Etude d'un clade de rétrotransposons Copia : les GalEa, au sein des génomes eucaryotes

Tifenn Donnart

Thèse de Doctorat de l'Université Pierre et Marie Curie

Ecole doctorale

Complexité du Vivant (ED 515)

Institut de Biologie Paris-Seine

UMR7138 "Evolution Paris-Seine" UPMC-CNRS

Equipe : Evolution du génome eucaryote

Etude d'un clade de rétrotransposons Copia: les GalEa, au sein des génomes eucaryotes

Par Tifenn Donnart

Thèse de doctorat de Génétique

Dirigée par Éric Bonnivard

Présentée et soutenue publiquement le 2 février 2015

Devant un jury composé de :

Emmanuelle LERAT, CR HDR, Université de Lyon I	Rapporteur
Hadi QUESNEVILLE. DR, INRA, Versailles	Rapporteur
Marie-Angèle GRANDBASTIEN, DR, INRA, Versailles	Examinatrice
Frédéric DEVAUX. Prof, Université Paris VI.	Examinateur
Éric BONNIVARD. MCU, HDR, Université Paris VI.	Directeur de thèse

Remerciements

Une thèse, c'est un long parcours que l'on ne sillonne jamais seul; J'aimerai remercier toutes les personnes qui m'ont accompagnée tout au long de ces 4 longues années.

Je remercie tout d'abord Hervé Le Guyader et Dominique Higuet de m'avoir accueillie au sein de leur unité.

Je remercie les membres du jury, mes rapporteurs Hadi Quesneville et Emmanuelle Lerat ainsi que mes examinateurs, Frédéric Devaux et Marie Angèle Grandbastien, qui ont accepté d'évaluer mon travail.

Je tiens aussi à remercier Laure Teysset, Aurélie Hua Van et de nouveau Hadi Quesneville pour avoir accepté de faire partie de mon comité de thèse.

Un grand merci à mon directeur de thèse, Éric Bonnivard. Merci d'avoir été aussi cool, aussi ouvert, de m'avoir permis de réaliser cette thèse dans les meilleures conditions qu'un thésard puisse rêver et de m'avoir soutenu durant toutes ces années. J'espère que je ne t'aurais pas trop désespéré et que tu auras de nombreux thésards par la suite, car tu m'as énormément appris.

Je voudrais remercier de nouveau Dominique Higuet qui m'a accueillie au sein de son équipe et qui a toujours été disponible pour des conversations scientifiques ou non, malgré un emploi du temps chargé.

Je tiens également à remercier Sophie Brouillet et Guillaume Achaz de l'ABI pour leurs énormes coups de main en bioinfo, sans lesquels j'aurais été bien perdue. Merci pour votre patience.

Je souhaiterais également remercier tous les thésards, passés et présents, de l'équipe : Mathieu pour tout et surtout pour tous tes coups de main en bioinfo et pour ta patience. Marie pour toutes ces journées passées ensemble à parler de tout et de rien, et surtout de tout. Kader, pour nous avoir appris le sens du mot modestie et nous avoir bien fait rire. Et Marguerite que je n'aurais pas connue si longtemps.

Je tiens à dire un grand merci à Paula, pour tout. Merci pour ta bonne humeur et tes sourires en toutes circonstances, pour les innombrables pauses café; Tu es indispensable à cette équipe !

Je remercie également tous les autres membres de l'équipe Génétique et Evolution, Denise, Denis, Hervé, Pierre, Laurent pour leurs précieux conseils.

Je remercie tous les thésards, du labo Systématique Adaptation Evolution, pour les déjeuners du lundi midi, enfin ces réunions, on va dire que c'était de vraies réunions, et les Journées Evolution si bien organisées.

Merci à Danielle de gérer toutes les questions administratives pour nous tous, tu es une mère pour nous. Merci aussi à Philipe.

Je remercie Gaëlle, le sourire de notre ancienne école doctorale, merci pour toute ton aide.

Merci à toutes les personnes du labo Systématique Adaptation Evolution, nouvellement renommée Evolution Paris Seine, pour leur disponibilité lors de mes nombreuses questions.

