal linear discriminant analysis, K nearest neighbors (for K=1 and 3), Nearest centroid, and Support vector machines) and a cutoff equal to 1E-08 for univariate feature selection. The 0.632+ cross validation method was used to compute mis-classification rate.

Detection of mutations and rearrangements

DNA and RNA isolation and cDNA synthesis

DNA and RNA were isolated using the guanidium isothiocyanate procedure (Trizol Reagent, Invitrogen Life Technologies, Gaithersburg, MD, USA). Quantification, degradation and DNA contamination of RNA were assessed using an RNA 6000 Nano Assay kit (Agilent

Technologies, Palo Alto, CA, USA). Reverse transcription was performed on 1µg of RNA with Advantage RT-for-PCR kit (Clontech Laboratory, Palo Alto, CA, USA) following the manufacturer's recommendations.

DNA and cDNA sequencing

The mutational status for BRAF, RET/PTC1, RET/PTC3 and RAS genes was determined for seven PTCs, five T-UM and five macrofollicular FTAs. The PCR reactions were performed on 5µl cDNA to look for RET/PTC rearrangements, or 5µl DNA for BRAF and RAS mutations using the HotGoldstar DNA polymerase according to the manufacturer's recommendations (Eurogentec, Seraing, Belgium). The primers used to amplify Ret/PTC1 were 5′-AGA-TAG-AGC-TGG-AGA-CCT-AC-3′ and 5′-TGC-AGG-CCC-CAT-ACA-ATT-TG-3′, and those used to amplify Ret/PTC3 were 5′-AGA-TAG-AGC-TGG-AGA-CCT-AC-3′ and 5′-CAT-GCC-AGA-GCA-GAA-GTC-A-3′. The primers flanking exon 15 of the BRAF gene were: 5′-TCC-TTT-ACT-TAC-ACC-TCA-G-3′ and CAT-CTC-AGG-GCC-AAA-AAT-3′. Exon 2 of the N-Ras gene was amplified using primer sequences described elsewhere (Di Cristofaro *et al.*, 2006).

Amplified fragments were purified and directly sequenced on a CEQ 8000 apparatus, using a CEQ DTCS Quick-start kit (Beckman Coulter, Fullerton, CA, USA) following the manufacturer's instructions.

Histological samples, Immunohistochemitry

All the histological diagnosis corresponding to the studied cases included in the study were reviewed (BF, ST).

T-UM tissues samples

The T-UM cases included in the study measured from 1,8 cm to 5 cm in largest diameter. According to the size of the tumour, 5 to 12 samples comprising the tumour /normal tissue interface were analysed (Kononen *et al*., 1998). In order to precisely determine whether or not the tumours classified T-UM in the study fit with the literature definitions (DeLellis *et al*., 2004) the following criteria -presence or absence of questionable PTC nuclear features, -questionable capsular and/or vascular invasion, -worrisome cellular and cyto architectural features were analysed, and quoted 0 if absent, 1 if suspicious, 2 if obvious. In addition was analysed by immuno histochemistry for each of the T-UM cases the expression of 4 markers

known as candidate markers for malignancy : HBME-1, Galectin-3, Cytokeratin 19, and quoted 0 if absent, 1 if express in less than 50% of tumor cells, 2 if more than 50% of the cells were positively stained and TPO quoted 0 if expressed in 80% or more of the tumour cells, 1 is expressed between 50 and <80% , and 2 when expressed < 50% (De Micco *et al*., 1999). All the datas are presented in the Supplementary Table 2, in addition with an individual score obtained by summing in each case the histological and immunohistochemical quoted variables.

Immunohistochemistry

The T-UM immunohistochemical study was performed on a tissue array prepared from the 24 specimens. Representative areas containing the tumour and the control tissue were selected (BF). Double Triplicate tissue cores with a diameter of 0.6mm were taken from each tumour and corresponding tissue specimen (Beecher Instruments, USA), and arrayed on a recipient paraffin block, using standard procedures (Kononen J, Nat Med 1998, 4,844-847.).Four µm consecutive tissue sections were cut from each arrayed paraffin block and prepared on pathological slides. Sections were deparaffinised in xylene followed by 0.3% hydrogen peroxide in methanol at room temperature for 20 minutes for blocking endogenous peroxidase. After rehydratation, antigen retrieval, immunostaining was performed with the following primary antibodies: anti Galectin-3 (diluted 1/300, from Novocastra, UK), anti HBME-1 mesothelial (diluted 1/50, from Dako, Glostrup, Denmark), anti CK19 (diluted 1/100, Dako, Denmark), TPO MoAb47 (diluted 1/10, Biocytex , France) incubated overnight at +4°C, and a universal streptavitin-biotin-peroxidase kit (LSAB, DAKO). For samples PTC152, T-UM 1 and T-UM 3, an anti 60 kD mitochondrial (clone 113-1) antibody from Biogenex (Dilution 1/300, San Ramon, CA) was used to explore the mitochondrial richness of these divergent tumours. The LSAB kit was used according to the manufacturer's instruction. Peroxydase activity was revealed with a commercial 3,3'diaminobenzidine-H2O2 kit (DAKO). For negative control slides, the primary antibody was either omitted or replaced by a suitable concentration of normal IgG of the same species.

Quantitative RT-PCR analysis

Real-time quantitative PCR was performed on cDNA originating from 16 thyroid tumour samples, independant from the samples used for microarray analysis. We quantified the expression of 9 selected genes on 6 T-UM, 5 FTA and 5 PTC samples, using SYBR Green I dye as fluorescent signal (iQ sybrGreen supermix, Biorad, Hercules, California, USA) and according to manufacturer's recommendations for the Chromo4 apparatus (Biorad). The ten

genes explored were : CDH3, CLDN1, ECM1, CITED1, MRC2, ABCC5, DPP4, ABCC3, and CAPN3. The amount of cDNA for the genes selected was normalized by the quantification of the β-ACTIN. Primers list used for quantitative PCR was supplied in Supplementary Table 3. Each sample was assayed in duplicate. Negative controls were included in the amplification reactions. The specificity of amplification from each primers pairs were attested by plotting the melting curves of products using the Opticon software.

The standard PCR products were generated from plasmids containing the appropriate cDNA insert as template. The sequence specific standard curve was plotted using serial dilutions of the target gene standard PCR product, and the same primers were used to amplify the cDNA Template.

Acknowledgments

We thank Marielle Mello, Dominique Couturier and Anne Coutoleau for technical help and data processing. We thank Kanaya Malkani for the critical reading of this paper. This work was supported by grants from the French Ministry of Research, the French National Institute for Medical Research (INSERM), the University Hospital of Angers, and the University of Angers (PHRC 03-10)

Supplementary Information accompanies the paper on the Oncogene website (http://www.nature.com/onc).

References

- Al-Shahrour F, Diaz-Uriarte R, Dopazo J. (2004). FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. Bioinformatics 20: 578-580.
- Baris O, Mirebeau-Prunier D, Savagner F, Rodien P, Ballester B, Loriod B, et al. (2005). Gene profiling reveals specific oncogenic mechanisms and signaling pathways in oncocytic and papillary thyroid carcinoma. Oncogene 24: 4155-4161.
- Baris O, Savagner F, Nasser V, Loriod B, Granjeaud S, Guyetant S, et al. (2004). Transcriptional profiling reveals coordinated up-regulation of oxidative metabolism genes in thyroid oncocytic tumors. J Clin Endocrinol Metab 89: 994-1005.
- Benjamini Y, Hochberg Y. (1995). Controlling the false discovery rate: a practical and powerfull approach to multiple testing. JRSSB 57: 289–300
- Bertucci F, Viens P, Hingamp P, Nasser V, Houlgatte R, Birnbaum D. (2003). Breast cancer revisited using DNA array-based gene expression profiling. Int J Cancer 103: 565-571.
- De la Torre NG, Buley I, Wass JA, Turner HE. (2006). Angiogenesis and lymphangiogenesis in thyroid proliferative lesions: relationship to type and tumour behaviour. Endocr Relat Cancer 13: 931-944.
- DeLellis R, Lloyd R, et al. (2004). World Health Organization Classification of Tumours: Pathology and Genetics of Tumours of Endocrine Organs: 320.
- De Micco C, Vassko V, Henry JF. (1999). The value of thyroid peroxidase immunohistochemistry for preoperative fine-needle aspiration diagnosis of the follicular variant of papillary thyroid cancer. Surgery 126:1200-4.
- Di Cristofaro J, Marcy M, Vasko V, Sebag F, Fakhry N, Wynford-Thomas D, et al. (2006) Molecular genetic study comparing follicular variant versus classic papillary thyroid carcinomas: association of N-ras mutation in codon 61 with follicular variant. Hum Pathol 37: 824-830.
- Durick K, Wu RY, Gill GN, Taylor SS. (1996). Mitogenic signaling by Ret/ptc2 requires association with enigma via a LIM domain. J Biol Chem 271: 12691-12694.
- Du Villard JA, Wicker R, Crespo P, Russo D, Filetti S, Gutkind JS, et al. (2000). Role of the cAMP and MAPK pathways in the transformation of mouse 3T3 fibroblasts by a TSHR gene constitutively activated by point mutation. Oncogene 19: 4896-4905.
- Edgar R, Domrachev M, Lash, AE. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 30: 207-210.
- Eisen MB, Spellman PT, Brown PO, Botstein D. (1998). Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A 95: 14863-14868.
- Eszlinger M, Wiench M, Jarzab B, Krohn K, Beck M, Lauter J, et al. (2006). Meta- and reanalysis of gene expression profiles of hot and cold thyroid nodules and papillary thyroid carcinoma for gene groups. J Clin Endocrinol Metab 91: 1934-1942.
- Finley DJ, Zhu B, Barden CB, Fahey TJ 3rd. (2004). Discrimination of benign and malignant thyroid nodules by molecular profiling. Ann Surg 240: 425-436.
- Finn SP, Smyth P, Cahill S, Streck C, O'regan EM, Flavin R, et al. (2007). Expression microarray analysis of papillary thyroid carcinoma and benign thyroid tissue: emphasis on the follicular variant and potential markers of malignancy. Virchows Arch 450: 249-260.
- Franc B, de la Salmoniere P, Lange F, Hoang C, Louvel A, de Roquancourt A, et al. (2003). Interobserver and intraobserver reproducibility in the histopathology of follicular thyroid carcinoma. Hum Pathol 34: 1092-1100.

- Ge Y, Dudoit S, Speed TP (2003). Resampling-based multiple testing for microarray data hypothesis. Test 12: 1-44.

- Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S et al. (2004). Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 5: R80.

- Giordano TJ, Kuick R, Thomas DG, Misek DE, Vinco M, Sanders D, et al. (2005). Molecular classification of papillary thyroid carcinoma: distinct BRAF, RAS, and RET/PTC mutation-specific gene expression profiles discovered by DNA microarray analysis. Oncogene 24: 6646-6656.

- Lloyd RV, Erickson LA, Casey MB, Lam KY, Lohse CM, Asa SL et al. (2004). Observer variation in the diagnosis of follicular variant of papillary thyroid Carcinoma. Am J Surg Pathol. 28:1336-40.

- Hirokawa M, Carney JA, Goellner JR, DeLellis RA, Heffess CS, Katoh R et al. (2002). Observer variation of encapsulated follicular lesions of the thyroid gland. Am J Surg Pathol. 26:1508-14.

- Huang Y, Prasad M, Lemon WJ, Hampel H, Wright FA, Kornacker K, et al. (2001). Gene expression in papillary thyroid carcinoma reveals highly consistent profiles. Proc Natl Acad Sci U S A 98: 15044-15049.

- Jarzab B, Wiench M, Fujarewicz K, Simek K, Jarzab M, Oczko-Wojciechowska M, et al. (2005). Gene expression profile of papillary thyroid cancer: sources of variability and diagnostic implications. Cancer Res 65: 1587-1597.

- Kononen J, Bubendorf L, Kallioniemi A, Barlund M, Schraml P, Leighton S, et al. (1998) Tissue microarrays for high-throughput molecular profiling of tumor specimens. Nat Med 4: 844-847.

- Lacroix L, Lazar V, Michiels S, Ripoche H, Dessen P, Talbot M, et al. (2005) Follicular thyroid tumors with the PAX8-PPARgamma1 rearrangement display characteristic genetic alterations. Am J Pathol 167: 223-231.

- Lubitz CC, Gallagher LA, Finley DJ, Zhu B, Fahey TJ, 3rd. (2005). Molecular analysis of minimally invasive follicular carcinomas by gene profiling. Surgery 138: 1042-1048.

- Mazzanti C, Zeiger MA, Costouros NG, Umbricht C, Westra WH, Smith D, et al. (2004). Using gene expression profiling to differentiate benign versus malignant thyroid tumors. Cancer Res 64: 2898-2903.

- Mirebeau-Prunier D, Guyetant S, Rodien P, Franc B, Baris O, Rohmer V, et al. (2004). Decreased expression of thyrotropin receptor gene suggests a high-risk subgroup for oncocytic adenoma. Eur J Endocrinol 150: 269-276.

- More E, Fellner T, Doppelmayr H, Hauser-Kronberger C, Dandachi N, Obrist P, et al. (2003). Activation of the MAP kinase pathway induces chicken ovalbumin upstream promoter-transcription factor II (COUP-TFII) expression in human breast cancer cell lines. J Endocrinol 176: 83-94.

- Nikiforova MN, Biddinger PW, Caudill CM, Kroll TG, Nikiforov YE. (2002). PAX8-PPARgamma rearrangement in thyroid tumors: RT-PCR and immunohistochemical analyses. Am J Surg Pathol 26: 1016-1023.

- Takahashi M, Yang XJ, Sugimura J, Backdahl J, Tretiakova M, Qian CN, et al. (2003). Molecular subclassification of kidney tumors and the discovery of new diagnostic markers. Oncogene 22: 6810-6818.

- Williams ED. (2000). Guest Editorial: Two Proposals Regarding the Terminology of Thyroid Tumors. Int J Surg Pathol 8: 181-183.

- Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, et al. (2002). Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res 30: e15.

Tables et figures

Table 1: Similarity of T-UM with the other tissue classes.

CLASSES	*SIMILARITY*	*P-VALUE*
T-UM- PTC	0.47	3.33E-05
T-UM- FTC	0.1	0.003
T-UM- FTA	0	0.456
T-UM- WT	-0.04	0.825
T-UM- AT/GD	-0.22	1
T-UM- OT	-0.29	1

Table 2: Mutational status of the samples.

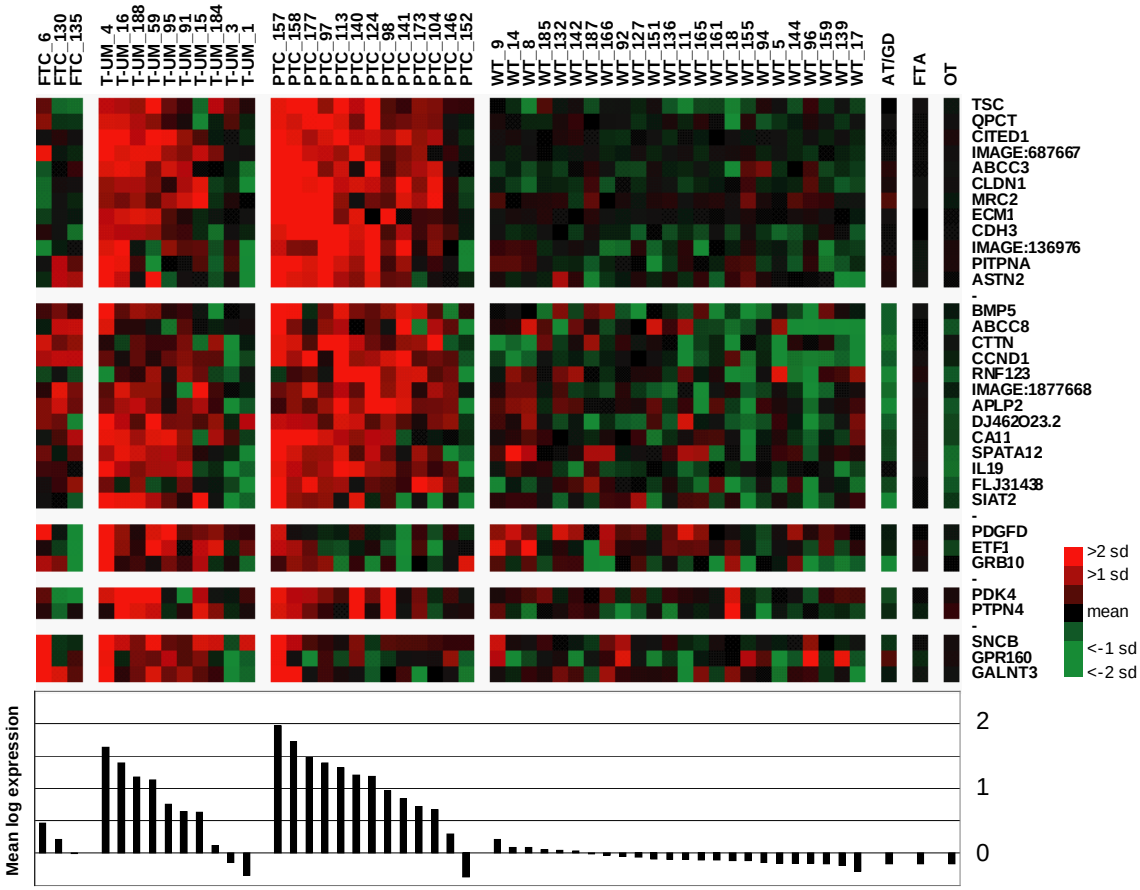
SAMPLE ID	*BRAF*	*RET/PTC1*	*RET/PTC3*	*RAS*	*PTC similarity*	*P-value*
PTC_124	-	+	-	-	0.79	3.33E-05
PTC_158	-	-	+	-	0.78	3.33E-05
PTC_177	-	-	-	-	0.76	3.33E-05
PTC_157	-	+	-	-	0.61	3.33E-05
PTC_141	+	-	-	-	0.59	3.33E-05
PTC_173	-	+	-	-	0.53	3.33E-05
PTC_140	-	-	-	-	0.50	3.35E-05
T-UM_16	-	-	-	-	0.45	3.01E-04
T-UM_4	-	-	-	-	0.34	9.97E-03
T-UM_59	-	+	-	-	0.33	1.35E-02
T-UM_15	-	+	-	-	0.27	3.42E-02
T-UM_95	-	+	-	-	0.18	6.85E-02
FTA_64	-	-	-	-	0.25	4.36E-02
FTA_147	-	-	-	-	0.08	7.78E-02
FTA_149	-	-	-	-	0.06	7.82E-02
FTA_28	-	-	-	-	0.05	7.89E-02
FTA_39	-	-	-	-	-0.13	1

Table 3: Histological characteristics of the 10 T-UM cases.

Sample ID	**Size * (cm)**	**Questionable PTC nuclear features**	**Questionable Capsular and/or vascular invasion**	**Worrisome cellular and cytoarchitectural features**
T-UM_1	3.5	±	-	+
T-UM_3	4.5	+	-	-
T-UM_4	4.5	±	-	+
T-UM_15	2.2	±	+	+
T-UM_16	4.5	+	+	-
T-UM_59	2.7	±	+	+
T-UM_91	1.8	+	-	+
T-UM_95	5	±	+	-
T-UM_184	3	+	-	- (oncocytic tumour)
T-UM_188	2	+	+	+

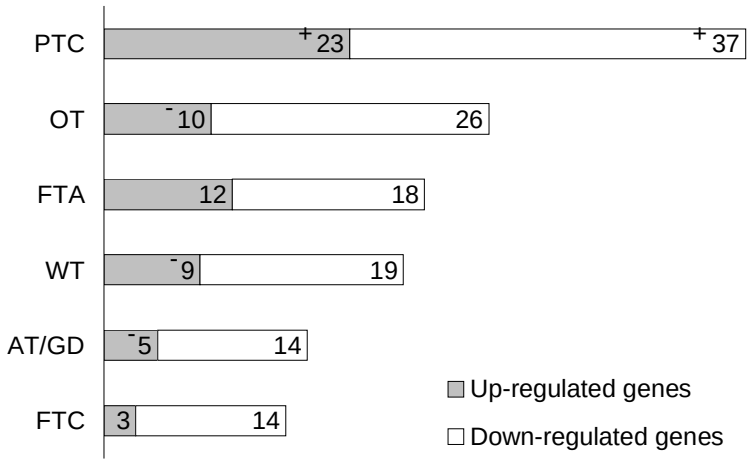
* The tumour sampling comprised 6 to 12 samples per case according to size. (-) absent, (±) suspicious, (+) obvious.

Figure 1: Clusters of correlated and upregulated genes for T-UM samples.



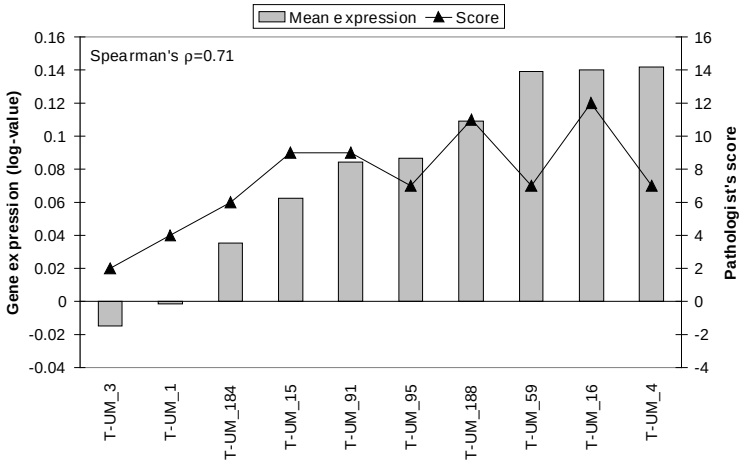
From the global hierarchical clustering of the gene-expression data, we extracted five clusters of correlated and upregulated genes in the T-UM samples. Gene-expression levels are colour-coded in the matrix from green (underexpression) to red (overexpression). Samples (columns) are grouped by class and ordered according to their mean log-level of expression (bar charts at the bottom). Only the mean signature is shown for the AT/GD, FTA and OT groups. AT/GD: autoimmune thyroiditis / Graves's disease, FTA: follicular thyroid adenoma, OT: oncocytic tumour.

Figure 2: Regulated genes of T-UM shared with other classes.



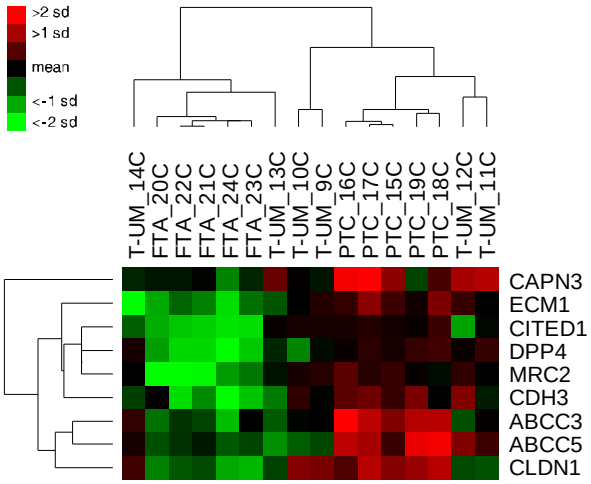
Among the differential genes deduced from the 12 classes of thyroid tissue, the T-UM class had 66 upregulated and 92 downregulated genes. The number of regulated genes shared with other classes is reported for the upregulated genes (gray bars) and the downregulated genes (white bars). Significant overrepresentation (+) and underrepresentation (-) were computed by the Fisher's exact test ($P<0.05$).

Figure 3: Diagnosis potential of predictive genes.



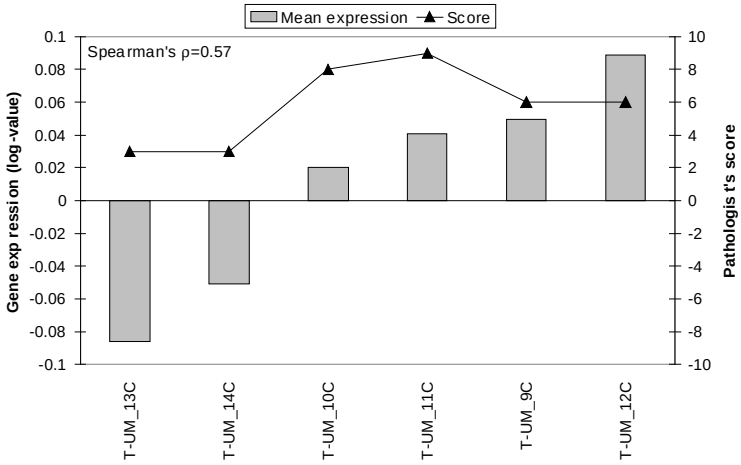
The mean log-level expression (grey bars) of the T-UM samples, computed on 16 marker genes, is compared to the pathologist's score (triangles). The marker genes were selected to classify PTC and T-UM samples simultaneously. The Spearman's rank correlation coefficient ρ between the 2 series of values is equal to 0.71.

Figure 4: Hierarchical clustering of an independent set of T-UM samples.



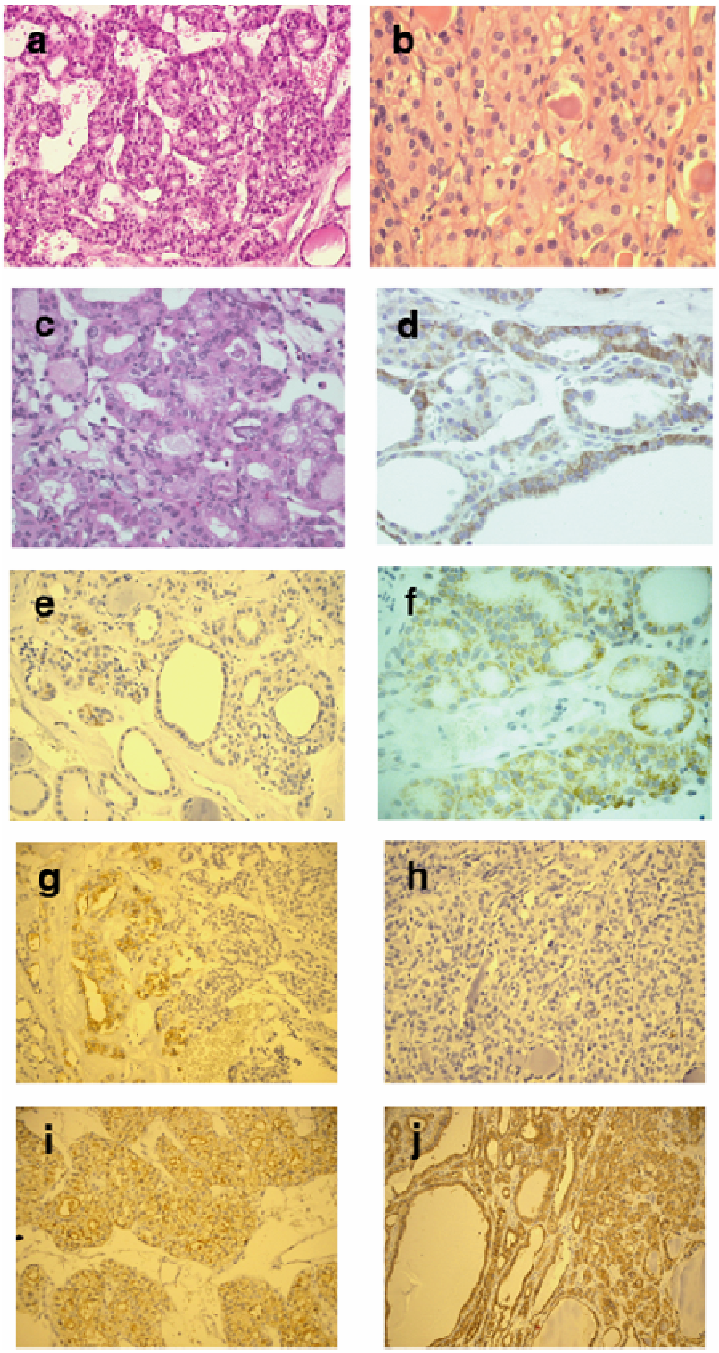
The gene-expression levels of 14 independent T-UM samples and 9 marker genes were measured by RT-PCR. The normalised and median-centred gene-expression levels were hierarchically clustered. Gene-expression levels are colour-coded in the matrix from green (underexpression) to red (overexpression). The separation of the tree sample in 2 clusters is supported by a robustness coefficient of 0.802.

Figure 5: validation of the diagnosis potential.



The mean log-level expression (grey bars) of the 6 independent T-UM samples, computed on 9 selected marker genes, is compared to the pathologist's score (triangles). The genes were selected from the list of 16 marker genes. The Spearman's rank correlation coefficient ρ between the 2 series of values is equal to 0.57.

Figure 6: Immunoperoxidase staining of thyroid tissue sections.



Nuclear features and immunohistochemistry of the two atypical adenomas (Samples 1 and 3) and the follicular variant of PTC (Sample 152) in which the cancer signatures were not correlated with follicular or papillary carcinoma. Original magnification X 40. Hematoxylin-eosin (a, Sample 1; b, Sample 3, c, Sample 152). Immunoperoxidase staining for the respiratory chain complex IV subunit (d, Sample 1; e, Sample 3, f, Sample 152), HBME-1 (g, Sample 1; h, Sample 3) and TPO (i, Sample 1; j, Sample 3).

Supplementary table 1: Specific up- and down-regulated genes in T-UM, FTA and PTC. P-values show the specific expression of each gene in T-UM, FTA and PTC classes. The 'regulation' column indicates the up- (+) or down-regulation (-) of the gene in a class.

Up-regulated genes and specific

FTA up-regulated and specific genes	FTA		T-UM		PTC	
Symbol	*P-value*	*regulation*	*P-value*	*regulation*	*P-value*	*regulation*
PRUNE	9.31E-07	+	8.09E-01		2.17E-02	-
GATA3	7.31E-07	+	4.06E-01		1.64E-01	
IMAGE:729542	1.77E-02	+	2.60E-01		1.50E-01	
DCI	1.35E-04	+	5.12E-01		2.46E-01	
EPHB6	3.74E-04	+	3.35E-01		1.73E-01	
FCGBP	1.58E-05	+	3.63E-01		4.33E-02	-
MBNL1	3.15E-03	+	5.72E-01		9.86E-01	
PPARD	1.16E-02	+	3.46E-01		4.34E-01	
NT5C2	3.69E-03	+	2.17E-01		2.23E-01	
SEMA4B	3.49E-03	+	8.94E-01		9.99E-02	
UBE1L	7.00E-03	+	5.60E-01		6.02E-01	
ASGR1	2.02E-05	+	2.95E-01		1.19E-03	-
IMAGE:365913	4.53E-03	+	6.86E-01		1.65E-03	-
IMAGE:666007	6.94E-04	+	7.73E-01		4.67E-03	-
CDH16	7.26E-03	+	1.31E-01		1.80E-03	-
RAD51L1	2.65E-05	+	1.79E-01		1.57E-02	-
IMAGE:668684	2.33E-03	+	8.93E-01		8.19E-01	
GALT	2.66E-03	+	6.27E-01		5.96E-01	
IMAGE:744055	1.46E-04	+	9.35E-01		7.13E-03	-
KIAA0789	8.88E-03	+	2.69E-01		7.29E-02	
KRT34	1.60E-05	+	7.74E-01		4.99E-02	-
RANBP10	1.21E-03	+	6.12E-01		2.81E-02	-
LOC253842	4.42E-03	+	6.10E-01		2.58E-01	
DDX56	1.38E-03	+	8.38E-01		2.70E-01	
IMAGE:741954	7.74E-04	+	2.01E-01		6.55E-01	
GM632	1.27E-04	+	4.33E-01		7.65E-02	
DAP	6.29E-03	+	3.76E-01		6.59E-01	
ZNF76	2.82E-04	+	4.10E-01		1.32E-01	
C9orf119	1.84E-02	+	2.91E-01		3.24E-01	
CCNE1	4.90E-04	+	8.88E-01		5.43E-02	
CDH1	1.92E-04	+	7.65E-01		3.83E-01	
HDAC9	6.76E-03	+	8.92E-01		7.79E-01	
SETD5	3.17E-02	+	5.46E-01		1.83E-01	
FBXW8	1.74E-02	+	8.39E-01		4.35E-02	-
IMAGE:725836	1.76E-02	+	4.99E-01		6.00E-01	
FGFRL1	1.07E-02	+	1.11E-01		8.79E-01	
PAX8	1.61E-03	+	2.62E-01		9.73E-01	
IMAGE:687852	3.39E-02	+	5.15E-01		9.22E-01	
NDUFA7	1.82E-02	+	9.99E-01		8.32E-01	
IL11RA	4.47E-03	+	7.53E-01		2.34E-01	
LOC400451	7.75E-03	+	7.79E-02		6.45E-01	
GLG1	3.43E-04	+	4.90E-01		2.15E-01	
CRLF3	2.02E-02	+	5.07E-01		2.09E-01	
GCLC	3.96E-02	+	2.23E-01		8.23E-01	
NEXN	3.57E-02	+	4.97E-01		8.13E-01	
IMAGE:744385	2.18E-02	+	2.65E-01		2.29E-02	-
IMAGE:738945	3.14E-03	+	6.49E-01		2.04E-02	-
LOC146909	1.11E-02	+	1.10E-01		9.86E-02	
BAT3	1.65E-03	+	5.93E-01		6.37E-01	
CTBP2	3.87E-02	+	1.07E-01		7.00E-01	
CABIN1	8.04E-05	+	5.38E-01		4.35E-02	-
MAPK15	1.41E-02	+	6.38E-01		3.37E-01	
MT1A	2.61E-05	+	3.87E-01		3.08E-02	-
EIF4G1	7.73E-04	+	4.79E-01		4.98E-01	
TSPAN5	5.02E-04	+	3.82E-01		2.95E-03	-
TOMM40	1.12E-02	+	1.15E-01		9.38E-01	
ARHGAP21	7.20E-03	+	1.91E-01		1.54E-01	
IMAGE:744657	3.90E-03	+	9.29E-01		4.72E-02	-

SGCG	2.38E-02	+	4.02E-01		1.12E-02	-
IMAGE:176543	6.48E-03	+	7.05E-01		5.02E-01	
E2F5	1.88E-03	+	8.55E-02		4.22E-02	-
AMPD2	9.90E-03	+	4.95E-01		2.30E-01	
IMAGE:472111	4.77E-03	+	2.18E-02	-	5.91E-02	
SH3RF2	4.61E-02	+	5.14E-01		2.92E-02	-
ID1	7.38E-04	+	1.43E-01		4.37E-01	
RBM25	4.85E-03	+	2.76E-01		6.21E-02	

FTA-aty	FTA		T-UM		PTC	
Symbol	*P-value*	*regulation*	*P-value*	*regulation*	*P-value*	*regulation*
TMEM125	1.22E-01		1.30E-02	+	3.62E-01	
IMAGE:146976	1.74E-01		4.57E-03	+	5.77E-01	
IMAGE:669136	6.85E-01		2.68E-03	+	1.21E-01	
NR2F2	2.61E-01		2.38E-02	+	5.50E-01	
ZMYM5	5.83E-02		1.22E-02	+	4.37E-01	
GRB10	2.31E-01		2.59E-02	+	5.89E-01	
TSHR	1.03E-01		4.96E-02	+	1.99E-01	
VEGFC	2.48E-01		5.72E-03	+	6.34E-02	
IMAGE:724276	2.91E-01		3.04E-02	+	4.95E-01	
IMAGE:666671	8.49E-01		1.47E-03	+	6.97E-01	
IMAGE:364510	1.27E-01		1.86E-02	+	6.35E-02	
CCL1	5.97E-01		2.98E-02	+	2.13E-01	
AQP4	1.53E-01		1.99E-02	+	8.56E-01	

PTC	FTA		T-UM		PTC	
Symbol	*P-value*	*regulation*	*P-value*	*regulation*	*P-value*	*regulation*
FUS	6.31E-02		6.64E-02		1.32E-02	+
IMAGE:738332	4.24E-01		8.47E-02		4.52E-03	+
ADRB3	5.62E-01		9.19E-02		1.01E-04	+
DPP4	2.76E-02	-	5.72E-01		1.48E-04	+
ARMCX3	9.65E-01		1.04E-01		2.84E-02	+
CLDN1	1.34E-02	-	1.41E-01		6.21E-05	+
LOC389906	5.55E-01		8.21E-01		1.99E-02	+
ABCC3	3.20E-01		5.55E-02		5.02E-04	+
BTBD14B	1.77E-01		5.41E-02		3.35E-02	+
ZNF217	3.28E-03	-	1.50E-01		1.37E-04	+
NOTCH1	1.60E-02	-	3.95E-01		5.59E-03	+
CDH3	1.21E-01		1.71E-01		1.98E-04	+
ETV4	4.98E-01		9.61E-01		8.93E-04	+
CTTN	6.75E-01		2.87E-01		1.46E-04	+
SLC22A18	6.59E-01		1.19E-01		3.76E-02	+
CCNE1	1.74E-01		3.49E-01		1.49E-04	+
RBBP4	2.88E-03	-	1.84E-01		5.01E-03	+
ECM1	2.15E-03	-	8.74E-02		5.03E-03	+
FLJ31438	2.53E-01		8.16E-01		1.09E-04	+
PET112L	4.08E-01		1.83E-01		7.96E-04	+
IMAGE:366159	9.43E-02		4.90E-01		1.03E-02	+
KRT18	3.10E-01		7.33E-02		2.82E-03	+
SHE	2.12E-01		9.12E-01		2.47E-03	+
SCEL	4.69E-02	-	9.57E-02		2.64E-03	+
IL19	2.21E-01		1.45E-01		8.99E-03	+
PITPNA	2.61E-03	-	8.56E-01		3.86E-03	+
IMAGE:24541	2.00E-01		2.57E-01		9.83E-04	+
COX15	8.87E-01		8.81E-01		1.63E-02	+
FMO5	9.37E-01		4.77E-01		3.37E-05	+
IMAGE:731298	4.59E-01		1.57E-01		1.56E-02	+
S100A6	5.49E-03	-	3.98E-01		3.24E-03	+
TMEM45B	3.74E-01		3.67E-01		5.36E-03	+
IMAGE:327182	9.40E-04	-	4.61E-01		6.83E-03	+
WHSC2	5.97E-01		3.58E-01		1.57E-02	+
SPP1	9.34E-01		1.98E-01		8.12E-03	+
VCL	5.44E-01		2.14E-01		9.91E-05	+
VTI1A	4.24E-01		5.65E-01		2.68E-02	+
HSPB7	2.32E-01		5.89E-01		1.74E-03	+
TEX13A	1.59E-01		3.95E-01		1.48E-02	+

FTA-aty/FTA	FTA		T-UM		PTC	
Symbol	*P-value*	*regulation*	*P-value*	*regulation*	*P-value*	*regulation*
PBXIP1	1.36E-02	+	2.11E-02	+	1.28E-01	
ARMCX4	4.13E-02	+	2.61E-02	+	4.43E-01	
LOC348761	8.65E-03	+	3.07E-02	+	6.91E-03	-
MESDC2	2.35E-03	+	2.66E-02	+	6.67E-01	
ARIH2	5.58E-04	+	3.18E-02	+	2.88E-01	
PTK2	3.26E-02	+	2.16E-02	+	3.65E-01	
IMAGE:744505	1.35E-02	+	3.74E-02	+	9.36E-01	
BCL2L11	1.10E-02	+	1.00E-02	+	1.55E-01	
OAZ2	6.82E-03	+	6.07E-03	+	4.53E-01	
FBXL16	3.89E-02	+	4.89E-02	+	7.80E-02	
IMAGE:731685	1.26E-02	+	3.08E-03	+	2.04E-01	

FTA-aty/PTC	FTA		T-UM		PTC	
Symbol	*P-value*	*regulation*	*P-value*	*regulation*	*P-value*	*regulation*
IGF2BP2	7.43E-01		4.57E-02	+	1.03E-04	+
NPAL3	1.92E-01		8.23E-04	+	6.86E-05	+
PDCD4	2.78E-01		2.89E-02	+	2.29E-03	+
TESC	7.97E-01		1.19E-02	+	7.06E-05	+
CAPN3	3.22E-01		4.39E-02	+	5.59E-04	+
IMAGE:687667	4.61E-04	-	1.36E-03	+	9.70E-04	+
PSAT1	9.16E-02		6.48E-03	+	1.23E-02	+
CITED1	7.52E-03	-	2.79E-02	+	9.57E-05	+
CD63	8.96E-01		4.44E-02	+	1.21E-03	+
CD4	8.58E-02		1.21E-02	+	1.89E-02	+
SNRPG	4.99E-01		7.77E-03	+	2.45E-02	+
KCNK5	9.34E-01		1.40E-02	+	4.49E-03	+
CA11	4.55E-01		1.13E-02	+	9.41E-03	+
CTNNA1	9.09E-01		4.89E-03	+	4.03E-03	+
SNCB	8.16E-02		1.77E-03	+	2.14E-02	+

FTA/PTC	FTA		T-UM		PTC	
Symbol	*P-value*	*regulation*	*P-value*	*regulation*	*P-value*	*regulation*
PDLIM1	4.26E-03	+	1.69E-01		5.61E-04	+
RPS6KA2	1.83E-02	+	6.19E-01		9.88E-04	+
LOC147650	3.81E-02	+	8.29E-02		1.07E-02	+
PLXNB2	4.95E-04	+	4.36E-01		3.09E-02	+
LGALS3BP	5.52E-03	+	2.66E-01		4.56E-04	+
PRKG1	1.53E-02	+	1.65E-01		4.10E-02	+

Down-regulated genes and specific

FTA	FTA		T-UM		PTC	
Symbol	*P-value*	*regulation*	*P-value*	*regulation*	*P-value*	*regulation*
C1QBP	1.71E-03	-	4.71E-01		1.36E-01	
IFI30	1.65E-02	-	7.15E-02		3.75E-01	
MLLT1	8.13E-03	-	6.91E-01		2.66E-02	+
VDAC1	9.84E-05	-	7.46E-01		4.17E-01	
IMAGE:666061	2.10E-02	-	8.06E-01		5.65E-01	
KIAA0427	2.09E-05	-	3.36E-01		8.72E-01	
MDH1	3.18E-02	-	7.40E-01		8.78E-02	
ZNF217	3.28E-03	-	1.50E-01		1.37E-04	+
CLGN	3.66E-04	-	8.55E-01		4.93E-01	
SENP6	1.27E-03	-	9.70E-01		2.07E-04	+
ZMYM6	4.53E-03	-	6.69E-01		1.38E-01	
RGS19	3.87E-02	-	1.35E-01		2.22E-01	
FAM117A	1.16E-02	-	7.11E-01		2.42E-01	
GPI	7.32E-03	-	8.29E-02		1.10E-01	
IMAGE:2113771	4.41E-03	-	3.56E-01		4.56E-04	+
S100A12	7.36E-03	-	7.56E-01		5.83E-01	
MAP2	2.01E-02	-	9.11E-02		6.29E-02	
ELK4	4.09E-04	-	7.63E-02		8.32E-01	
COX7B	1.67E-03	-	4.74E-01		9.43E-01	
IMAGE:731357	4.29E-02	-	7.78E-01		5.85E-01	
KLRK1	6.03E-03	-	2.58E-01		2.05E-01	
KPNA2	1.35E-02	-	3.11E-02	+	5.53E-03	+
C10orf6	8.26E-03	-	6.94E-01		5.89E-01	

KLF4	1.69E-03	-	2.84E-01		1.94E-02	+
TTC7B	4.79E-03	-	7.97E-02		4.23E-01	
GLRX2	1.97E-02	-	9.38E-01		1.70E-01	
MAN1A1	8.56E-04	-	9.44E-01		5.07E-01	
IMAGE:742919	2.26E-02	-	6.08E-01		1.18E-01	
WAS	6.81E-05	-	8.18E-01		9.39E-02	
CD47	3.50E-03	-	3.14E-01		3.78E-03	+
TMEM14B	2.01E-02	-	2.38E-01		6.16E-02	
CASP4	4.12E-02	-	8.95E-02		5.71E-01	
MYC	9.38E-04	-	5.90E-01		5.55E-01	
SPDEF	9.11E-04	-	3.66E-01		1.83E-01	
RPS6KA5	1.37E-03	-	4.28E-01		1.15E-01	
EPRS	1.63E-02	-	4.78E-01		8.04E-01	
FABP5	3.86E-02	-	9.54E-02		8.06E-01	
FAM107A	3.34E-02	-	9.70E-02		1.26E-03	+
S100A6	5.49E-03	-	3.98E-01		3.24E-03	+
CLDN7	4.43E-02	-	3.47E-02	+	1.04E-01	
IMAGE:327182	9.40E-04	-	4.61E-01		6.83E-03	+
BNIP3	1.76E-03	-	5.13E-01		5.75E-01	
PITPNA	2.61E-03	-	8.56E-01		3.86E-03	+
IMAGE:382423	1.18E-02	-	4.65E-02	+	8.68E-02	

FTA-aty	FTA		T-UM		PTC	
Symbol	*P-value*	*regulation*	*P-value*	*regulation*	*P-value*	*regulation*
IFITM3	3.06E-01		4.01E-02	-	5.66E-01	
SLPI	6.43E-02		1.35E-02	-	1.02E-03	+
ITGA4	8.97E-02		6.64E-03	-	9.43E-01	
IMAGE:743615	4.57E-02	+	4.29E-03	-	4.53E-01	
IMAGE:472111	4.77E-03	+	2.18E-02	-	5.91E-02	
TRAF5	1.74E-01		5.60E-03	-	5.87E-01	

PTC	FTA		T-UM		PTC	
Symbol	*P-value*	*regulation*	*P-value*	*regulation*	*P-value*	*regulation*
MATN2	2.92E-01		1.40E-01		1.81E-04	-
CCL28	2.11E-01		4.25E-01		5.75E-05	-
IRAK1	1.01E-01		5.74E-01		5.23E-05	-
CDH16	7.26E-03	+	1.31E-01		1.80E-03	-
FLJ11184	6.79E-01		1.46E-01		5.07E-04	-
FLT4	1.93E-04	+	8.31E-02		1.10E-03	-
IMAGE:667527	7.90E-02		1.25E-01		1.18E-03	-
BAK1	2.31E-01		2.03E-01		1.24E-04	-
BMP5	6.03E-01		6.68E-02		3.47E-04	-
IMAGE:727289	9.89E-02		2.13E-01		7.57E-03	-
GPD1	2.00E-01		2.26E-01		1.13E-02	-
JUP	1.20E-01		1.60E-01		2.88E-02	-
PTPN4	6.68E-01		8.01E-01		3.63E-03	-
PSCD3	5.28E-01		7.46E-01		1.01E-02	-
MAPK3	1.64E-01		1.06E-01		6.76E-04	-
IMAGE:485104	5.66E-02		7.75E-02		3.89E-02	-
CBL	2.83E-01		6.75E-02		9.81E-03	-
IMAGE:743422	6.29E-01		9.62E-01		6.44E-04	-
FLJ23577	4.51E-01		4.69E-01		2.05E-02	-
CD34	9.01E-02		3.40E-01		2.87E-02	-
LOC388284	7.04E-01		5.78E-02		3.17E-03	-
IMAGE:383718	8.18E-01		2.18E-01		1.36E-03	-
MT1A	2.61E-05	+	3.87E-01		3.08E-02	-
RUNX3	2.83E-01		5.01E-02		1.18E-02	-
IMAGE:363955	3.66E-01		6.75E-01		3.01E-02	-
MATN1	9.25E-01		1.11E-01		2.21E-03	-
ME3	4.40E-01		2.63E-01		4.12E-03	-
DIO1	1.78E-01		7.02E-01		6.76E-03	-
STYXL1	2.64E-01		9.11E-02		3.71E-03	-
MFAP3L	3.24E-01		2.57E-01		1.17E-02	-
TSPAN5	5.02E-04	+	3.82E-01		2.95E-03	-
KLHL9	5.52E-01		5.13E-02		8.00E-03	-
F5	8.66E-01		2.39E-01		6.73E-03	-
TRIP12	8.75E-01		9.18E-02		4.40E-02	-
RNF32	7.53E-01		8.49E-02		1.31E-02	-
IMAGE:744657	3.90E-03	+	9.29E-01		4.72E-02	-
CD33	1.47E-01		2.75E-01		1.10E-02	-

E2F5	1.88E-03	+	8.55E-02		4.22E-02	-
KCNQ2	1.91E-01		7.52E-02		1.93E-02	-
LOC440934	3.34E-01		4.18E-01		1.87E-03	-
LHB	8.09E-01		1.77E-01		7.26E-04	-
HOXD11	4.62E-01		6.74E-01		2.07E-02	-
MLPH	9.60E-01		3.21E-01		2.66E-03	-

FTA-aty/FTA	FTA		T-UM		PTC	
Symbol	*P-value*	*regulation*	*P-value*	*regulation*	*P-value*	*regulation*
S100A10	3.57E-03	-	1.36E-02	-	2.71E-02	+
ATR	2.80E-03	-	4.50E-02	-	7.01E-01	
HCLS1	1.19E-02	-	1.58E-02	-	6.30E-03	+

FTA-aty/PTC	FTA		T-UM		PTC	
Symbol	*P-value*	*regulation*	*P-value*	*regulation*	*P-value*	*regulation*
KLF1	3.11E-02	+	3.96E-03	-	2.18E-05	-
C20orf19	1.70E-01		4.72E-02	-	3.49E-03	-
PPIL4	7.07E-01		2.50E-02	-	1.98E-02	-
IL24	8.16E-02		4.68E-02	-	4.03E-04	-
TWIST1	3.61E-01		1.20E-02	-	5.35E-03	-
KLK2	2.52E-01		1.60E-02	-	9.55E-03	-
ODAM	8.94E-01		3.87E-02	-	3.19E-02	-
FRZB	1.82E-01		5.44E-03	-	4.43E-02	-
ROGDI	3.87E-01		4.65E-02	-	1.13E-02	-
VAV2	4.98E-01		1.65E-03	-	1.90E-03	-

FTA/PTC	FTA		T-UM		PTC	
Symbol	*P-value*	*regulation*	*P-value*	*regulation*	*P-value*	*regulation*
KLRD1	1.85E-02	-	5.02E-01		3.52E-03	-
PICK1	4.08E-02	-	8.15E-01		1.97E-02	-
IMAGE:742837	3.37E-03	-	5.67E-01		1.61E-05	-
ABCA8	4.34E-03	-	3.57E-01		1.33E-02	-
GFPT2	3.56E-03	-	6.63E-01		1.53E-02	-
DCAMKL1	3.85E-02	-	2.96E-01		1.08E-02	-
SYP	2.73E-02	-	2.28E-01		4.47E-03	-
SPI1	4.32E-02	-	3.29E-01		3.77E-03	-
IMAGE:745512	4.78E-02	-	1.24E-01		7.44E-03	-
CCL11	2.55E-02	-	4.22E-01		1.96E-02	-

ALL 3	FTA		T-UM		PTC	
Symbol	*P-value*	*regulation*	*P-value*	*regulation*	*P-value*	*regulation*
CD274	2.67E-02	-	5.75E-04	-	3.16E-03	-
IGSF10	1.54E-02	-	2.94E-02	-	1.48E-02	-
IL2RA	4.10E-03	-	8.52E-03	-	2.53E-02	-
PTGES	1.17E-02	-	1.40E-03	-	6.43E-03	-
PLCG2	3.33E-02	-	3.03E-03	-	2.40E-03	-
CD38	4.13E-02	-	1.54E-02	-	1.98E-03	-
PCSK6	1.99E-02	-	1.15E-02	-	7.63E-03	-
BAP1	3.39E-02	-	1.93E-02	-	2.28E-02	-
LILRB3	2.69E-02	-	4.93E-03	-	3.71E-03	-

Supplementary table 2: histological features and markers on immunohistochemestry for the T-UM samples. The histological criteria (presence or absence of questionable PTC nuclear features, questionable capsular and/or vascular invasion, worrisome cellular and cytoarchitectural architectural features) are quoted 0 if absent, 1 if suspicious, 2 if obvious. The expression of the HBME-1, Galectin-3 and Cytokeratin 19 markers by immuno histochemistry are quoted 0 if absent, 1 if expressed in less than 50% of tumour cells, 2 if expressed in more than 50% of the cells. The TPO marker is quoted 0 if expressed in 80% or more of the tumour cells, 1 if expressed between 50 and <80%, and 2 when expressed < 50%. (*) Samples included in the microarray study. (**) Samples included in the quantitative RT-PCR study. (NA) Not available.

Symbol	% Cross Validation support	CloneID	UniGene Cluster	Locus Link ID	Name
IMAGE:4811759	100	687667	Hs.537002		CDNA clone IMAGE:4811759
TESC	99	745490	Hs.525709	54997	Tescalcin
CITED1	99	265558	Hs.40403	4435	Cbp/p300-interacting transactivator, with Glu/Asp-rich carboxy-terminal domain, 1
ECM1	90	301122	Hs.81071	1893	Extracellular matrix protein 1
MRC2	95	235882	Hs.7835	9902	Mannose receptor, C type 2
CDH3	91	359051	Hs.554598	1001	Cadherin 3, type 1, P-cadherin (placental)
CLDN1	88	594279	Hs.439060	9076	Claudin 1
ABCC3	76	208097	Hs.463421	8714	ATP-binding cassette, sub-family C (CFTR/MRP), member 3
NPAL3	87	738970	Hs.523442	57185	NIPA-like domain containing 3
DPP4	74	343987	Hs.368912	1803	Dipeptidyl-peptidase 4 (CD26, adenosine deaminase complexing protein 2)
CACNB2	74	173841	Hs.59093	783	Calcium channel, voltage-dependent, beta 2 subunit
CAPN3	47	757248	Hs.143261	825	Calpain 3, (p94)
ABCC5	43	212366	Hs.368563	10057	ATP-binding cassette, sub-family C (CFTR/MRP), member 5
CCND1	42	324079	Hs.523852	595	Cyclin D1
CD5	31	356841	Hs.58685	921	CD5 molecule
SNCB	27	50202	Hs.90297	6620	Synuclein, beta

Supplementary Table 3: Specific primers used for Real-Time quantitative PCR.

	Questionable PTC nuclear features	Questionable Capsular and or vascular invasion	Worrisome cellular and cytoarchitectural features	HBME-1	Gal-3	CK-19	TPO	Sum
1*	1	0	2	0	0	0	0	3
3*	2	0	0	0	0	0	0	2
4*	1	0	2	2	2	NA	NA	7
15*	1	2	2	0	2	0	1	8
16*	2	2	0	2	2	2	2	12
59*	1	2	2	1	0	0	1	7
91*	2	0	2	1	2	0	2	9
95*	1	2	0	1	2	0	1	7
184*	2	0	0 (oncocytic tumour)	0	2	0	2	6
188*	2	2	2	2	1	1	1	11
1C	2	2	2	2	1	2	2	13
2C	0	0	2	0	0	0	0	2
3C	0	0	2	0	0	0	0	2
4C	1	0	2	0	2	0	2	7
5C	2	0	2	0	2	0	0	6
6C	2	0	2	1	1	0	0	6
7C	1	2	2	2	0	2	0	9
8C	1	0	2 (oncocytic component)	0	0	0	0	3
9C**	2	0	2	1	1	0	0	6
10C**	1	2	2	2	1	0	0	8
11C**	2	0	2	1	2	1	1	9
12C**	2	0	2	0	2	0	NA	6
13C**	1	0	2	0	0	0	0	3
14C**	1	0	2	0	0	0	0	3

Supplementary Table 4: Similarity of the FTA, T-UM and PTC samples to the PTC class. The P-values were computed from permutation tests (30000 permutations). (FTA-a) macrofollicular adenoma, (FTA-b), microfollicular adenoma, (MNG) multinodular goitre.

SAMPLES	SIMILARITY	P-VALUE
FTA-a_147	0.08	7.78E-02
FTA-a_149	0.06	7.82E-02
FTA-a_167	0.07	7.78E-02
FTA-a_168	0.28	2.76E-02
FTA-a_179	-0.08	1.00E+00
FTA-a_19	-0.12	1.00E+00
FTA-a_20	0.14	7.47E-02
FTA-a_21	-0.08	1.00E+00
FTA-a_26	-0.22	1.00E+00
FTA-a_27	0.18	6.84E-02
FTA-a_28	0.05	7.89E-02
FTA-a_29	-0.22	1.00E+00
FTA-a_30	-0.07	1.00E+00
FTA-a_34	0.24	4.92E-02
FTA-a_36	0.11	7.64E-02
FTA-a_39	-0.13	1.00E+00
FTA-a_41	-0.29	1.00E+00
FTA-a_42	-0.15	1.00E+00
FTA-a_43	-0.02	4.20E-01
FTA-a_47	0.08	7.76E-02
FTA-a_48	0.11	7.67E-02
FTA-a_49	-0.09	1.00E+00
FTA-a_50	0.07	7.80E-02
FTA-a_54	-0.13	1.00E+00
FTA-a_55	-0.17	1.00E+00
FTA-a_64	0.25	4.36E-02
T-UM_1	-0.27	1.00E+00
T-UM_15	0.27	3.41E-02
T-UM_16	0.45	3.00E-04
T-UM_184	0.13	7.55E-02
T-UM_188	0.37	5.26E-03
T-UM_3	-0.10	1.00E+00
T-UM_4	0.34	9.95E-03
T-UM_59	0.33	1.35E-02
T-UM_91	0.27	3.38E-02
T-UM_95	0.18	6.86E-02
FTA-b_128	-0.26	1.00E+00
FTA-b_154	-0.15	1.00E+00
FTA-b_170	-0.15	1.00E+00
FTA-b_172	-0.04	9.61E-01
FTA-b_176	0.02	8.16E-02
FTA-b_180	-0.12	1.00E+00
FTA-b_182	-0.24	1.00E+00
FTA-b_191	-0.13	1.00E+00
FTA-b_25	-0.22	1.00E+00
FTA-b_35	0.06	7.83E-02
FTA-b_38	-0.30	1.00E+00

SAMPLES	SIMILARITY	P-VALUE
FTA-b_44	-0.18	1.00E+00
FTA-b_52	-0.37	1.00E+00
FTA-b_61	-0.13	1.00E+00
FTA-b_67	-0.13	1.00E+00
FTA-b_68	-0.09	1.00E+00
FTA-b_73	-0.19	1.00E+00
FTC_130	0.01	8.58E-02
FTC_135	-0.07	1.00E+00
FTC_6	0.20	6.28E-02
MNG_131	0.18	6.94E-02
MNG_137	0.09	7.74E-02
MNG_138	0.14	7.50E-02
MNG_143	-0.14	1.00E+00
MNG_145	-0.19	1.00E+00
MNG_163	0.05	7.90E-02
MNG_178	-0.07	1.00E+00
MNG_181	0.05	7.86E-02
MNG_22	-0.01	1.96E-01
MNG_23	0.09	7.72E-02
MNG_24	-0.18	1.00E+00
MNG_31	-0.04	9.84E-01
MNG_32	0.07	7.81E-02
MNG_33	-0.30	1.00E+00
MNG_37	-0.07	1.00E+00
MNG_40	0.14	7.47E-02
MNG_46	0.04	7.90E-02
MNG_53	-0.13	1.00E+00
MNG_62	0.17	6.98E-02
MNG_63	-0.12	1.00E+00
MNG_65	-0.23	1.00E+00
MNG_66	-0.08	1.00E+00
MNG_69	0.18	6.79E-02
MNG_72	-0.09	1.00E+00
PTC_104	0.66	3.33E-05
PTC_113	0.81	3.33E-05
PTC_124	0.79	3.33E-05
PTC_140	0.50	3.33E-05
PTC_141	0.59	3.33E-05
PTC_146	0.22	5.52E-02
PTC_152	-0.08	1.00E+00
PTC_157	0.61	3.33E-05
PTC_158	0.78	3.33E-05
PTC_173	0.53	3.33E-05
PTC_177	0.76	3.33E-05
PTC_97	0.77	3.33E-05
PTC_98	0.48	9.98E-05

Supplementary table 5: Specific primers used for Real-Time quantitative PCR

Gene	Primer sense (5' – 3')	Primer antisense (5' – 3')	Annealing (°C)
CDH3	GTG-GAC-CTC-TCG-GGT-CTC-T	GAC-TGT-CTC-GCC-ATT-CCG-C	**60**
CLDN1	GCC-CCA-GTG-GAG-GAT-TTA-CT	CAC-CCC-CAA-TGA-CAG-CCA-T	**58**
ECM1	GCC-TCT-GAG-GGA-GGC-TTC-A	GAG-CTG-GCG-TTC-CTT-CCT-T	**58**
CITED1	GCG-AAG-GAG-GAT-GCC-AAC-C	CAG-CTG-CAT-ACT-AGC-CAG-CA	**58**
MRC2	CTT-CCT-CAT-CTT-CAG-CCA-TG	CAA-TGC-CAG-CGA-AGA-TTC-AG	**55**
ABCC5	GGG-TAT-AGA-AGT-GTG-AGG-GA	CAA-GCC-ATG-ATG-GTA-CTT-TC	**55**
DPP4	TTG-TCA-CCA-TCA-TCA-CCG-TG	CTA-AGA-GAA-TAA-ACT-GCC-CAT	**55**
ABCC3	GGC-TCC-AAG-TTC-TGG-GAC-T	CCC-AGG-ACC-ATC-TTG-AGC-TT	**55**
CAPN3	GAT-GGA-GCC-AAC-AGA-ACT-GA	GCA-TAA-GCC-TTC-TCC-AGC-AG	**55**
ACTIN	CAC-CCT-GTG-CTG-CTC-ACC-GA	CCA-GGG-AGG-AAG-AGG-ATG-C	**58**

3.3.3 Discussion

Les profils associés aux tumeurs de malignité incertaine (T-UM) montrent globalement une forte corrélation avec les carcinomes folliculaires et papillaires. Huit tumeurs sur dix (80%) se sont révélées des variants folliculaires de carcinomes papillaires. Cette dernière sous-classe est homogène avec les papillaires classiques. Les deux autres cas (20%) sont validés comme des tumeurs bénignes par les marqueurs immuno-histochimiques.

Les résultats obtenus sont validés de manière indépendante par leur cohérence avec la littérature. Des gènes bien connus pour être spécifiques de la classe papillaire sont aussi mis en évidence dans notre analyse: CITED1, CTNNA1, DPP4 and CDH3 (Huang *et al.*, 2001; Jarzab *et al.*, 2005). La démarche statistique est aussi validée par l'étude des mutations spécifiques aux tumeurs papillaires, RET/PTC, que l'on retrouve dans les T-UM. Ces mutations ne sont d'ailleurs ni systématiquement associées aux tumeurs papillaires et T-UM, ni associées aux seuls meilleurs représentants de ces classes.

La corrélation des profils d'expression au score d'un pathologiste expérimenté démontre la possibilité de mettre en place un outil pour le diagnostic des T-UM malins. Nous avons caractérisé les relations fonctionnelles entre ces T-UM et les 2 principaux types de cancers de la thyroïde. Les données d'expression à grande échelle nous permettent d'établir un diagnostic lorsque l'inspection morphologique est difficile et sujette à la subjectivité des critères de décision. Pour continuer dans cette voie, il faudrait envisager des études incluant un nombre encore plus grand d'échantillons testés, évaluer les coûts diagnostiques et les bénéfices en termes de fiabilité diagnostique par rapport aux outils existants. Les T-UM semblent représenter un état précoce de la transformation des nodules en cancers papillaires. Comprendre la régulation des gènes différentiels spécifiques à ces tumeurs permettrait de définir des cibles thérapeutiques adaptées.

3.4 Article 3

3.4.1 Introduction

Titre : Up-regulation of DAP3 in the thyroid oxyphilic cells adenomas and carcinomas

Nous avons travaillé dans les deux premières études sur la classification des tumeurs de la thyroïde par une approche pan génomique. Certaines analyses, comme la classification, peuvent sélectionner des gènes efficaces sans que leur pertinence biologique ne soit élevée.

Nous avons étudié les données d'expression d'un gène d'intérêt, DAP3, et de sa protéine. Ce gène déjà étudié dans le laboratoire a d'abord été impliqué dans l'apoptose (Kissil *et al.,* 1995). Cette protéine exercerait une fonction pro-apoptotique dans la matrice mitochondriale. Elle y resterait toujours localisée, mais ce point reste discuté (Miyazaki *et al.,* 2002, Berger and Kretzler 2002). DAP3 joue aussi un rôle dans la scission mitochondriale (Mukamel *et al*) qui serait en relation avec l'apoptose (Perfettini *et al.,* 2005). La mise en évidence de DAP3 comme protéine constitutive de la petite sous-unité du mitoribosome en fait un candidat de choix pour une fonction indispensable au processus de traduction mitochondriale (Cavdar Koc *et al.,* 2001, Suzuki *et al.,* 2001).

Nous avons mesuré l'expression du messager de Dap3 par RT-PCR pour 12 adénomes et 5 carcinomes oncocytaires en paire avec leur tissu sain avoisinant. Nous avons regroupé des données immuno-histochimique de biopuces à tissus de 97 tumeurs thyroïdiennes associées à leur tissu sain avoisinant. Cet ensemble contenait 51 carcinomes (33 papillaires, 9 oncocytaires, 5 folliculaires, 3 peu différenciés, 1 anaplasique) et 46 adénomes (33 oncocytaires, 8 folliculaires, 5 T-UMs).

Nous avons réalisé une méta-analyse bioinformatique de données d'expression publiques associée à l'analyse de séquences génomiques pour déduire les meilleurs candidats à la régulation de dap3. Nous avons d'abord établi la liste des gènes co-régulés avec dap3 à partir de 100 jeux de données publiques. La recherche des sites potentiels de fixation de facteurs de transcription s'est faite par l'intermédiaire d'une approche probabiliste avec une grande attention sur les modèles de références utilisés. Nous avons associé nos résultats à la recherche de sites de fixations dans les séquences promotrices des gènes du mitoribosome.

3.4.2 Article

Title: Up-regulation of DAP3 in the thyroid oxyphilic cells adenomas and carcinomas

Authors:

Caroline JACQUES[1-3*], Jean Fred FONTAINE[1-2*], Brigitte FRANC4[4], Delphine PRUNIER-MIREBEAU[1-3], Frédérique SAVAGNER[1-3], Yves MALTHIERY[1-3]

*: The two first authors contributed equally to this work.

1: INSERM, U694, Angers, F-49033 France

2: Université d'Angers, Faculté de Médecine, Angers, F-49033 France

3: CHU Angers, Laboratoire de Biochimie, Angers, F-49033 France

4: Laboratoire d'Anatomie Pathologique, Hôpital A Paré, 92104 Boulogne, France

Corresponding author:

Caroline Jacques, Inserm U 694, Laboratoire de Biochimie, CHU, 4 rue Larrey, 49033

Angers, France, tel : +33 241 35 33 14, Fax : +33 241 35 40 17, caroline.jacques@univ-angers.fr.

Running Title: Up-regulation of DAP3 in thyroid oncocytoma

Abstract

The Death Associated Protein 3 (DAP3), a GTP-binding constituent of the small subunit of the mitochondrial ribosome, is implicated in the first steps of apoptosis inducing pathways and is also involved in the maintenance of the mitochondrial network. We have investigated the mitochondrial role of DAP3 by analyzing its mRNA and protein expression in thyroid oxyphilic cells tumors: the thyroid oncocytomas. Benignant and malignant thyroid oncocytomas showed an up-regulation of the DAP3 mRNA and protein expression. We analysed in silico the promoter sequence of 337 potential DAP3 co-regulated genes, 71/85 of the mitochondrial ribosome genes and the DAP3 gene. Twelve transcriptional factors were highlighted presenting putative motifs in these gene promoters. These transcriptional regulators can be considered as potential regulators of the DAP3 gene expression. These factors are mainly implicated in cellular growth (ELK1, ELK4, RUNX1, HOX11-CTF1, TP53, TAL1-TCF3) and mitochondrial biogenesis (NRF1, GABPA, PPARG-RXRA, ERRalpha). Considering the DAP3 protein is a component of the mitochondrial ribosome, these data suggest a role for this protein in the mitochondrial translation. The DAP3 protein can be considered as a new link between the mitochondrial homeostasy and the tumorigenesis process.

Introduction

Performing most of the cellular ATP synthesis within the oxidative phosphorylation and being implicated in the intrinsic pathway of apoptosis induction, the mitochondrion plays a major role in life and death of the cell (Regula et al., 2003 for review). Mitochondrial defects have been implicated in a wide variety of pathologies as degenerative diseases, cancer, but also in physiological process as aging (Schapira et al., 2006).

The mitochondrion possesses its own genome, the mitochondrial DNA (mtDNA), which encodes for 37 genes: 2 rRNA, 22 tRNA and 13 proteins. The 13 proteins are translated by the mitochondrial ribosomes in the mitochondrial matrix and are all essential catalytic components of four of the five complexes of the OXPHOS system. The mitochondrial ribosome is a two-asymmetric subunits ribonucleoproteic complex. The large subunit of the mitochondrial ribosome is constituted of a 16S rRNA molecule associated with 48 proteins, and the small subunit of a 12S rRNA molecule associated with 29 proteins. The 77 mitochondrial ribosomal proteins (MRPs) are encoded by nuclear genes. After transcription of the mitochondrial ribosomal genes and translation of the corresponding peptides in the cytoplasm, mitochondrial

addressing signals drive the proteins in the appropriate mitochondrial compartment where they can associate with the rRNA molecules (O'Brien et al., 2002 ; Mears et al., 2006 for review).

The DAP3 protein has been demonstrated to be a constituent of the small subunit of the mitochondrial ribosome (Suzuki et al., 2001 ; O'Brien T et al., 2005). This protein has no counterpart in the bacterial or cytoplasmic ribosomes (Cavdar Koc et al., 2001). DAP3 has been revealed by its implication in the programmed cell death in 1995 (Kissil et al., 1995). Its mitochondrial addressing is driven by a N terminal sequence (Morgan et al., 2001 ; Mukamel et al., 2004). Multiple roles have been suggested for this protein. First identified as a pro-apoptotic factor, this protein is suspected to interact with the death domain of the TNF-related apoptosis-inducing ligand (TRAIL) receptors and the Fas-associated death domain protein (FADD) in the cytosol, linking FADD and TRAIL (Kissil et al., 1995 and 1999 ; Miyazaki et al., 2001). Knocking down this protein by siRNA has been shown to reduce the mitochondrial network fragmentation, suggesting a role for DAP3 in the fission of the mitochondrial network (Mukamel, 2004).

The implication of this protein in the first steps of apoptosis induction suggests a modulation of expression of the DAP3 mRNA and/or protein in pathological contexts as the development of tumors. The expression of DAP3 was first explored in the glioblastoma multiforme (GBM) cells and revealed to be over-expressed in the invasive GBM cells. In vitro, the highly motile glioma cells over-expressing DAP3 were also more resistant to apoptosis compared to their parental cells (Mariani et al., 2001). In the thymoma, the DAP3 mRNA expression was correlated with the stage described in the WHO classification, the stage IV thymomas displaying significantly higher expression compared to the stage I thymomas (Sasaki et al., 2004). In the adult asthma, specific variants of DAP3 (polymorphism) are also suspected to be associated with airway inflammation and remodelling of bronchial epithelium (Hirota et al., 2004).

Most of the tumor types display preferentially a glycolytic metabolism. However, in our laboratory, we have investigated the thyroid oncocytoma, a mitochondrial rich thyroid tumor characterised by an oxidative metabolism (Baris et al., 2004 ; Savagner et al., 2003). The reduced expression of the DAP3 protein has recently been described in senescent cells induced by oxidative stress. Using DAP3-shRNA, Murata et al. showed a resistance to oxidative stress and a decrease of intracellular reactive oxygen species (ROS) production (Murata et al., 2006). As regards to its mitochondrial localisation, its implication in the mitochondrial ribosome as well as its role in modulation of the mitochondrial network and the oxidative stress, it would be

of interest to study the DAP3 protein expression in a model displaying an oxidative metabolism. The DAP3 protein is a component of the mitochondrial ribosome. This suggests its implication in the structure or the function of this nucleoribosomic complex, and points DAP3 as a potential actor or regulator of the mitochondrial translation.

To explore the role of DAP3 in the mitochondrial translation, we studied its expression by RT-PCR and IHC techniques in the thyroid oncocytoma. We analysed multi-type micro-arrays and genomics data to identify co-regulated genes and potential regulators of the DAP3 gene. Our results support a role for DAP3 in mitochondrial translation.

Results

Expression of DAP3 in oncocytic tumors by RT-PCR

We measured the DAP3 mRNA expression in 12 oncocytic adenomas and 5 oncocytic carcinomas. The mean expression level of DAP3 in oncocytic adenomas and carcinomas was significantly different from the paired normal tissue (respectively, P<3.46E-03 and P<3.34E-02). Figure 1 shows DAP3 expression fold change in tumors relatively to the normal tissue. The distribution of expression levels in oncocytic adenomas had a median value of 1.19, a minimum of 0.84 and a maximum of 3.23. Only two levels were below 1 (0.84 and 0.91). The distribution in oncocytic carcinomas had a median value of 2.24, a minimum of 1.06 and a maximum of 2.57.

Immunohistochemical staining

The DAP3 protein expression was evaluated on immunohistochemical staining for 99 tumors and 97 normal tissue samples (Table 1). In all histological types, the DAP3 expression showed various intensities and was localized into the cytoplasm, either close to the nucleus or over all the cytoplasm with granular aspects (Figure 3). The normal thyroid tissue was mainly negative for DAP3, with only 4 positives (4.12%) out of 97 samples. There were several significant over-representations of DAP3-positive cells in 4 types of carcinomas compared to the normal tissue.

Eighteen (51.43%) oncocytic adenomas out of 35 were positives (P<5.43E-06). Twelve (36.36%) papillary carcinomas out of 33 were positives (P<6E-04). Five (55.55%) oncocytic carcinomas out of 9 were positives (P<3.3E-03). Two (66.66%) poorly differentiated

carcinomas out of 3 were positives (P<5E-02). Other tumor types showed either any or not over-represented DAP3-positive cells.

Identification and categorisation of DAP3 co-expressed genes

Microarray transcriptomic data can identify co-expressed genes but deduced gene-expression correlations vary between studies. Variations involve technical and biological reasons like the chip technology or the nature of the samples. Methodologies were proposed to circumvent these biases and to analyse several heterogeneous datasets (Becker et al., 2001 ; Miles et al., 2001). We used the TMM web server to find co-expressed genes with DAP3 by comparing the gene correlations of 100 publicly available datasets. As shown by the authors, a correlation found in at least 3 datasets is significant, the higher the better. We used this criterion to select 337 correlated genes (Supplemental Table 1). Each gene was found correlated with DAP3 in at least 3 and up to 9 datasets.

We characterized the list by clustering GO co-annotated genes. One hundred and eighty six (55%) genes out of 337 were divided into 22 significant clusters (P<0.05) which shared at least 80% of their annotations (Supplemental table 2). One cluster was removed (cluster n°16) because of no significant P-values for its GO terms. We removed intra-cluster redundancy as respect to the GO tree and grouped the clusters as respect to their similarity. The similarity was either 0 or 100% (shared genes) thus each cluster was contained into a bigger one, except the biggest one (Cluster A, intracellular, n=186). Clusters were related to intracellular membrane-bound organelle (mitochondrion, oxidoreductase activity and catalytic activity), primary metabolism, cellular biosynthesis and ligase activity, macromolecule metabolism transcription, and protein metabolism (biopolymer metabolism, ATP binding and protein biosynthesis).

Bioinformatics search for regulators of the co-expresssed genes

Assuming that correlated genes have a common regulatory mechanism, we searched for transcription factors which could have been involved into the regulation of the DAP3 co-expressed genes. We examined the gene promoters using the 123 JASPAR database TFBS position-weight matrices and two additional matrices defining the NRF-1 and ERRalpha binding sites (Figure 2). The gene promoter sequences were extracted from –1000 to +1000 from the transcription starting site and scanned for motif over-representation using the Clover software. Significant motif over-representation was considered at 5% risk comparing to 2 background models: a set of human promoters and a set of CpG islands.

Table 2 shows the over-represented TFBS motifs in the DAP3 co-expressed genes promoters. Two of the 8 corresponding transcriptional factors, NRF1 and GABPA, are activators of the mitochondrial biogenesis (Scarpulla et al., 2006). STAF is a transcriptional activator of a large number of genes in the Human genome, these target genes comprising DNA binding proteins and transcription factors (23%), protein synthesis degradation turnover and modification (21%), and genes implicated in cell growth (Myslinski et al., 2006). E74A is a transcriptional activator of genes implicated in autophagic cell death (Lee et al., 2002). ELK1 and ELK4 are members of the Ets family of transcription factors and of the ternary complex factor (TCF) subfamily. The TCF proteins act primarily through a ternary complex by binding to the serum response factor and the serum reponse element in the promoter of immediate early class genes such as the c-fos proto-oncogene (Yang et al., 2003). The RUNX1 (also called RUNX1/AML1) gene encodes an important regulator of hematopoiesis and is the target of several genetic alterations during leukemogenesis (Mikhail et al., 2006). HOX11 and CTF1 are two transcriptional factors which up-regulation and interaction are implicated in the hematopoietic precursor cells immortalisation (Zhang et al., 1999).

Regulators of the mitochondrial ribosome

We computed an independent TFBS search in promoter sequences of the mitochondrial ribosome. We separately search motifs in gene promoters of the large subunit (44 sequences) and in gene promoters of the small subunit (27 sequences). Table 2 showed the statistically significant over-represented motifs.

The large subunit analysis revealed several transcriptional factors in common with the 337 “co-regulated” genes previously explored: STAF, NRF1, E74A, ELK4, ELK1 and GABPA. Two additional transcriptional factors where significantly represented in the sequences analysed. PPARG and RXRA are two transcriptional factors respectively implicated in the adipocytes differenciation and lipogenesis (Metzger et al., 2005). TP53 is transcriptional factor implicated in numerous proscesses as cell cycle, apoptosis, angiogenesis, senescence and DNA repair (Bode et al., 2004).

The small subunit analysis revealed three transcriptional factors in common with the TMM genes previously explored and the mitoribosomal large subunit; E74A, NRF1 and GABPA. Two supplementary transcriptional factors where also significantly represented specifically in the small subunit sequences. ERR alpha is a regulator of beta-oxidation which acts via the control of the medium chain acyl-coenzyme A dehydrogenase (MCAD) promoter. ERRalpha is

also implicated in PGC-1a-induced mitochondrial biogenesis (Scarpulla et al., 2006). TAL1 is a basic helix-loop-helix (bHLH) transcription factor required for normal hematopoiesis. Its aberrant expression leads to T-cell acute lymphoblastic leukemia (Palomero et al., 2006).

Best candidate TFBS into the DAP3 gene promoter

As DAP3 was the crossing point between all our previous transcriptional factors sites explorations, we searched for motifs of all the above mentioned tr anscriptional factors sites in the DAP3 promoter sequence. Its scanning by the Possum program for each candidate TFBS gave score distributions (data not shown). Highly scored sites detached from the background were visually selected (Figure 4 et Table 3). All of the previously cited transcriptional factors showed highly scored TFBS in the DAP3 gene promoter.

TFBS were distributed in the promoter region from –679 to 998 on both strands. No sites were found between –1000 to –680. We found at least 2 sites per motif (Hox11-CTF1, RUNX1 and TAL1-TCF3) and at most 5 sites (E74A). There were site overlaping between motifs. Several sites for E74A, ELK1, ELK4 and HOX11-CTF1 overlapped in 3 areas ([-458;-449], [344;354], [-79;-69]). Two ERRalpha sites overlapped with ELK1 in [-79;-69] and RUNX1 in [-276;-262]. Two GABPA sites overlapped with E74A in [898;907] and STAF in [185;210]. NRF-1 and TAL1-TCF3 did not overlap with other motif sites. Remarkably, two highly scored TAL1-TCF3 sites were strictly localized at –399 to –386 on both strands.

Discussion

Considering the multiple possible functions for the DAP3 protein, we studied its expression in thyroid oncocytomas, mitochondria rich benignant or malignant tumors. Our results show a significant over-expression of the DAP3 mRNA in the thyroid oncocytic adenoma and thyroid oncocytic carcinoma. We then looked at the expression of the DAP3 protein in thyroid oncocytoma and found an over-expression of DAP3. We previously described the expression of the mRNA DAP3 in rho° cells, but the absence of expression of the protein DAP3 (Jacques et al., 2006). We suggested a post-translational down regulation of the protein, argumented by the absence of the RNA 12S molecule in the mtDNA depleted cells. Conversely, the up-regulation of DAP3 revealed here in the thyroid oncocytomas suggests the need of this protein for a particular function in these tumors.

A previous study on various types (gastric, lung, colorectal and hepatocellular) and grades of carcinomas (359 tumors) did not show any mutation in the P-loop domain of DAP3 (Woo Lee

et al., 2006). This suggests that the mutation of this GTPasic motif is not essential in the development of these types of carcinomas. Then the functional proapoptotic capacity of DAP3 is not an obstacle in the balance between the proliferation and the apoptosis of cells in these cancers. These data and the previously published DAP3 expression analysis in two types of tumors (glioblastoma, thymoma) are not in favour of a major pro-apoptotic role of DAP3 in these cancers.

In the normal thyroid tissue, the cell death is almost inexistent. In the normal thyrocytes, the Fas protein receptor is expressed but the apoptosis process is stopped at the pro-caspase 3 step, which is not activated in caspases 3 (Mezosi et al., 2005). In the thyroid carcinomas, the Fas receptor is expressed but the induction of apoptosis by Fas ligand is stopped before activation of the caspase 8 (Mitsiades et al., 2000). Conversely, the Fas expression has also been suggested to be re-oriented and exploited in these tumors for a cell growth advantage (Mitsiades et al., 2006). The TRAIL receptors R1 and R2 are expressed in the thyroid carcinomas and this death inducing pathway is functional in these tumors but not in the normal thyroid (Mitsiades et al., 2001). These data are also not in favour of a required proapoptotic role of DAP3 in the thyroid tumors.

We propose a different approach of the role of the DAP3 protein in the mitochondria. In the thyroid oncocytomas, the mitochondrial subunits of the respiratory chain complexes encoded by the mitochondrial DNA (mtDNA) are up-regulated and the respiratory chain complexes do not display any deficient activity (Savagner et al., 2001). This admits an efficient mitochondrial translation. We showed that DAP3 is up-regulated in these tumors. This protein being also a component of the mitoribosome, we postulate for a role of DAP3 in the mitochondrial translation, which is essential for the mitochondrial biogenesis observed in thyroid oncocytomas. The implication of DAP3 in the efficacy or fidelity of mitochondrial translation requires a functional DAP3 protein in these tumors. We postulate the overexpression of the DAP3 mRNA and protein in the thyroid oncocytic tumors is linked to the mitochondrial biogenesis observed in the oncocytes. We then propose the thyroid oncocytomas as a model to analyse the function of the DAP3 protein in the mitochondrial translation.

Transcriptomic analysis of thyroid oncocytomas were previously performed in our laboratory. Screening the expression of 6720 genes in 29 thyroid oncocytomas by microarray revealed the overexpression of numerous mitochondrial genes in the tumors compared to normal tissue (Baris et al., 2004). Amoung 126 up-regulated genes discriminating the thyroid oncocytomas from the other samples, genes coding for the four mitochondrial respiratory chain complexes

subunits (11) and the ATP synthase (2) were represented. The mitochondrial ribosome was represented with the MRPL49 gene, the 5th most overexpressed gene amoung the 126 genes, and which encodes a component of the large mitoribosomal subunit. In a two-step (differential display and macroarray) analysis on 6 thyroid oncocytic adenomas we also showed a two-fold or more up-regulation of 30 genes in the tumors. Twelve out of these 30 genes are mitochondrial mtDNA encoded genes and several mtDNA or nuclear encoded genes highlighted are redundant with the previously mentioned microarray results (Jacques et al., 2005). As postulated previously for the renal and salivary oncocytomas, our data in the thyroid oncocytomas are in favour of a coordinated regulation of the mitochondrial nuclear and mtDNA encoded genes in the oncocytic tumors (Heddi et al., 1996),(Baris et al., 2005).

We then postulated a step toward the exploration of a DAP3 function in the mitochondrial translation would be the analysis of the promoter sequence of the DAP3 gene and a comparison with the promoter sequence of other mitochondrial nuclear encoded genes.

We first analysed the gene correlations of 100 publicly available datasets using the TMM web server to find DAP3 co-expressed genes. We found 337 correlated genes in at least 3 and up to 9 datasets. These genes are mainly related to the nucleic acid metabolism, transcription and RNA processing, protein metabolism and mitochondrion. We postulated these genes have a common regulatory mechanism. Using 125 TFBS position-weight matrices, we searched for transcription factors which could have been involved in the regulation of the DAP3 co-expressed genes. Eight transcriptional factors were revealed. These factors are mainly implicated in tumorigenesis (ELK1, ELK4, RUNX1 and HOX11-CTF1) and mitochondrial biogenesis (NRF1, GABPA). E74 is related to the autophagic cell death, and STAF is a transcriptional factor of large spectrum.

The analysis of 44 out of the 52 large subunit proteins and 27 out of 33 small subunit proteins showed 4 supplementary transcriptional factors related to a potential regulation of the large or small mitochondrial ribosome subunits constituents (respectively PPARG-RXRA and TP53, ERRalpha and TAL1-TCF3). These factors are implicated in tumorigenesis (TP53 and TAL1-TCF3) or mitochondrial biogenesis (PPARG-RXRA and ERRalpha). The use of 2 background models (6461 human promoters and a set of 27 555 CpG islands) and the selection of the TFBS motifs significantly represented at 5% risk in the two models is highly selective. The redundancy of the results obtained with the 337 DAP3 potential co-regulated genes and the 71/85 mitochondrial ribosome proteins can be also considered as a validation of each analysis. We postulate that the DAP3 gene is regulated by transcriptional factors implicated mainly in

cellular growth (ELK1, ELK4, RUNX1, HOX11-CTF1, TP53, TAL1-TCF3) and mitochondrial biogenesis (NRF1, GABPA, PPARG-RXRA, ERRalpha).

The DAP3 mRNA and protein are over-expressed in thyroid oncocytoma, a mitochondria rich and oxydative metabolism tumor type. In a general biological context, the DAP3 gene can be suspected to be regulated by transcriptional factors implicated in the tumorigenesis and mitochondrial biogenesis. These hypothesis of gene regulation have however to be confirmed by experimentation, which are currently running. These data and the implication of the DAP3 protein in the mitochondrial ribosome small subunit suggest an implication of the DAP3 protein in the mitochondrial translation. DAP3 can also be considered as a link between the mitochondrial homeostasy and the tumorigenesis process. If a correct and increased mitochondrial physiology is required for the tumorigenesis process, one can expect a default in one of the mitochondrial ribosome proteins can disrupt the tumorigenesis process and direct the cell to death.

Materials and methods

Quantitative PCR

The expression level for the DAP3 gene was explored by quantitative real time polymerase chain reaction (RT-PCR) analysis using the actin gene as a reference and using the Light Cycler technology (Roche, Basel, Switzerland), as previously described (Jacques et al., 2006). The standards were obtained by PCR performed on normal thyroid tissue total cDNA. The DAP3 standard was obtained using the forward primer 5′-GCTGGGAAAGGAAGGATTTG-3′ and the reverse primer 5′-TTCGCGTTACTTAGGAACAG-3′. The actin standard was obtained using the forward primer 5'-CGACATGGAGAAAATCTGGC-3' and the reverse primer 5'-AGGTCCAGACGCAGGATGG-3'. The quantifications were performed using the same primers. Differences between mean DAP3 expression in tumor and normal tissues were tested by the one-tailed paired Student’s test. Significance was considered at 5% risk.

Immunohistochemistry

A total of 104 tumors were analysed comprising 35 papillary carcinomas, 10 oxyphilic cells carcinomas and 40 oxyphilic cells adenomas. These samples were used for tissue array construction, as described by Kononen et al. (Kononen et al., 1998). A triplet of tumor tissue and a triplet of the normal counterpart were arrayed for each sample. Immunostaining was performed using the standard avidin-biotin peroxidase technique. The primary antibody was the

monoclonal anti-DAP3 from Transduction Laboratories (Lexington, UK). Diaminobenzidin was used as the chromogen and hematoxylin as the nuclear counterstain. For negative controls, the primary antibody was either omitted or replaced by a suitable concentration of normal IgG of the same species. Over-representation of positive cells in tumors compared to their paired normal tissue was assessed by the two-tailed Fisher's exact test with Benjamini and Hochberg correction for multiplicity at 5% risk (Benjamini and Hochberg, 1995).

Bioinformatics analysis

DAP3 co-expressed genes were searched by the TMM web server (www.bioinformatics.ubc.ca/tmm) (Lee et al., 2004). Only significant genes were retained, i.e. genes which were correlated in at least 3 out of 100 microarray datasets. Gene Ontology (GO) clustering was performed by the GOproxy web server (http://crfb.univ-mrs.fr/GOToolBox/index.php) (Martin et al., 2004). Significant clusters (P<5%) of at least 10 genes which shared at least 80% of their GO annotations were selected. P-values were computed from the hypergeometric distribution and corrected by the Bonferroni method.

We collected 123 transcription factor binding sites (TFBS) motifs from the Jaspar database (http://jaspar.cgb.ki.se/cgi-bin/jaspar_db.pl). We aligned 9 sequences of NRF-1 binding sites to construct a position weight matrix (de Sury et al., 1998 ; Au et al., 1998 ; Elbehti-Green et al., 1998 ; Hirawake et al., 1999). The ERRalpha position weight matrix was described by Sladek et al. (Sladek et al., 1997). Visualisation was done by WEBLOGO web server (http://weblogo.berkeley.edu/). Promoter sequences were extracted by the Promoser web server (http://biowulf.bu.edu/zlab/promoser/) from –1000 to +1000 nucleotide bases from the transcription start site. Known repeats in sequences were marked with N characters. Duplicate promoters were removed from lists and entries that mapped to more than one locus were excluded. TFBS over-representation was searched by the Clover program (http://zlab.bu.edu/clover/) by using 2 background models. The first consisted in 6461 randomly chosen gene promoter sequences of the human genome from –1000 to +1000 nucleotide bases from the transcription start site. The second was composed of 27 555 sequences from CpG Islands regions. Significance was considered relatively to the two models at 5% risk. The best sites into the DAP3 promoter were defined by the Possum program (http://zlab.bu.edu/~mfrith/possum/) with the Jaspar motifs and the NRF-1 and ERRalpha previously described matrices. A visual inspection of the score distributions was used to select the best sites relatively to the background.

References

- Au, H. C. andI. E. Scheffler (1998). "Promoter analysis of the human succinate dehydrogenase iron-protein gene--both nuclear respiratory factors NRF-1 and NRF-2 are required." Eur J Biochem 251(1-2): 164-74.

- Baris, O.,F. Savagner, V. Nasser, B. Loriod, S. Granjeaud, S. Guyetant, B. Franc, P. Rodien, V. Rohmer, F. Bertucci, D. Birnbaum, Y. Malthiery, P. Reynier and R. Houlgatte (2004) "Transcriptional profiling reveals coordinated up-regulation of oxidative metabolism genes in thyroid oncocytic tumors." J Clin Endocrinol Metab 89(2): 994-1005.

- Baris, O.,D. Mirebeau-Prunier,F. Savagner,P. Rodien,B. Ballester,B. Loriod,S. Granjeaud,S. Guyetant,B. Franc,R. Houlgatte,P. Reynier andY. Malthiery (2005). "Gene profiling reveals specific oncogenic mechanisms and signaling pathways in oncocytic and papillary thyroid carcinoma." Oncogene 24(25): 4155-61.

- Becker, K. G. (2001) “The sharing of cDNA microarray data.” Nat Rev Neurosci. Jun 2(6):438-40. Review.

- Benjamini, Y., Hochberg, Y. (1995) “Controlling the false discovery rate: a practical and powerful approach to multiple testings.” JRSSB 57:289-300

- Bode, A. M. andZ. Dong (2004). "Post-translational modification of p53 in tumorigenesis." Nat Rev Cancer 4(10): 793-805.

- Cavdar Koc, E.,W. Burkhart,K. Blackburn,A. Moseley andL. L. Spremulli (2001). "The small subunit of the mammalian mitochondrial ribosome. Identification of the full complement of ribosomal proteins present." J Biol Chem 276(22): 19363-74.

- De Sury, R.,P. Martinez,V. Procaccio,J. Lunardi andJ. P. Issartel (1998). "Genomic structure of the human NDUFS8 gene coding for the iron-sulfur TYKY subunit of the mitochondrial NADH:ubiquinone oxidoreductase." Gene 215(1): 1-10.

- Elbehti-Green, A.,H. C. Au,J. T. Mascarello,D. Ream-Robinson andI. E. Scheffler (1998). "Characterization of the human SDHC gene encoding of the integral membrane proteins of succinate-quinone oxidoreductase in mitochondria." Gene 213(1-2): 133-40.

- Heddi, A.,H. Faure-Vigny,D. C. Wallace andG. Stepien (1996). "Coordinate expression of nuclear and mitochondrial genes involved in energy production in carcinoma and oncocytoma." Biochim Biophys Acta 1316(3): 203-9.

- Hirawake, H.,M. Taniwaki,A. Tamura,H. Amino,E. Tomitsuka andK. Kita (1999). "Characterization of the human SDHD gene encoding the small subunit of cytochrome b (cybS) in mitochondrial succinate-ubiquinone oxidoreductase." Biochim Biophys Acta 1412(3): 295-300.

- Hirota, T.,K. Obara,A. Matsuda,M. Akahoshi,K. Nakashima,K. Hasegawa,N. Takahashi,M. Shimizu,H. Sekiguchi,M. Kokubo,S. Doi,H. Fujiwara,A. Miyatake,K. Fujita,T. Enomoto,F. Kishi,Y. Suzuki,H. Saito,Y. Nakamura,T. Shirakawa andM. Tamari (2004). "Association between genetic variation in the gene for death-associated protein-3 (DAP3) and adult asthma." J Hum Genet 49(7): 370-5.

- Jacques, C.,O. Baris,D. Prunier-Mirebeau,F. Savagner,P. Rodien,V. Rohmer,B. Franc,S. Guyetant,Y. Malthiery andP. Reynier (2005). "Two-step differential expression analysis reveals a new set of genes involved in thyroid oncocytic tumors." J Clin Endocrinol Metab 90(4): 2314-20.

- Jacques, C.,A. Chevrollier,D. Loiseau,L. Lagoutte,F. Savagner,Y. Malthiery andP. Reynier (2006). "mtDNA controls expression of the Death Associated Protein 3." Exp Cell Res 312(6): 737-45.

- Kissil, J. L.,L. P. Deiss,M. Bayewitch,T. Raveh,G. Khaspekov andA. Kimchi (1995). "Isolation of DAP3, a novel mediator of interferon-gamma-induced cell death." J Biol Chem 270(46): 27932-6.

- Kissil, J. L., O. Cohen, T. Raveh andA. Kimchi (1999). "Structure-function analysis of an evolutionary conserved protein, DAP3, which mediates TNF-alpha- and Fas-induced cell death." Embo J 18(2): 353-62.

- Kononen, J.,L. Bubendorf,A. Kallioniemi,M. Barlund,P. Schraml,S. Leighton,J. Torhorst,M. J. Mihatsch,G. Sauter andO. P. Kallioniemi (1998). "Tissue microarrays for high-throughput molecular profiling of tumor specimens." Nat Med 4(7): 844-7.

- Lee, C. Y.,C. R. Simon,C. T. Woodard andE. H. Baehrecke (2002). "Genetic mechanism for the stage- and tissue-specific regulation of steroid triggered programmed cell death in Drosophila." Dev Biol 252(1): 138-48.

- Lee, H. K., Hsu A. K., Sajdak J., Qin J., Pavlidis P. (2004) “Coexpression analysis of human genes across many microarray data sets.” Genome Res. 14(6):1085-94.

- Mariani, L.,C. Beaudry,W. S. McDonough,D. B. Hoelzinger,E. Kaczmarek,F. Ponce,S. W. Coons,A. Giese,R. W. Seiler andM. E. Berens (2001). "Death-associated protein 3 (Dap-3) is overexpressed in invasive glioblastoma cells in vivo and in glioma cell lines with induced motility phenotype in vitro." Clin Cancer Res 7(8): 2480-9.

- Martin, D., Brun, C., Remy, E., Mouren, P., Thieffry, D., Jacq, B. (2004) “GOToolBox: functional analysis of gene datasets based on Gene Ontology.” Genome Biol. 5(12):R101. Epub 2004 Nov 26.

- Mears, J. A.,M. R. Sharma,R. R. Gutell,A. S. McCook,P. E. Richardson,T. R. Caulfield,R. K. Agrawal andS. C. Harvey (2006). "A structural model for the large subunit of the mammalian mitochondrial ribosome." J Mol Biol 358(1): 193-212.

- Metzger, D.,T. Imai,M. Jiang,R. Takukawa,B. Desvergne,W. Wahli andP. Chambon (2005). "Functional role of RXRs and PPARgamma in mature adipocytes." Prostaglandins Leukot Essent Fatty Acids 73(1): 51-8.

- Mezosi, E.,S. H. Wang,S. Utsugi,L. Bajnok,J. D. Bretz,P. G. Gauger,N. W. Thompson andJ. R. Baker, Jr. (2005). "Induction and regulation of Fas-mediated apoptosis in human thyroid epithelial cells." Mol Endocrinol 19(3): 804-11.

- Mikhail, F. M.,K. K. Sinha,Y. Saunthararajah andG. Nucifora (2006). "Normal and transforming functions of RUNX1: a perspective." J Cell Physiol 207(3): 582-93.

- Miles M. F. (2001) “Microarrays: lost in a storm of data?” Nat Rev Neurosci. 2(6):441-3. Review.

- Mitsiades, N.,V. Poulaki,S. Tseleni-Balafouta,D. A. Koutras andI. Stamenkovic (2000). "Thyroid carcinoma cells are resistant to FAS-mediated apoptosis but sensitive to tumor necrosis factor-related apoptosis-inducing ligand." Cancer Res 60(15): 4122-9.

- Mitsiades, N.,V. Poulaki,C. S. Mitsiades,D. A. Koutras andG. P. Chrousos (2001). "Apoptosis induced by FasL and TRAIL/Apo2L in the pathogenesis of thyroid diseases." Trends Endocrinol Metab 12(9): 384-90.

- Mitsiades, C. S.,V. Poulaki,G. Fanourakis,E. Sozopoulos,D. McMillin,Z. Wen,G. Voutsinas,S. Tseleni-Balafouta and N. Mitsiades (2006). "Fas signaling in thyroid carcinomas is diverted from apoptosis to proliferation." Clin Cancer Res 12(12): 3705-12.

- Miyazaki, T. andJ. C. Reed (2001). "A GTP-binding adapter protein couples TRAIL receptors to apoptosis-inducing proteins." Nat Immunol 2(6): 493-500.

- Morgan, C. J., Jacques, C., Savagner, F., Tourmen, Y., Mirebeau, D. P., Malthiery, Y., Reynier, P. (2001) “A conserved N-terminal sequence targets human DAP3 to mitochondria.” Biochem Biophys Res Commun. 280(1):177-81.

- Mukamel, Z. andA. Kimchi (2004). "Death-associated protein 3 localizes to the mitochondria and is involved in the process of mitochondrial fragmentation during cell death." J Biol Chem 279(35): 36732-8.

- Murata, Y.,T. Wakoh,N. Uekawa,M. Sugimoto,A. Asai,T. Miyazaki andM. Maruyama (2006). "Death-associated protein 3 regulates cellular senescence through oxidative stress response." FEBS Lett.

- Myslinski, E.,M. A. Gerard,A. Krol andP. Carbon (2006). "A genome scale location analysis of human Staf/ZNF143-binding sites suggests a widespread role

for human Staf/ZNF143 in mammalian promoters." J Biol Chem 281(52): 39953-62.

- O'Brien, T. W. (2002). "Evolution of a protein-rich mitochondrial ribosome: implications for human genetic disease." Gene 286(1): 73-9.

- O'Brien, T.,B. O'Brien andR. Norman (2005). "Nuclear MRP genes and mitochondrial disease." Gene(In press).

- Palomero, T.,D. T. Odom,J. O'Neil,A. A. Ferrando,A. Margolin,D. S. Neuberg,S. S. Winter,R. S. Larson,W. Li,X. S. Liu,R. A. Young andA. T. Look (2006). "Transcriptional regulatory networks downstream of TAL1/SCL in T-cell acute lymphoblastic leukemia." Blood 108(3): 986-92.

- Regula, K. M.,K. Ens andL. A. Kirshenbaum (2003). "Mitochondria-assisted cell suicide: a license to kill." J Mol Cell Cardiol 35(6): 559-67.

- Sasaki, H.,N. Ide,H. Yukiue,Y. Kobayashi,I. Fukai,Y. Yamakawa andY. Fujii (2004). "Arg and DAP3 expression was correlated with human thymoma stage." Clin Exp Metastasis 21(6): 507-13.

- Savagner, F.,B. Franc,S. Guyetant,P. Rodien,P. Reynier andY. Malthiery (2001). "Defective mitochondrial ATP synthesis in oxyphilic thyroid tumors." J Clin Endocrinol Metab 86(10): 4920-5.

- Savagner F, Mirebeau D, Jacques C, Guyetant S, Morgan C, Franc B, Reynier P, Malthiery Y (2003) “PGC-1-related coactivator and targets are upregulated in thyroid oncocytoma.” Biochem Biophys Res Commun. 310(3):779-84.

- Scarpulla, R. C. (2006). "Nuclear control of respiratory gene expression in mammalian cells." J Cell Biochem 97(4): 673-83.

- Schapira, A. H. (2006). "Mitochondrial disease." Lancet 368(9529): 70-82.

- Sladek, R.,J. A. Bader andV. Giguere (1997). "The orphan nuclear receptor estrogen-related receptor alpha is a transcriptional regulator of the human medium-chain acyl coenzyme A dehydrogenase gene." Mol Cell Biol 17(9): 5400-9.

- Suzuki, T.,M. Terasaki,C. Takemoto-Hori,T. Hanada,T. Ueda,A. Wada andK. Watanabe (2001). "Proteomic analysis of the mammalian mitochondrial ribosome. Identification of protein components in the 28 S small subunit." J Biol Chem 276(35): 33181-95.

- Woo Lee, J.,Y. Hwa Soung,S. Young Kim,S. Woo Nam,W. Sang Park,J. Young Lee,N. Jin Yoo andS. Hyung Lee (2006). "Mutational analysis of proapoptotic death associated protein 3 (DAP3) P-loop domain in common human carcinomas." Acta Oncol 45(4): 489-90.

- Yang, S. H.,E. Jaffray,R. T. Hay andA. D. Sharrocks (2003). "Dynamic interplay of the SUMO and ERK pathways in regulating Elk-1 transcriptional activity." Mol Cell 12(1): 63-74.

- Zhang, N.,W. Shen,R. G. Hawley andM. Lu (1999). "HOX11 interacts with CTF1 and mediates hematopoietic precursor cell immortalization." Oncogene 18(13): 2273-9.

Tables et figures

Table 1 : DAP3 immunohistochemical staining in tumor and normal thyroid tissues.

Class	n	Positive cases			Quantity				Intensity		
					0%	<30%	30-60%	>60%	0	1+	2+
Oncocytic adenoma	33	18	51.43%	***	15	2	11	5	15	15	3
Papillary carcinoma	33	12	36.36%	**	20	7	2	3	20	10	2
Oncocytic carcinoma	9	5	55.55%	*	4	0	2	3	4	5	0
Poorly differentiated carcinoma	3	2	66.66%	*	1	1	1	0	1	1	1
Follicular adenoma	8	0	0.00%		8	0	0	0	8	0	0
Atypical adenoma	5	1	20.00%		4	0	0	1	4	1	0
Follicular carcinoma	5	0	0.00%		5	0	0	0	5	0	0
Anaplastic carcinoma	1	0	0.00%		1	0	0	0	1	0	0
Normal tissue	97	4	4.12%		93	3	0	1	93	3	1

Immunohistochemistry staining of normal thyroid tissue, oncocytic adenoma and carcinoma for the DAP3 protein. The overexpression of the DAP3 protein is revealed by the brown staining in the cytoplasm of the cells. * significant, P<5E-02 by the Fisher's exact test (Benjamini and Hochberg correction). ** significant, P<6E-04. *** significant, P<5.43E-06.

Table 2 : Enriched motifs in promoters

	Motif name	*Motif family*	*Raw score*	*P-values*	
				Promoters	*Cpg*
(A) DAP3 Correlated genes					
	GABPA	ETS	4.83	0.001	0
	ELK1	ETS	1.62	0	0.006
	NRF1		1.61	0.009	0
	ELK4	ETS	-1.32	0	0
	RUNX1	RUNT	-2.66	0.042	0.046
	Hox11-CTF1	HOMEO/CAAT	-3.01	0.041	0.02
(B) Mitochondrial small subunit					
	ERRalpha		3.29	0	0.021
	NRF1		3.14	0.002	0
	TAL1-TCF3	bHLH	2.25	0	0.006
	GABPA	ETS	1.82	0.022	0.012
(C) Mitochondrial big subunit					
	NRF1		9.91	0	0
	ELK4	ETS	2.97	0	0
	GABPA	ETS	2.88	0.001	0.001
	ELK1	ETS	1.99	0	0.005
	PPARG-RXRA	NUCLEAR RECEPTOR	0.598	0.042	0.003
	TP53	P53	-2.74	0.01	0.038

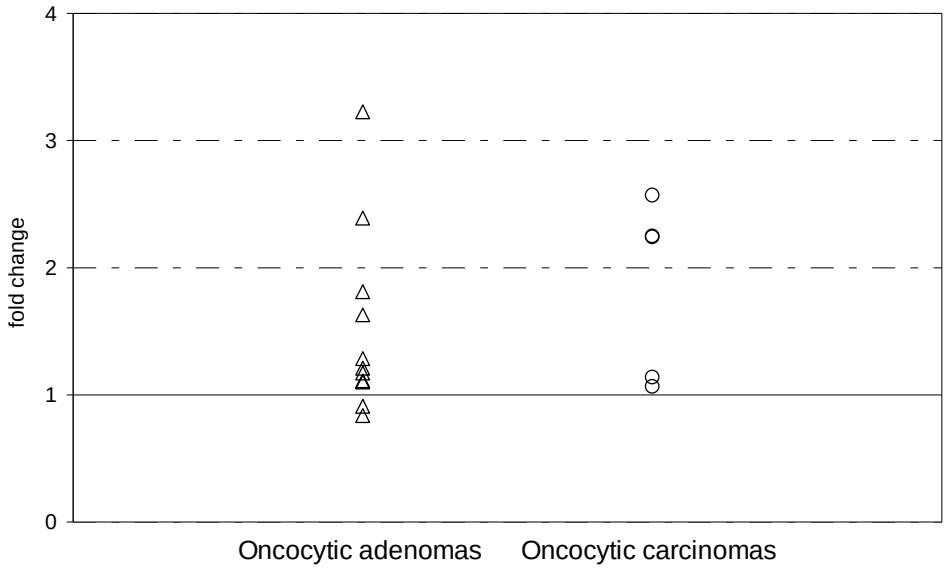
We found enriched TFBS motifs in gene promoter sequences of DAP3 co-regulated genes (A), of the mitochondrial small (B) and large (C) subunits. The raw score and the p-values were computed relatively to a set of human promoters and a set of CpG Island sequences.

Table 3 : transcription factor binding sites in dap3 promoter

motif	*sequence*	*start pos.*	*end pos.*	*strand*	*score*
ELK1	cttacggaaa	-458	-449	+	5.92
	tttcccctcg	345	354	-	5.51
	tgtcagggac	-79	-70	-	5.49
ELK4	tccgggagt	-93	-85	+	4.87
	actggaatt	938	946	+	4.64
	gtttcccct	344	352	-	4.17
ERRalpha	tgtcctcga	-232	-224	-	7.48
	agaccttgt	-270	-262	-	6.41
	tgacattta	-345	-337	-	5.8
	tcagggaca	-77	-69	+	5.02
GABPA	agaggaaggg	898	907	+	7.18
	ctcctccgct	185	194	-	6.99
	ccctcccgct	215	224	-	6.56
	cgcggcaggg	-287	-278	+	6.44
Hox11-CTF1	agggagggagctaa	418	431	+	3.27
	tttcccctcgcaaa	345	358	+	3.11
NRF1	cgcgcgtgcgcc	833	844	+	5.57
	cgcgcgtgcgcc	833	844	-	5.49
	agcgcaggccct	690	701	-	4.75
RUNX1	gaccacaac	-679	-671	-	6.05
	caccacaga	-276	-268	-	4.74
TAL1-TCF3	agaccatctgtc	-399	-388	+	9.93
	accatctgtctg	-397	-386	-	7.18
PPARG-RXRA	ttttggcccttcacatttac	-736	-717	-	2.6
TP53	aaaaacatctctttccatgt	-378	-359	+	2.71

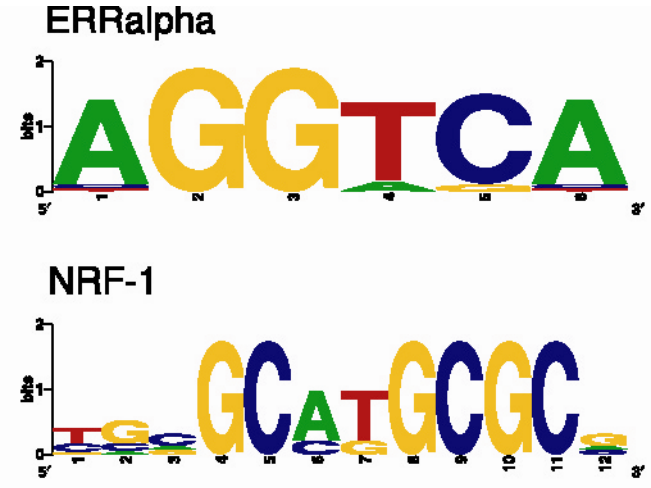
Scores are computed by the POSSUM program with motifs from the JASPAR database and the compiled ERRalpha and NRF1 motifs.

Figure 1: DAP3 mRNA expression levels in oncocytic tumours



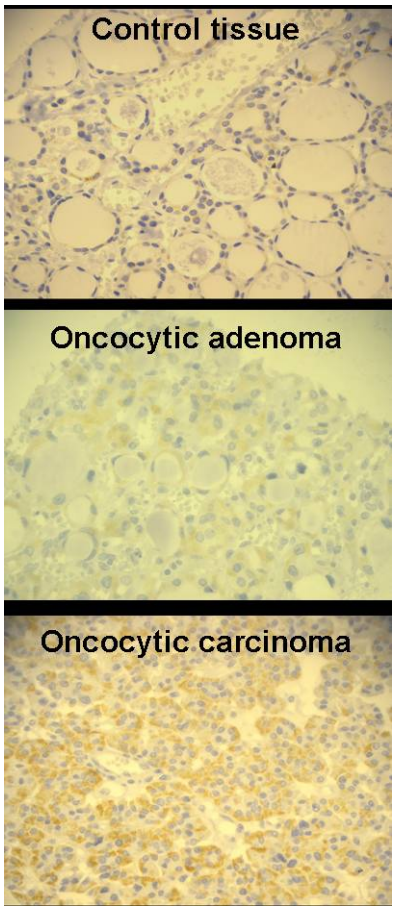
The figure shows the DAP3-expression levels explored by quantitative RT-PCR analysis for twelve thyroid oncocytic adenomas (triangles) and 5 thyroid oncocytic carcinomas (circles). Each level is normalized by the paired normal tissue thus 1 represents iso-expression.

Figure 2: logo motifs of ERRalpha and NRF-1



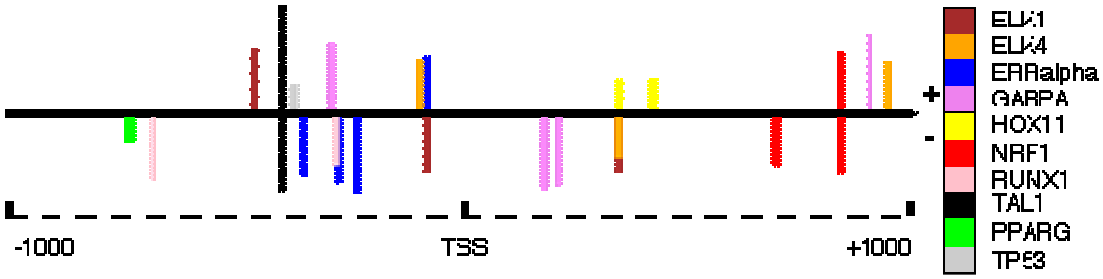
These graphical representation of the ERRalpha and NRF-1 motifs show the 6 and 12 positions of their position-weight matrices. For each position, present DNA bases are proportionally drawn according to their relative frequency. The overall height at a position indicates the sequence conservation.

Figure 3: DAP3 immunohistochemestry staining



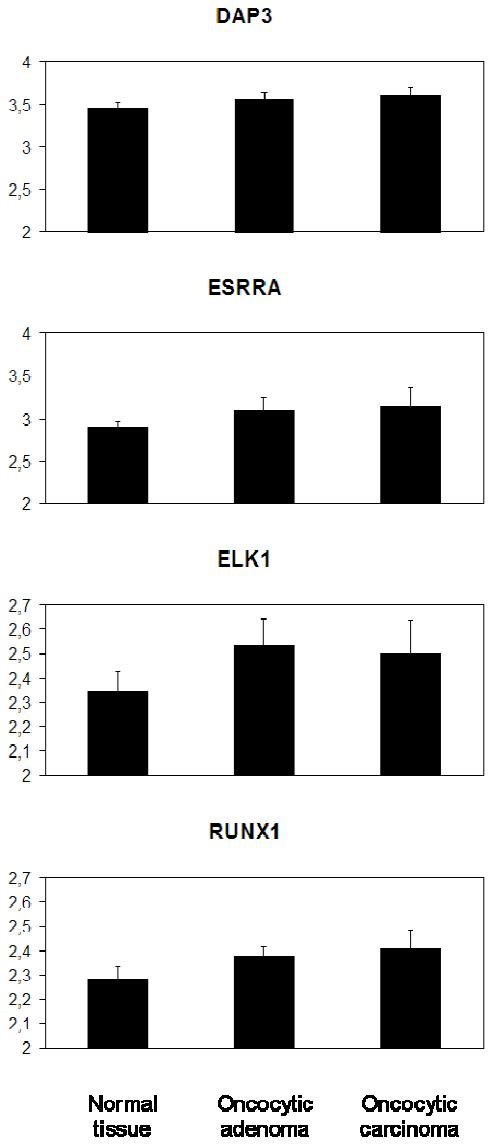
DAP3 protein localisation in thyroid the normal tissue (up), oncocytic adenomas (middle) and oncocytic carcinomas (bottom).

Figure 4: Best Sites for each candidate regulators of the DAP3 gene



Highly scored TFBS by the POSSUM program are shown by colored bars. Each color correspond to a potential regulator and the bar height is proportional to the score. Sites are distributed from -1000 to +1000 base pairs around the transcription start site (TSS).

Figure 5: Gene-expression in an indépendant dataset



We used a public dataset to independently validate the DAP3 over-expression in oncocytic adenomas and carcinomas. Three of the candidate regulators of DAP3 were also significantly over-expressed in oncocytic tumours as compared to the normal tissue ($p<0.05$).

Supplemental table 1: DAP3 co-regulated genes.

Genes correlated with DAP3 in at least 3 public microarray datasets (#ds) are reported in this table. The score is given by the TMM web server.

Symbol	*Name*	*#ds*	*score*
CCT3	Chaperonin containing TCP1, subunit 3 (gamma)	9	0.76
PSMB4	Proteasome (prosome, macropain) subunit, beta type, 4	9	0.73
EPRS	Glutamyl-prolyl-tRNA synthetase	9	0.69
SDHC	Succinate dehydrogenase complex, subunit C, integral membrane protein, 15kDa	8	0.72
SSBP1	Single-stranded DNA binding protein 1	7	0.79
SNRPE		7	0.71
KARS	Lysyl-tRNA synthetase	6	0.88
PSMA6	Proteasome (prosome, macropain) subunit, alpha type, 6	6	0.83
GA17		6	0.81
APEX1	APEX nuclease (multifunctional DNA repair enzyme) 1	6	0.77
PSMD4	Proteasome (prosome, macropain) 26S subunit, non-ATPase, 4	6	0.76
CCT6A	Chaperonin containing TCP1, subunit 6A (zeta 1)	6	0.75
XRCC5	X-ray repair complementing defective repair in Chinese hamster cells 5 (double-strand-break rejoining; Ku autoantigen, 80kDa)	6	0.75
PA2G4		6	0.74
CBX3	Chromobox homolog 3 (HP1 gamma homolog, Drosophila)	6	0.7
PSMA1	Proteasome (prosome, macropain) subunit, alpha type, 1	5	0.85
PSMB1	Proteasome (prosome, macropain) subunit, beta type, 1	5	0.84
UQCRC2	Ubiquinol-cytochrome c reductase core protein II	5	0.83
BTF3	Basic transcription factor 3	5	0.81
PSMA5	Proteasome (prosome, macropain) subunit, alpha type, 5	5	0.81
RABGGTB	Rab geranylgeranyltransferase, beta subunit	5	0.81
AATF	Apoptosis antagonizing transcription factor	5	0.8
PSMD14	Proteasome (prosome, macropain) 26S subunit, non-ATPase, 14	5	0.8
TARS	Threonyl-tRNA synthetase	5	0.8
CCT2	Chaperonin containing TCP1, subunit 2 (beta)	5	0.78
CLNS1A	Chloride channel, nucleotide-sensitive, 1A	5	0.78
CBFB	Core-binding factor, beta subunit	5	0.77
COPS5	COP9 constitutive photomorphogenic homolog subunit 5 (Arabidopsis)	5	0.77
COX7B	Cytochrome c oxidase subunit VIIb	5	0.77
EIF3S1	Eukaryotic translation initiation factor 3, subunit 1 alpha, 35kDa	5	0.77
KIAA0117		5	0.77
PWP1	PWP1 homolog (S. cerevisiae)	5	0.77
PSMA3	Proteasome (prosome, macropain) subunit, alpha type, 3	5	0.75
DKC1	Dyskeratosis congenita 1, dyskerin	5	0.74
NAP1L1		5	0.74
SNRPF	Small nuclear ribonucleoprotein polypeptide F	5	0.74
CCT4	Chaperonin containing TCP1, subunit 4 (delta)	5	0.71
DDX21	DEAD (Asp-Glu-Ala-Asp) box polypeptide 21	5	0.71
EIF3S7	Eukaryotic translation initiation factor 3, subunit 7 zeta, 66/67kDa	5	0.71
SHFM1	Split hand/foot malformation (ectrodactyly) type 1	5	0.71
SCAMP3	Secretory carrier membrane protein 3	5	0.7
GSPT1	G1 to S phase transition 1	5	0.69
NICE-4		5	0.68
RFC4	Replication factor C (activator 1) 4, 37kDa	5	0.67
FLJ10326		5	0.66
PSMA2	Proteasome (prosome, macropain) subunit, alpha type, 2	4	0.86
DKFZP564M182		4	0.84
POLR2B	Polymerase (RNA) II (DNA directed) polypeptide B, 140kDa	4	0.83
UQCRH	Ubiquinol-cytochrome c reductase hinge protein	4	0.83
MCTS1	Malignant T cell amplified sequence 1	4	0.82
POLR2F	Polymerase (RNA) II (DNA directed) polypeptide F	4	0.82
SET	SET translocation (myeloid leukemia-associated)	4	0.81
C1QBP	Complement component 1, q subcomponent binding protein	4	0.79
COPS3	COP9 constitutive photomorphogenic homolog subunit 3 (Arabidopsis)	4	0.79
PPP2R5E	Protein phosphatase 2, regulatory subunit B (B56), epsilon isoform	4	0.79
PSMA4	Proteasome (prosome, macropain) subunit, alpha type, 4	4	0.79
PSMC1	Proteasome (prosome, macropain) 26S subunit, ATPase, 1	4	0.79
AK074970		4	0.78
COX7A2L	Cytochrome c oxidase subunit VIIa polypeptide 2 like	4	0.78
EIF4A1	Eukaryotic translation initiation factor 4A, isoform 1	4	0.78
ETFA	Electron-transfer-flavoprotein, alpha polypeptide (glutaric aciduria II)	4	0.78
H2AFY		4	0.78
SFRS2	Splicing factor, arginine/serine-rich 2	4	0.78
COX7C	Cytochrome c oxidase subunit VIIc	4	0.77
GART	Phosphoribosylglycinamide formyltransferase, phosphoribosylglycinamide synthetase, phosphoribosylaminoimidazole synthetase	4	0.77
HSPA4	Heat shock 70kDa protein 4	4	0.77
M11S1		4	0.77
PTD004		4	0.77
VBP1	Von Hippel-Lindau binding protein 1	4	0.77
CCT7	Chaperonin containing TCP1, subunit 7 (eta)	4	0.76
KIAA0179	KIAA0179	4	0.76
NME1	Non-metastatic cells 1, protein (NM23A) expressed in	4	0.76
RBMX		4	0.76
TFAM	Transcription factor A, mitochondrial	4	0.76
TRIP15		4	0.76
DDX18	DEAD (Asp-Glu-Ala-Asp) box polypeptide 18	4	0.75
FNBP3		4	0.75
FUSIP1		4	0.75
PPP1CC	Protein phosphatase 1, catalytic subunit, gamma isoform	4	0.75

SNRPA1	Small nuclear ribonucleoprotein polypeptide A'	4	0.75
SUMO1	SMT3 suppressor of mif two 3 homolog 1 (S. cerevisiae)	4	0.75
TCEA1		4	0.75
TRAP1	TNF receptor-associated protein 1	4	0.75
CCT5	Chaperonin containing TCP1, subunit 5 (epsilon)	4	0.74
EIF2S3	Eukaryotic translation initiation factor 2, subunit 3 gamma, 52kDa	4	0.74
MAPRE1	Microtubule-associated protein, RP/EB family, member 1	4	0.74
PRDX3	Peroxiredoxin 3	4	0.74
GDI2	GDP dissociation inhibitor 2	4	0.73
MTIF2	Mitochondrial translational initiation factor 2	4	0.73
RBBP4	Retinoblastoma binding protein 4	4	0.73
SFRS10	Splicing factor, arginine/serine-rich 10 (transformer 2 homolog, Drosophila)	4	0.73
SNRPD3	Small nuclear ribonucleoprotein D3 polypeptide 18kDa	4	0.73
SRPK1	SFRS protein kinase 1	4	0.73
YARS	Tyrosyl-tRNA synthetase	4	0.73
GNPAT	Glyceronephosphate O-acyltransferase	4	0.72
IMMT	Inner membrane protein, mitochondrial (mitofilin)	4	0.72
MAK3		4	0.72
NPM1	Nucleophosmin (nucleolar phosphoprotein B23, numatrin)	4	0.72
NSEP1		4	0.72
XPO1	Exportin 1 (CRM1 homolog, yeast)	4	0.72
ENSA		4	0.71
LRPPRC	Leucine-rich PPR-motif containing	4	0.71
NASP	Nuclear autoantigenic sperm protein (histone-binding)	4	0.71
NUP153		4	0.71
PDCD2		4	0.71
DEK	DEK oncogene (DNA binding)	4	0.7
DKFZP547E1010		4	0.7
HNRPAB		4	0.7
PSMB7	Proteasome (prosome, macropain) subunit, beta type, 7	4	0.7
SRP68	Signal recognition particle 68kDa	4	0.7
USP10	Ubiquitin specific peptidase 10	4	0.7
ADAR	Adenosine deaminase, RNA-specific	4	0.69
CKS1B	CDC28 protein kinase regulatory subunit 1B	4	0.69
METAP2	Methionyl aminopeptidase 2	4	0.69
MRPL9	Mitochondrial ribosomal protein L9	4	0.69
SSR2	Signal sequence receptor, beta (translocon-associated protein beta)	4	0.69
XTP2		4	0.69
H2AFV		4	0.68
SSB	Sjogren syndrome antigen B (autoantigen La)	4	0.68
HSPA9B		4	0.67
TCFL1		4	0.64
C14ORF166	Chromosome 14 open reading frame 166	3	0.92
EIF3K		3	0.88
EIF3S3	Eukaryotic translation initiation factor 3, subunit 3 gamma, 40kDa	3	0.88
GTF3A	General transcription factor IIIA	3	0.87
AF271775		3	0.86
DRG1	Developmentally regulated GTP binding protein 1	3	0.86
MRPS18B		3	0.86
SOD1	Superoxide dismutase 1, soluble (amyotrophic lateral sclerosis 1 (adult))	3	0.85
APG12L		3	0.83
ATP5J	ATP synthase, H+ transporting, mitochondrial F0 complex, subunit F6	3	0.83
C6ORF49	Chromosome 6 open reading frame 49	3	0.83
COX5A	Cytochrome c oxidase subunit Va	3	0.83
HNRPF	Heterogeneous nuclear ribonucleoprotein F	3	0.83
NDUFB5	NADH dehydrogenase (ubiquinone) 1 beta subcomplex, 5, 16kDa	3	0.83
PPP6C	Protein phosphatase 6, catalytic subunit	3	0.83
RPL4		3	0.83
TXNL1	Thioredoxin-like 1	3	0.83
ESD	Esterase D/formylglutathione hydrolase	3	0.82
EXOSC8	Exosome component 8	3	0.82
HSA9761		3	0.82
HSPC111	Hypothetical protein HSPC111	3	0.82
LOC91137		3	0.82
POLD2	Polymerase (DNA directed), delta 2, regulatory subunit 50kDa	3	0.82
ATP6V0B	ATPase, H+ transporting, lysosomal 21kDa, V0 subunit b	3	0.81
CCND3	Cyclin D3	3	0.81
CCT8	Chaperonin containing TCP1, subunit 8 (theta)	3	0.81
FBL	Fibrillarin	3	0.81
HNRPC		3	0.81
MFN2	Mitofusin 2	3	0.81
MRPL16	Mitochondrial ribosomal protein L16	3	0.81
NDUFV2	NADH dehydrogenase (ubiquinone) flavoprotein 2, 24kDa	3	0.81
NIPA2	Non imprinted in Prader-Willi/Angelman syndrome 2	3	0.81
PSMC4	Proteasome (prosome, macropain) 26S subunit, ATPase, 4	3	0.81
PSMC6		3	0.81
SFRS3	Splicing factor, arginine/serine-rich 3	3	0.81
UBE3A	Ubiquitin protein ligase E3A (human papilloma virus E6-associated protein, Angelman syndrome)	3	0.81
BCL2A1	BCL2-related protein A1	3	0.8
COX4I1	Cytochrome c oxidase subunit IV isoform 1	3	0.8
NCL	Nucleolin	3	0.8
NSMAF	Neutral sphingomyelinase (N-SMase) activation associated factor	3	0.8
PANK2	Pantothenate kinase 2 (Hallervorden-Spatz syndrome)	3	0.8
SLBP	Stem-loop (histone) binding protein	3	0.8
SRRM1	Serine/arginine repetitive matrix 1	3	0.8
TCERG1	Transcription elongation regulator 1	3	0.8
UBE2N		3	0.8
ABCE1	ATP-binding cassette, sub-family E (OABP), member 1	3	0.79
API5	Apoptosis inhibitor 5	3	0.79
CSNK2A1		3	0.79
CTPS	CTP synthase	3	0.79
DPM1	Dolichyl-phosphate mannosyltransferase polypeptide 1, catalytic subunit	3	0.79
GLUD1	Glutamate dehydrogenase 1	3	0.79

HNRPA3		Heterogeneous nuclear ribonucleoprotein A3	3	0.79
HSPD1			3	0.79
LSM3		LSM3 homolog, U6 small nuclear RNA associated (S. cerevisiae)	3	0.79
MGC15396			3	0.79
NHP2L1		NHP2 non-histone chromosome protein 2-like 1 (S. cerevisiae)	3	0.79
PCMT1		Protein-L-isoaspartate (D-aspartate) O-methyltransferase	3	0.79
SNRPD2		Small nuclear ribonucleoprotein D2 polypeptide 16.5kDa	3	0.79
TIAL1		TIA1 cytotoxic granule-associated RNA binding protein-like 1	3	0.79
TXNDC			3	0.79
ZRF1		Zuotin related factor 1	3	0.79
ATP5C1		ATP synthase, H+ transporting, mitochondrial F1 complex, gamma polypeptide 1	3	0.78
BC066990			3	0.78
CPSF6		Cleavage and polyadenylation specific factor 6, 68kDa	3	0.78
ILF3		Interleukin enhancer binding factor 3, 90kDa	3	0.78
MRPL3		Mitochondrial ribosomal protein L3	3	0.78
PAICS			3	0.78
SLC25A6		Solute carrier family 25 (mitochondrial carrier; adenine nucleotide translocator), member 6	3	0.78
SLC35B1		Solute carrier family 35, member B1	3	0.78
SUMO2			3	0.78
VDP		Vesicle docking protein p115	3	0.78
APPBP1		Amyloid beta precursor protein binding protein 1	3	0.77
C14ORF32		Chromosome 14 open reading frame 32	3	0.77
C7ORF28A		Chromosome 7 open reading frame 28A	3	0.77
DDX1		DEAD (Asp-Glu-Ala-Asp) box polypeptide 1	3	0.77
EIF2S2			3	0.77
EIF3S10			3	0.77
FXR1		Fragile X mental retardation, autosomal homolog 1	3	0.77
HNRPU		Heterogeneous nuclear ribonucleoprotein U (scaffold attachment factor A)	3	0.77
IMPDH2		IMP (inosine monophosphate) dehydrogenase 2	3	0.77
KIAA0052			3	0.77
KPNA1		Karyopherin alpha 1 (importin alpha 5)	3	0.77
MRPS31		Mitochondrial ribosomal protein S31	3	0.77
PRKDC		Protein kinase, DNA-activated, catalytic polypeptide	3	0.77
PTMA		Prothymosin, alpha (gene sequence 28)	3	0.77
RARS		Arginyl-tRNA synthetase	3	0.77
SND1		Staphylococcal nuclease domain containing 1	3	0.77
SRP72		Signal recognition particle 72kDa	3	0.77
YME1L1		YME1-like 1 (S. cerevisiae)	3	0.77
BCLAF1			3	0.76
CKS2		CDC28 protein kinase regulatory subunit 2	3	0.76
ME2		Malic enzyme 2, NAD(+)-dependent, mitochondrial	3	0.76
MRPS17		Mitochondrial ribosomal protein S17	3	0.76
NOL1		Nucleolar protein 1, 120kDa	3	0.76
OAT		Ornithine aminotransferase (gyrate atrophy)	3	0.76
PSMD13		Proteasome (prosome, macropain) 26S subunit, non-ATPase, 13	3	0.76
RAN			3	0.76
RG9MTD1		RNA (guanine-9-) methyltransferase domain containing 1	3	0.76
SNRPA		Small nuclear ribonucleoprotein polypeptide A	3	0.76
UBA2			3	0.76
ZNF9			3	0.76
AHCY		S-adenosylhomocysteine hydrolase	3	0.75
AHSA1		AHA1, activator of heat shock 90kDa protein ATPase homolog 1 (yeast)	3	0.75
ATIC		5-aminoimidazole-4-carboxamide ribonucleotide formyltransferase/IMP cyclohydrolase	3	0.75
BMS1L		BMS1-like, ribosome assembly protein (yeast)	3	0.75
BUB3		BUB3 budding uninhibited by benzimidazoles 3 homolog (yeast)	3	0.75
CAPZA1		Capping protein (actin filament) muscle Z-line, alpha 1	3	0.75
COG2		Component of oligomeric golgi complex 2	3	0.75
CPOX		Coproporphyrinogen oxidase	3	0.75
DHX9		DEAH (Asp-Glu-Ala-His) box polypeptide 9	3	0.75
EIF2B1		Eukaryotic translation initiation factor 2B, subunit 1 alpha, 26kDa	3	0.75
G3BP			3	0.75
LOC134218			3	0.75
NDUFS2		NADH dehydrogenase (ubiquinone) Fe-S protein 2, 49kDa (NADH-coenzyme Q reductase)	3	0.75
NUDT5		Nudix (nucleoside diphosphate linked moiety X)-type motif 5	3	0.75
NUP50		Nucleoporin 50kDa	3	0.75
PGK1		Phosphoglycerate kinase 1	3	0.75
PNN		Pinin, desmosome associated protein	3	0.75
PP591			3	0.75
PSMA7		Proteasome (prosome, macropain) subunit, alpha type, 7	3	0.75
PSMC5		Proteasome (prosome, macropain) 26S subunit, ATPase, 5	3	0.75
PSME2			3	0.75
RNF138			3	0.75
SLC25A5		Solute carrier family 25 (mitochondrial carrier; adenine nucleotide translocator), member 5	3	0.75
SNRPD1		Small nuclear ribonucleoprotein D1 polypeptide 16kDa	3	0.75
ST13		Suppression of tumorigenicity 13 (colon carcinoma) (Hsp70 interacting protein)	3	0.75
SUCLA2		Succinate-CoA ligase, ADP-forming, beta subunit	3	0.75
YWHAQ		Tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, theta polypeptide	3	0.75
	39331		3	0.74
ADIPOR1		Adiponectin receptor 1	3	0.74
AF116699			3	0.74
AK2		Adenylate kinase 2	3	0.74
CUL4A		Cullin 4A	3	0.74
DLD		Dihydrolipoamide dehydrogenase	3	0.74
EEF1E1		Eukaryotic translation elongation factor 1 epsilon 1	3	0.74
FUBP1		Far upstream element (FUSE) binding protein 1	3	0.74
HBS1L		HBS1-like (S. cerevisiae)	3	0.74
NCBP2		Nuclear cap binding protein subunit 2, 20kDa	3	0.74
PABPC1		Poly(A) binding protein, cytoplasmic 1	3	0.74
PRPF4B		PRP4 pre-mRNA processing factor 4 homolog B (yeast)	3	0.74
RNPS1			3	0.74
TEBP			3	0.74
TXNDC9		Thioredoxin domain containing 9	3	0.74
CCNH		Cyclin H	3	0.73

EIF3S9	Eukaryotic translation initiation factor 3, subunit 9 eta, 116kDa	3	0.73
FLJ20303		3	0.73
GTF3C3	General transcription factor IIIC, polypeptide 3, 102kDa	3	0.73
HNRPR	Heterogeneous nuclear ribonucleoprotein R	3	0.73
LEREPO4		3	0.73
NONO	Non-POU domain containing, octamer-binding	3	0.73
PAI-RBP1		3	0.73
PES1	Pescadillo homolog 1, containing BRCT domain (zebrafish)	3	0.73
PPP1CB		3	0.73
SCYE1	Small inducible cytokine subfamily E, member 1 (endothelial monocyte-activating)	3	0.73
SRP9		3	0.73
SSR1	Signal sequence receptor, alpha (translocon-associated protein alpha)	3	0.73
SYNCRIP		3	0.73
TNPO1	Transportin 1	3	0.73
CAD	Carbamoyl-phosphate synthetase 2, aspartate transcarbamylase, and dihydroorotase	3	0.72
CGI-48		3	0.72
DYT1		3	0.72
EIF3S2	Eukaryotic translation initiation factor 3, subunit 2 beta, 36kDa	3	0.72
ERP70		3	0.72
HNRPA2B1	Heterogeneous nuclear ribonucleoprotein A2/B1	3	0.72
SMAD2		3	0.72
SMARCA5	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily a, member 5	3	0.72
SMARCE1	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily e, member 1	3	0.72
SNRPG		3	0.72
UBE2V2	Ubiquitin-conjugating enzyme E2 variant 2	3	0.72
XPOT	Exportin, tRNA (nuclear export receptor for tRNAs)	3	0.72
C10ORF7		3	0.71
GHITM	Growth hormone inducible transmembrane protein	3	0.71
MEP50		3	0.71
MTHFD2	Methylenetetrahydrofolate dehydrogenase (NADP+ dependent) 2, methenyltetrahydrofolate cyclohydrolase	3	0.71
PPM1G	Protein phosphatase 1G (formerly 2C), magnesium-dependent, gamma isoform	3	0.71
RAD21	RAD21 homolog (S. pombe)	3	0.71
RCN1	Reticulocalbin 1, EF-hand calcium binding domain	3	0.71
SFRS9	Splicing factor, arginine/serine-rich 9	3	0.71
SNRPB	Small nuclear ribonucleoprotein polypeptides B and B1	3	0.71
TBCA	Tubulin folding cofactor A	3	0.71
ACTL6A	Actin-like 6A	3	0.7
AFURS1		3	0.7
ANAPC7	Anaphase promoting complex subunit 7	3	0.7
APTX	Aprataxin	3	0.7
ATP5G3	ATP synthase, H+ transporting, mitochondrial F0 complex, subunit C3 (subunit 9)	3	0.7
BC026067		3	0.7
C1ORF37	Chromosome 1 open reading frame 37	3	0.7
NOLC1	Nucleolar and coiled-body phosphoprotein 1	3	0.7
PAK2	P21 (CDKN1A)-activated kinase 2	3	0.7
PPP2R5C	Protein phosphatase 2, regulatory subunit B (B56), gamma isoform	3	0.7
PTDSS1	Phosphatidylserine synthase 1	3	0.7
REA		3	0.7
RIC-8		3	0.7
SF3B1	Splicing factor 3b, subunit 1, 155kDa	3	0.7
SKB1		3	0.7
SNRPC	Small nuclear ribonucleoprotein polypeptide C	3	0.7
TDG		3	0.7
UBL4		3	0.7
VPS26		3	0.7
FLJ20729		3	0.69
RPL24		3	0.69
RPP40	Ribonuclease P 40kDa subunit	3	0.69
HIP2	Huntingtin interacting protein 2	3	0.68
SF3A3	Splicing factor 3a, subunit 3, 60kDa	3	0.68
WTAP	Wilms tumor 1 associated protein	3	0.68
DHX40	DEAH (Asp-Glu-Ala-His) box polypeptide 40	3	0.67
GRPEL1	GrpE-like 1, mitochondrial (E. coli)	3	0.67
RBM14		3	0.67
ANP32E	Acidic (leucine-rich) nuclear phosphoprotein 32 family, member E	3	0.66
FLJ14668	Hypothetical protein FLJ14668	3	0.66
IPO9	Importin 9	3	0.66
POLR3C	Polymerase (RNA) III (DNA directed) polypeptide C (62kD)	3	0.58
VMP		3	-0.52
SFMBT2	Scm-like with four mbt domains 2	3	-0.57
ATXN1	Ataxin 1	3	-0.58
AB020684		3	-0.79

Supplemental table 2: GO clusters

Cluster	Genes	P-value	GO terms
1	22	9.70E-25	RNA processing
2	11	6.56E-14	ligase activity
3	51	5.81E-34	protein metabolism
4	17	1.07E-17	oxidoreductase activity
5	18	2.48E-07	biopolymer metabolism
		1.80E-04	catalytic activity
6	187	0	intracellular
7	15	0	protein folding
		0	unfolded protein binding
		1.09E-10	ATP binding
8	66	8.25E-18	intracellular membrane-bound organelle
9	18	3.50E-07	protein complex
10	10	1.15E-05	RNA binding
11	12	3.39E-14	carrier activity
		3.39E-14	electron transport
		4.66E-07	transport
12	21	1.77E-15	mitochondrion
13	111	2.41E-38	primary metabolism
14	14	7.12E-13	protein biosynthesis
		5.35E-06	nucleic acid binding
15	16	1.80E-20	transcription
17	26	7.40E-08	catalytic activity
		3.10E-07	cytoplasm
18	10	9.67E-13	intracellular transport
		7.42E-07	cell organization and biogenesis
		1.76E-06	membrane
19	97	0	macromolecule metabolism
		2.95E-21	binding
20	17	1.51E-14	ribonucleoprotein complex
21	14	5.17E-09	cellular biosynthesis
		6.48E-05	nucleobase, nucleoside, nucleotide and nucleic acid metabolism
		7.07E-03	catalytic activity
22	38	1.34E-20	nucleobase, nucleoside, nucleotide and nucleic acid metabolism
		1.09E-18	nucleus

The 330 DAP3 co-regulated genes are clustered according to their GO annotations. Each cluster is composed of at least 10 genes which share at least 80% of their annotations. Only significant clusters were retained at 5% risk computed by the hyper-geometric distribution with Bonferroni correction.

3.4.3 Discussion

Les oncocytomes thyroïdiens présentent une surexpression des sous-unités de la chaine respiratoire alors que l'activité des complexes n'est pas déficiente (Savagner *et al.*, 2001). La traduction protéique doit alors être efficace. Le gène et la protéine DAP3, ayant une fonction dans la mort cellulaire, sont surexprimés dans les tumeurs oncocytaires de la thyroïde. Les tumeurs oncocytaires sont caractérisées par peu d'apoptose, la surexpression de DAP3 doit être impliquée dans un autre processus. La protéine DAP3 étant un constituant du mitoribosome, elle doit avoir un rôle dans la traduction mitochondriale. Sa surexpression doit être liée à la biogenèse mitochondriale. Les oncocytomes thyroïdiens pourront servir de modèle à l'étude précise du rôle de DAP3 sur la traduction.

L'analyse bioinformatique des séquences régulatrices associées à DAP3 a révélé le rôle potentiel de 10 facteurs transcriptionnels. Ils sont principalement impliqués dans la tumorigenèse (ELK1, ELK4, RUNX1 and HOX11-CTF1) et la biogenèse mitochondriale (NRF1, GABPA). La redondance des résultats obtenus par l'analyse de 337 gènes co-régulés ou des gènes du mitoribosome nous permettent une validation croisée bioinformatiques. De plus, ces analyses sont en partie validées par un jeu de données indépendant. Des validations expérimentales plus avancées pourraient utiliser les technologies de mutation, transfection ou d'immuno-précipitation pour évaluer les interactions ADN-protéines impliquées dans les régulations.

DAP3 peut aussi être considérée comme un lien entre l'homéostasie mitochondriale et le processus de tumorigenèse. Si une physiologie mitochondriale correcte et augmentée est requise pour la tumorigenèse, on peut s'attendre à un défaut de l'une des protéines du mitoribosome qui peut stopper la tumorigenèse et enclencher la mort cellulaire.

4 Discussion

Nous avons abordé l'étude des données d'expression à haut débit pour la classification des tumeurs différenciées de la thyroïde. Les technologies à haut débit nous ont permis de classer les lésions de la thyroïde les unes par rapport aux autres avec une prise en compte globale des gènes spécifiques. La recherche de marqueurs spécifiques à chaque type de lésion à été possible grâce à l'intégration de plus de 90% des lésions du même organe. Les applications industrielles relevant de ces résultats seront mise en place grâce à l'invention déposée au bureau européen des brevets. La détermination précise de la nature des T-UM montre l'intérêt d'une approche pan-génomique associée à une forte collaboration avec un anatomopathologiste, ainsi qu'à des méthodes statistiques adaptées, pour le diagnostic clinique. Il nous a aussi été possible d'utiliser ces données massives ainsi que des données publiques, d'expression et de séquences, pour raisonner au niveau d'un gène particulier et de sa protéine, DAP3.

Les études des tissus thyroïdiens par biopuces se sont intéressées soit à la séparation des tissus bénins et malins, soit à la sous classification d'un type tumoral en fonction de ses mutations spécifiques. Elles incluent de 2 à 6 classes de tissus au maximum. Les marqueurs révélés par ces études ne peuvent donc pas être considérés comme parfaitement spécifiques. Les algorithmes de prédiction automatique utilisés souffrent de la même limite même s'ils ont pu se montrer très efficaces. Nous avons intégré à nos expérimentations sur biopuces plus de 90% des lésions thyroïdiennes différenciées représentant 11 classes de lésions. Grâce à des méthodes statistiques adaptées nous avons établi une classification fonctionnelle globale et simultanée de toutes les pathologies de cet organe. Les marqueurs sélectionnés dans ce jeu de données pour chaque classe font partie du brevet européen que nous avons déposé. Ils permettent de classer automatiquement et de façon précise les échantillons. Cette application est destinée aux cliniciens mais une étape de validation à grande échelle sur cytoponction sera nécessaire pour une application de routine.

Les T-UM ont toujours donné des difficultés aux spécialistes. La détermination d'un diagnostic malignité étant impossible, cela peut entrainer des chirurgies de précaution qui ne seront pas forcément justifiées. Nous avons résolu la classification de ces tumeurs en analysant les données transcriptomiques des 12 classes de tissus thyroïdiens. Nous trouvons que 80% des cas T-UM sont liés aux PTCs alors qu'aucun cas n'a été associé aux FTCs. Néanmoins, l'intégration de plus de cas FTC donnerait des résultats plus puissants. Nous montrons et confirmons qu'il existe des marqueurs spécifiques à ces 2 types de carcinomes. Cette spécificité

biologique prend en compte toutes les autres pathologies différenciées comme contrôle, ainsi que le tissu sain. Nous émettons l'hypothèse que les T-UM sont un état précoce de cancers. Les gènes que nous trouvons spécifiquement exprimés dans ces tumeurs sont un point de départ pour la compréhension du phénomène de transformation des nodules en PTCs. Cependant, l'évolution lente de ces tumeurs fait qu'il est difficile d'obtenir le suivi à 10-15 ans des patients pour s'assurer du potentiel évolutif des tumeurs en cancers.

Par ailleurs, les analyses de données produites à haut débit nous ont permis de mettre en évidence le rôle potentiel dans la traduction mitochondriale de la protéine DAP3 du mitoribosome, aussi impliquée dans la mort cellulaire. Nous avons montré que DAP3 est impliquée dans la traduction mitochondriale. L'analyse bioinformatique des données génomiques et fonctionnelles révèlent une implication directe du gène dans la tumorigenèse et la biogenèse mitochondriale. La relation entre ces deux fonctions est explorée au laboratoire.

Le design imposé des expérimentations nous a posé des problèmes lors de l'utilisation des méthodes statistiques. Dans le cas idéal, les mesures effectuées doivent suivre une loi normale et avoir des variances égales entre les gènes. Cela implique que les groupes à comparer soient au mieux de même taille. Malheureusement, aucune de ces 3 conditions n'étaient complètement respectée. Les tests statistiques sont plutôt robustes face à des distributions non normales (Kariya *et al*., 1986). La non-égalité des variances est prise en compte dans des versions alternatives des tests classiques. Le déséquilibre dans la taille des groupes pose plus de problèmes, il peut fortement influer sur le nombre de gènes significatifs (Yang *et al.*, 2007). Les tests permutés se sont souvent révélé les mieux adaptés à nos données. Ainsi, il est nécessaire de pouvoir adapter les méthodes et tests statistiques au jeu de données étudié. Il n'est pas toujours possible de modifier le code des programmes ou d'implémenter des méthodes compliquées. Le choix existe dans les méthodes et les plus intuitives se montrent parfois les meilleures (Dudoit *et al*., 2002 ; Wessels *et al*., 2005).

L'étude des tumeurs humaines nous ont imposé une vue statique des phénomènes moléculaires spécifiques aux cellules thyroïdiennes. Nous avons alors réalisé des études comparatives principalement de type différentiel. Sans expérimentation dédiées, nous n'avons pas pu établir d'études des réseaux de régulation impliqués dans ces pathologies. Les données génomiques et les méthodes bioinformatiques, avec validation croisée, nous ont permis de trouver les régulateurs potentiels de DAP3. Nous avons utilisé des méthodes très sélectives (plusieurs références simultanées, seuils faibles, validation croisée) mais des validations expérimentales supplémentaires sont en cours.

Nous allons nous intéresser de façon plus précise aux adénomes micro-folliculaires et à leur lien significatif avec les tumeurs oncocytaires. Leur signature spécifique en comparaison des autres adénomes et carcinomes laisse penser qu'ils pourraient suivre une voie de tumorigenèse différente des adénomes macro-folliculaires. Nous allons aussi nous intéresser à la dynamique de tumorigenèse dans les oncocytomes. Nous allons étudier et rechercher des gènes impliqués dans la progression cellulaire et mitochondriale pour découvrir des motifs continus.

Les régulations moléculaires directes des pathologies seront étudiées à l'aide d'autres expérimentations dédiées. Nous réalisons en collaboration avec la plateforme Transcriptome de la Génopole Ouest l'étude de facteurs de transcription d'intérêt par immuno-précipitation de la chromatine sur biopuces (ChIP-chip). De plus, un nouveau projet est en cours pour étudier l'influence des microARNs sur tous les types de tumeurs thyroïdiennes. Cette vue dynamique des pathologies de la thyroïde nous permettra de comprendre les relations de progression cellulaires et mitochondriales. Un objectif sera de pouvoir agir sur les nœuds de ces réseaux pour réprimer la tumorigenèse.

L'utilisation de toutes les données publiques pourra nous servir à l'amélioration des outils de prédiction automatique. Il faudra d'abord résoudre les problèmes de normalisation des données qui empêche toute analyse transversale. Il faudra aussi considérer les méthodes qualitatives qui contournent ce problème. Les méthodes multivariées de sélection de gènes devraient induire de meilleures performances.

Nous allons commencer un projet de validation à grande échelle des signatures prédictives associées au brevet européen. Il s'agit d'utiliser 1000 échantillons sur 5 sites nationaux (Tours, Nancy, Paris, Angers et Poitiers) et de mettre en place une biopuce dédiée. Après cette étape, une application industrielle sera possible dans une société de biotechnologie.

En conclusion, l'analyse des données d'expression pan-génomiques de 166 tissus thyroïdiens nous a permis d'établir la classification fonctionnelle de 90% des pathologies de la thyroïde. Les marqueurs que nous avons définis nous permettent d'effectuer des classements automatiques précis. Ces méthodes de diagnostic sont protégées par un brevet européen. Nous avons résolu la classification de la majorité des T-UM en cancer papillaires. L'analyse associée des données transcriptomiques et des séquences biologiques nous ont permis de déduire les meilleurs candidats à la régulation d'un gène particulier impliqué dans la tumorigenèse et la biogenèse mitochondriale. Dans le futur, les analyses des tumeurs de la thyroïde pourront intégrer des données d'interaction (micro ARNs, ChIP-chip) ainsi que des données génomiques

(Comparative Genome Hybridization, CGH, et Single nucleotide polymorphism, SNP) pour mieux prendre en compte les différences individuelles au sein des groupes de patients.

5 Matériels et Méthodes

5.1 Tissus biologiques

Les échantillons biologiques des 2 premiers articles ont été obtenus à partir de 132 tumeurs thyroïdiennes humaines et 34 tissus de contrôle. Les tumeurs comportaient 26 adénomes macrofolliculaires, 17 adénomes microfolliculaires et 10 T-UMs. Ces derniers étaient définis par leurs spécificités nucléaires, et la présence de modifications cellulaires et vasculaires ; ils ne montraient pas clairement d'invasion capsulaire ou vasculaire. De plus, nous avons utilisé 24 adénomes provenant du plus grand nodule de goitres multi nodulaire. Ils présentaient tous une architecture macrofolliculaire sauf 1 échantillon microfolliculaire. Nous avons aussi examiné 30 adénomes oncocytaires et 5 adénomes oncocytaires atypiques définis sur les mêmes critères que les adénomes non oncocytaires. Nous avons inclus 3 types de carcinomes : folliculaires (3 cas), oncocytaires (4 cas) et papillaires (13 cas). Les échantillons de contrôle étaient 24 tissus sains, 5 thyroïdites auto-immunes et 5 maladies de Grave.

Les diagnostics ont été faits en accord avec la classification de l'Organisation Mondiale de la Santé (DeLellis *et al*., 2004). Quatre-vingt sept échantillons anonymes ont été obtenus à l'hôpital Ambroise Paré (APHP, Boulogne sur Seine, France), et 79 échantillons anonymes au C.H.U. d'Angers, France.

5.2 Méthodes pour déterminer la taille des groupes :

Nous disposons des données d'expression de 258 gènes pour 166 prélèvements thyroïdiens. Nous pouvons calculer la variance observée de chaque gène dans chacun des 9 groupes de tumeurs. L'écart type observé de la population peut aussi être calculé pour évaluer l'effet à prendre en compte. Nous utilisons la formule suivante de calcul de la taille des groupes pour le test t (Seo *et al.*, 2006) :

$$N = 2\left\{\frac{s(z_{1-\alpha} + z_{1-\beta})}{\mathrm{ES}}\right\}^2,$$

où z représente la distribution normale standard, α est l'erreur de type I, β est l'erreur de type II, ES est la taille de l'effet (effect size), s est la déviation standard poolée définie comme :

$$s = \sqrt{\frac{(n_1 - 1)\sigma_1^2 + (n_2 - 1)\sigma_2^2}{n_1 + n_2 - 2}},$$

où n1 est la taille du groupe 1, n2 est la taille du groupe 2, σ_1^2 est la variance du groupe 1 et σ_2^2 est la variance du groupe 2.

6 Références

- Aldred MA, Huang Y, Liyanarachchi S, Pellegata NS, Gimm O, Jhiang S, Davuluri RV, de la Chapelle A, Eng C. Papillary and follicular thyroid carcinomas show distinctly different microarray expression profiles and can be distinguished by a minimum of five genes. J Clin Oncol. 2004 Sep 1;22(17):3531-9.

- Alwine, J. C., Kemp, D. J. & Stark, G. R. (1977). "Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes." Proc Natl Acad Sci U S A 74(12): 5350-4.

- Ambroise, C. and McLachlan, G.J. (2002) Selection bias in gene extraction on the basis of microarray gene-expression data. Proc. Natl Acad. Sci. USA, 99, 6562–6566

- Armstrong NJ, van de Wiel MA. Microarray data analysis: from hypotheses to conclusions using gene expression data. Cell Oncol. 2004;26(5-6):279-90. Review.

- Asa SL. My approach to oncocytic tumours of the thyroid. J Clin Pathol. 2004 Mar; 57(3):225-32. Review.

- Barden CB, Shister KW, Zhu B, Guiter G, Greenblatt DY, Zeiger MA, Fahey TJ 3rd. Classification of follicular thyroid tumors by molecular signature: results of gene profiling. Clin Cancer Res. 2003 May;9(5):1792-800.

- Barnard M (1935) The secular variations of skull characters in four series of Egyptian skulls. Ann. Eugenics 6, pp. 352–371.

- Bartel DP (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. Cell 116: 281–297.

- Bartolazzi A, Gasbarri A, Papotti M, Bussolati G, Lucante T, Khan A, et al; Thyroid Cancer Study Group. (2001) Application of an immunodiagnostic method for improving preoperative diagnosis of nodular thyroid lesions. Lancet. 357(9269):1644-50.

- Beckner ME, Heffess CS, Oertel JE (1995). Oxyphilic papillary thyroid carcinomas. Am J Clin Pathol 103(3): 280-7.

- Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: Apractical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B*, **57**: 289–300

- Benjamini,Y. and Yekutieli,D. (2001) The control of the false discovery rate under dependency. Ann Stat., 29, 1165–1188.

- Blanchette M, Bataille AR, Chen X, Poitras C, Laganiere J, Lefebvre C, Deblois G, Giguere V, Ferretti V, Bergeron D, Coulombe B, Robert F. Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression. Genome Res. 2006 May; 16(5):656-68. Epub 2006 Apr 10.

- Berho M, Suster S (1997) The oncocytic variant of papillary carcinoma of the thyroid: a clinicopathologic study of 15 cases. Hum Pathol 28(1): 47-53.

- Carlin JB, Doyle LW. Statistics for clinicians: 4: Basic concepts of statistical reasoning: hypothesis tests and the t-test. J Paediatr Child Health. 2001;37:72–77.

- Chapman S., Schenk P., Kazan K.,Manners J. (2002). Using biplots to interpret gene expression patterns in plants. Bioinformatics 18: 202-204.

- Chen D, Liu Z, Ma X, Hua D. Selecting genes by test statistics. J Biomed Biotechnol. 2005 Jun 30;2005(2):132-8.

- Chérié-Challine L et les membres du comité de rédaction. Surveillance sanitaire en France en lien avec l'accident de Tchernobyl. Bilan actualisé sur les cancers thyroïdiens et études épidémiologiques en cours en 2006. *Institut de veille sanitaire*, 2006, ISBN : 2-11-096297-6, http://www.invs.sante.fr/publications/2006/tchernobyl/

- Chilingaryan A., Gevorgyan N., Vardanyan A., Jones D., Szabo A. (2002). Multivariate approach for selecting sets of differentially expressed genes. Math. Biosci. 176: 59-72.
- Churchill GA. Fundamentals of experimental design for cDNA microarrays. Nat Genet. 2002 Dec;32 Suppl:490-5. Review.
- Collins, M. T.,N. J. Sarlis,M. J. Merino,J. Monroe,S. E. Crawford,J. A. Krakoff,L. C. Guthrie,S. Bonat,P. G. Robey andA. Shenker (2003). Thyroid carcinoma in the McCune-Albright syndrome: contributory role of activating Gs alpha mutations. J Clin Endocrinol Metab 88(9): 4413-7.
- DeLellis, R., Lloyd R., Heitz P. and Eng C. (2004). "World Health Organization Classification of Tumours: Pathology and Genetics of Tumours of Endocrine Organs." (ISBN 92 832 2416 7): 320p.
- Delisle MJ, Schwartz C, Theobald S, Maes B, Vaudrey C, Pochart JM. Les cancers de la thyroïde. Intérêt d'un registre de 627 patients diagnostiqués traités et suivis par une même équipe multi-disciplinaire. Ann. Endocrinol. (Paris) 1996 ; 57: 41-49.
- De Micco C, Ruf J, Chrestian MA, Gros N, Henry JF, Carayon P. (1991) Immunohistochemical study of thyroid peroxidase in normal, hyperplastic, and neoplastic human thyroid tissues. Cancer. 67(12):3036-41.
- Domingos, P. and Pazzani, M.J. (1997) On the Optimality of the simple Bayesian classifier under zero-one loss. Mach. Learn., 29, 103–130
- Dudoit, S., et al. (2002a) Comparison of discrimination methods for the classification of tumors using gene expression data. JASA, 97, 77–87.
- DudoitS, Yang YH, Callow MJ, Speed TP (2002b).Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments, Statistica Sinica, 2(1):111-139.
- Duggan, D. J., Bittner, M., Chen, Y., Meltzer, P. & Trent, J. M. (1999). "Expression profiling using cDNA microarrays." Nat Genet 21(1 Suppl): 10-4.
- Efron, B. (1983) Estimating the error rate of a prediction rule: improvements in cross-validation. JASA, 72, 316–331.
- Eisen M.B., Spellman P.T., Brown P.O., Botstein D. (1998). Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. 95: 14863-14868.
- Eszlinger M, Krohn K, Kukulska A, Jarzab B, Paschke R. Perspectives and limitations of microarray-based gene expression profiling of thyroid tumors. Endocr Rev. 2007 May;28(3):322-38. Epub 2007 Mar 12. Review.
- Fagin JA. (2002) Perspective: lessons learned from molecular genetic studies of thyroid cancer--insights into pathogenesis and tumor-specific therapeutic targets. Endocrinology. 143(6):2025-8.
- Fellenberg K., Hauser N.C., Brors B., Neutzner A., Hoheisel J.D., Vingron M. (2001). Correspondence analysis applied to microarray data. Proc. Natl. Acad. Sci. 96: 10781-10786.
- Fontaine JF, Mirebeau D, Franc B, Triau S, Rodien P, Houlgatte R, Malthièry Y & Savagner F (2007) Microarray analysis refines classification of non medullary thyroid tumours of uncertain malignancy. *Oncogene*. In press.
- Franc B, de la Salmoniere P, Lange F, Hoang C, Louvel A, de Roquancourt A, Vilde F, Hejblum G, Chevret S, Chastang C. Interobserver and intraobserver reproducibility in the histopathology of follicular thyroid carcinoma. Hum Pathol. 2003 Nov;34(11):1092-100.
- Franc B (2007) Histologie et cytologie de la thyroïde. Traité d'endocrinologie, ISBN-13: 978-2257120052: 123-130

- Getz G., Levine E., Domany E. (2000). Coupled two-way clustering analysis of gene microarray data. Proc. Natl. Acad. Sci. 97: 12079-12084.
- Gharib H. (1997) Changing concepts in the diagnosis and management of thyroid nodules. Endocrinol.Metab Clin.North Am. 26, 777-800.
- Ghosh D., Chinnaiyan A.M. (2002). Mixture modelling of gene expression data from microarray experiments. Bioinformatics 18: 275-286.
- Gimm O, Dziema H, Brown J, Hoang-Vu C, Hinze R, Dralle H, et al (2001) Over-representation of a germline variant in the gene encoding RET co-receptor GFRalpha-1 but not GFRalpha-2 or GFRalpha-3 in cases with sporadic medullary thyroid carcinoma. Oncogene. 20(17):2161-70.
- Gordon A, Glazko G, Qiu X, Yakovlev A. Control of the mean number of false discoveries, bonferroni and stability of multiple testing. *The Annals of Applied Statistics* 2007,Vol.1,No.1,179–190
- Gozu, H.,M. Avsar,R. Bircan,M. Claus,S. Sahin,O. Sezgin,O. Deyneli,R. Paschke,B. Cirakoglu andS. Akalin (2005). "Two novel mutations in the sixth transmembrane segment of the thyrotropin receptor gene causing hyperfunctioning thyroid nodules." Thyroid 15(4): 389-97.
- GuhaThakurta D. Computational identification of transcriptional regulatory elements in DNA sequence. Nucleic Acids Res. 2006 Jul 19;34(12):3585-98. Print 2006. Review.
- Hastie T., Tibshirani R., Botstein D., Brown P. (2001). Supervised harvesting of expression trees. Genome Biology 2(1): research0003.1-0003.12.
- Hastie T., Tibshirani R., Eisen M.B., Alizadeh A., Levy R., Staudt L., Chan W.C., Botstein D., Brown P. (2000). 'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns. Genome Biology 1(2):research0003.1-0003.21.
- Hedinger C, Williams ED, Sobin LH (1988). WHO: histological typing of thyroid tumours. New York, Springer, 70 pages
- Herrera MF, Hay ID, Wu PS, Goellner JR, Ryan JJ, Ebersold JR, Bergstralh EJ, Grant CS (1992). "Hurthle cell (oxyphilic) papillary thyroid carcinoma: a variant with more aggressive biologic behavior." World J Surg 16(4): 669-74; discussion 774-5
- Hirokawa M, Carney JA, Goellner JR, DeLellis RA, Heffess CS, Katoh R, Tsujimoto M, Kakudo K. Observer variation of encapsulated follicular lesions of the thyroid gland. Am J Surg Pathol. 2002 Nov;26(11):1508-14.
- Holm S (1979) A simple sequentially rejective multiple test procedure. Scand. J. Statist. 6: 665–70.
- Huang Y, Prasad M, Lemon WJ, Hampel H, Wright FA, Kornacker K, LiVolsi V, Frankel W, Kloos RT, Eng C, Pellegata NS, de la Chapelle A. Gene expression in papillary thyroid carcinoma reveals highly consistent profiles. Proc Natl Acad Sci U S A. 2001 Dec 18;98(26):15044-9.
- Hwang D, Schmitt WA, Stephanopoulos G, Stephanopoulos G. Determination of minimum sample size and discriminatory expression patterns in microarray data. Bioinformatics. 2002 Sep;18(9):1184-93.
- Inza I, et al. (2004) Filter versus wrapper gene selection approaches in DNA microarray domains. Artif. Intell. Med., 31, 91–103
- Jafari P, Azuaje F. An assessment of recently published gene expression data analyses: reporting experimental design and statistical factors. BMC Med Inform Decis Mak. 2006 Jun 21;6:27. Review.
- Jeffery IB, Higgins DG, Culhane AC. Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data. BMC Bioinformatics. 2006 Jul 26;7:359.

- Kariya T, Sinha BK, Giri NC. Robustness of t-Test. Defense Technical Information Center, Technical rept. Oct 1986. Accession Number : ADA176972
- Kim BS, Kim I, Lee S, Kim S, Rha SY, Chung HC. Statistical methods of translating microarray data into clinically relevant diagnostic information in colorectal cancer. Bioinformatics. 2005 Feb 15;21(4):517-28. Epub 2004 Sep 16.
- Krajewski P, Bocianowski J. Statistical methods for microarray assays. J Appl Genet. 2002;43(3):269-78. Review.
- Kreil DP, Russell RR. There is no silver bullet--a guide to low-level data transforms and normalisation methods for microarray data. Brief Bioinform. 2005 Mar;6(1):86-97. Review.
- Kuruvilla F.G., Park P.J., Schreiber S.L. (2002). Vector algebra in the analysis of genome-wide expression data. Genome Biology 3(3): research 0011.1-0011.11.
- Li, T., et al. (2004) A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. Bioinformatics, 20, 2429–2437
- Liang, P. & Pardee, A. B. (1992). "Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction." Science 257(5072): 967-71.
- Lloyd RV, Erickson LA, Casey MB, Lam KY, Lohse CM, Asa SL, Chan JK, DeLellis RA, Harach HR, Kakudo K, LiVolsi VA, Rosai J, Sebo TJ, Sobrinho-Simoes M, Wenig BM, Lae ME. Observer variation in the diagnosis of follicular variant of papillary thyroid carcinoma. Am J Surg Pathol. 2004 Oct;28(10):1336-40.
- Lockhart, D. J., Dong, H., Byrne, M. C., Follettie, M. T., Gallo, M. V., Chee, M. S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H. & Brown, E. L. (1996). "Expression monitoring by hybridization to high-density oligonucleotide arrays." Nat Biotechnol 14(13): 1675-80.
- Malthiery Y, Savagner F. [Energy metabolism of the cancer cell: example of mitochondria-rich endocrine tumors] Ann Endocrinol (Paris). 2006 Jun;67(3):205-13. Review. French.
- Mclachlan G.J., Bean R.W., Peel D. (2002). A mixture model-based approach to the clustering of microarray expression data. Bioinformatics 18: 413-422.
- Moher D, Dulberg CS, Wells GA. Statistical power, sample size, and their reporting in randomized controlled trials. JAMA. 1994;272:122–124.
- Mount DW, Pandey R. Using bioinformatics and genome analysis for new therapeutic interventions. Mol Cancer Ther. 2005 Oct;4(10):1636-43. Review.
- Niccoli-Sire P, Conte-Devolx B (2007) Cancer médullaire de la thyroïde. Traité d'endocrinologie, ISBN-13: 978-2257120052: 184-189
- Palstra RJ, Tolhuis B, Splinter E, Nijmeijer R, Grosveld F, de Laat W. The beta-globin nuclear compartment in development and erythroid differentiation. Nat Genet. 2003 Oct;35(2):190-4. Epub 2003 Sep 21.
- Pan W, Lin J, Le CT. How many replicates of arrays are required to detect gene expression changes in microarray experiments? A mixture model approach. Genome Biol. 2002;3(5):research0022. Epub 2002 Apr 22.
- Pavlidis P, Li Q, Noble WS. The effect of replication on gene expression microarray experiments. Bioinformatics. 2003 Sep 1;19(13):1620-7.
- Qin LX, Kerr KF; Contributing Members of the Toxicogenomics Research Consortium. Empirical evaluation of data transformations and ranking statistics for microarray analysis. Nucleic Acids Res. 2004 Oct 12;32(18):5471-9. Print 2004. Erratum in: Nucleic Acids Res. 2004 Dec;32(22):6718. Nucleic Acids Res. 2004 Nov 8;32(19):5972.
- Salvatore, G.,R. Giannini,P. Faviana,A. Caleo,I. Migliaccio,J. A. Fagin,Y. E. Nikiforov,G. Troncone,L. Palombini,F. Basolo andM. Santoro (2004). "Analysis of BRAF point mutation and RET/PTC rearrangement

refines the fine-needle aspiration diagnosis of papillary thyroid carcinoma." J Clin Endocrinol Metab 89(10): 5175-80.

- Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. (1995). "Quantitative monitoring of gene expression patterns with a complementary DNA microarray." Science 270(5235): 467-70.

- Schena, M., Shalon, D., Heller, R., Chai, A., Brown, P. O. & Davis, R. W. (1996). "Parallel human genome analysis: microarray-based expression monitoring of 1000 genes." Proc Natl Acad Sci U S A 93(20): 10614-9.

- Seldrup J. Whatever happened to the t-test? Drug Inf J. 1997;31:745–750.

- Seo J, Gordish-Dressman H, Hoffman EP. An interactive power analysis tool for microarray hypothesis testing and generation. Bioinformatics. 2006 Apr 1;22(7):808-14. Epub 2006 Jan 17.

- Simon, R., et al. (2003) Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. J. Natl Cancer Inst., 95, 14–18

- Tibshirani R. A simple method for assessing sample sizes in microarray experiments. BMC Bioinformatics. 2006 Mar 2;7:106.

- Tronche F, Ringeisen F, Blumenfeld M, Yaniv M, Pontoglio M. Analysis of the distribution of binding sites for a tissue-specific transcription factor in the vertebrate genome. J Mol Biol. 1997 Feb 21;266(2):231-45.

- Van 't Veer, L.J., et al. (2002) Gene expression profiling predicts clinical outcome of breast cancer. Nature, 415, 530–536

- Vapnik, V. Statistical Learning Theory, (1999) , New York John Wiley and Sons.

- Velculescu, V. E., Zhang, L., Vogelstein, B. & Kinzler, K. W. (1995). "Serial analysis of gene expression." Science 270(5235): 484-7.

- Wall M.E., Dyck P.A., Brettin T.S. (2001). SVDMAN – singular value decomposition analysis of microarray data. Bioinformatics 17: 566-568.

- Wasserman WW, Sandelin A. Applied bioinformatics for the identification of regulatory elements. Nat Rev Genet. 2004 Apr;5(4):276-87. Review.

- Wessels LF, Reinders MJ, Hart AA, Veenman CJ, Dai H, He YD, van't Veer LJ. A protocol for building and evaluating predictors of disease state based on microarray data. Bioinformatics. 2005 Oct 1;21(19):3755-62. Epub 2005 Apr 7.

- Westfall PH and Young SS. Resampling-Based Multiple Testing, Wiley, New York (1993).

- Williams JL, Hathaway CA, Kloster KL, Layne BH. Low power, type II errors, and other statistical problems in recent cardiovascular research. Am J Physiol. 1997;273:487–493.

- Wu J, Smith LT, Plass C, Huang TH. ChIP-chip comes of age for genome-wide functional analysis. Cancer Res. 2006 Jul 15;66(14):6899-902. Review.

- Wu FX, Zhang WJ, Kusalik AJ. Determination of the minimum number of microarray experiments for discovery of gene expression patterns. BMC Bioinformatics. 2006 Dec 12;7 Suppl 4:S13.

- Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, Lindblad-Toh K, Lander ES, Kellis M. Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. Nature. 2005 Mar 17;434(7031):338-45. Epub 2005 Feb 27.

- Yang MC, Yang JJ, McIndoe RA, She JX. Microarray experimental design: power and sample size considerations. Physiol Genomics. 2003 Dec 16;16(1):24-8.

- Yang K, Li J, Gao H. The impact of sample imbalance on identifying differentially expressed genes. BMC Bioinformatics. 2006 Dec 12;7 Suppl 4:S8.

- Yang YH, Speed T. Design issues for cDNA microarray experiments. Nat Rev Genet. 2002 Aug;3(8):579-88. Review.

- Yeung K.Y., Fraley C., Murua A., Raftery A.E., Ruzzo W.L. (2001). Model-based clustering and data transformation for gene expression data. Bioinformatics 17: 977-987.

- Yeung K.Y., Ruzzo W.L. (2001). Principal component analysis for clustering gene expression data. Bioinformatics 17: 763-774.

- Yukinawa N, Oba S, Kato K, Taniguchi K, Iwao-Koizumi K, Tamaki Y, Noguchi S, Ishii S. A multi-class predictor based on a probabilistic model: application to gene expression profiling-based diagnosis of thyroid tumors. BMC Genomics. 2006 Jul 27;7:190

Résumé

L'accident de Tchernobyl, en 1986, a mis en lumière le cancer de la thyroïde. Ce cancer, assez rare, est pourtant en forte augmentation, et cela avant même 1986. Le diagnostic des tumeurs folliculaires thyroïdiennes par analyses anatomopathologiques est parfois difficile. Ce travail s'intéresse à la classification moléculaire de ces tumeurs ainsi qu'à la recherche de marqueurs spécifiques à chaque pathologie. En regroupant et en analysant le transcriptome de plus de 90% des types de lésions folliculaires de la thyroïde, nous avons établi une classification moléculaire globale de ces lésions, nous avons défini des marqueurs spécifiques faisant l'objet d'un brevet, et nous avons permis de préciser le diagnostic de la majorité des tumeurs de malignité incertaine. L'étude bioinformatique des données issues de biopuces apporte des précisions à la classification internationale des tumeurs, particulièrement pour les tumeurs oncocytaires, microfolliculaires et de malignité incertaine.

English title

CHARACTERISATION AND MOLECULAR CLASSIFICATION OF THYROID PATHOLOGIES BY TRANSCRIPTOMIC APPROACH

Abstract

In 1986, the Tchernobyl accident highlighted cancers of the thyroid. This rather rare cancer is still in progression, and before 1986. Diagnosis of follicular thyroid tumours by pathologists is sometimes difficult. This work is about the molecular classification of these tumours, and also the search for specific gene markers for each of the pathologies. By gathering and analysing the transcriptome of more than 90% of the thyroid follicular lesions, we have established a global molecular classification of these lesions, we have defined specific markers in a patent, and we have brought precisions on the diagnosis of tumours of uncertain malignancy. The bioinformatics study of microarray data brings precisions to the international classification of the tumours, particularly for oncocytic tumours, microfollicular tumours and tumours of uncertain malignancy.

Discipline : Biologie Cellulaire

Mots clés : bioinformatique, cancer, thyroïde, puces à ADN, classification