Merci à la meilleure des docteurs (je dirais même un Docteur de Choque ! ah ah ah !), Elodie, qui a fait beaucoup pour ce manuscrit, et surtout les relectures. Merci pour toutes tes corrections et de me redonner confiance quand j'en avais besoin.

Un grand merci aussi à tous mes amis qui m'ont soutenue pendant ces 4 années de thèse ou depuis bien plus longtemps, Clément , Sabrina, Anne-so, Céline, Fanny, Flora, Laure, Louma, Marie, Mélanie, Basile et j'en oublie sûrement..

Merci aux FT ! Bientôt 30 !!!

Enfin, je voudrai dire merci à ma famille, mes parents, ma sœur, bandes de fou vous n'auriez jamais dû me laisser faire tout ça !!!! A mes oncles et tantes, oui on va en boire du champagne !

A Alex, les derniers mois ont été difficiles mais c'est bon, c'est fini :) on va pouvoir profiter maintenant.

Et en dernier à Ana, d'être une bébé si parfaite, pour tous tes sourires pendant cette dure rédaction, ça l'a rendue plus agréable à vivre… continue de sourire comme ça tous les jours.

À Ana

Sommaire

RESUME

Résumé

Les éléments transposables jouent un rôle majeur dans l'évolution des génomes eucaryotes. La connaissance de la distribution des éléments transposables entre différentes espèces au sein d'un même taxon est une condition essentielle pour étudier leur dynamique et mieux comprendre leur rôle dans l'évolution des espèces. Compte tenu de leur abondance, de leur diversité spécifique et de milieu de vie, les crustacés sont un excellent modèle pour étudier la génomique comparative des rétrotransposons. C'est notamment chez les Galathées qu'a été défini le clade GalEa des éléments de la superfamille des Copia. Nous avons étudié la distribution de deux superfamilles de rétrotransposons à LTR bien connus: les Gypsy et les Copia, au sein des crustacés. En combinant des PCRs avec amorces dégénérées et des analyses *in silico*, nous avons identifié 35 familles de rétrotransposons Copia et 46 familles de rétrotransposons Gypsy dans respectivement 15 et 18 espèces de crustacés (principalement des malacostracés : crabes, crevettes, krill...). Ces éléments présentent une distribution et une diversité différentes au sein des crustacés. Les éléments Gypsy apparaissent relativement fréquents et diversifiés dans toutes les espèces. A l'inverse, les éléments Copia semblent rares, donc difficilement détectables, et sont largement dominés par les éléments du clade GalEa. Ces résultats suggèrent deux stratégies différentes de dynamique pour les rétrotransposons Gypsy (théorie de la Reine Rouge) et les rétrotransposons GalEa ('domino days spreading' branching process). De plus, les éléments GalEa présentent un grand succès évolutif en étant largement distribués dans de nombreuses branches de métazoaires. Ils sont aussi présents chez quelques algues rouges et nous en avons également détectés chez des champignons. Profitant des nombreuses données génomiques disponibles, nous avons donc étudié la distribution des éléments GalEa de Champignon, dans le but de comparer celle-ci aux résultats obtenus chez les crustacés. En fait, ils n'apparaissent qu'au sein d'un grand embranchement d'ascomycètes, les pezizomycotina, et ils forment un groupe monophylétique au sein des GalEa. Enfin, chez les champignons, les éléments GalEa ne sont pas majoritaires parmi les rétrotransposons Copia. Nous avons donc initié une nouvelle étude chez les mollusques, afin de définir si les résultats obtenus chez les crustacés sont une caractéristique des éléments GalEa, des malacostracés ou des métazoaires.

Mots clés : [Rétrotransposons à LTR, Gypsy, Copia, GalEa, Crustacés, Rhodophytes, Champignons, Analyse des génomes]

Abstract

Transposable elements play a major role in the evolution of eukaryotic genomes. Knowing the distribution of transposable elements between different species within the same taxon is essential to study their dynamics and to better understand their role in the evolution of species. Given their abundance, species diversity and living environment, crustaceans are an excellent model for studying comparative genomics of retrotransposons. It is notably in the squat lobsters that the GalEa clade of Superfamily Copia was defined. We studied the distribution of two well-known LTR retrotransposons superfamilies: Gypsy and Copia, in crustaceans. By combining PCRs with degenerate primers and *in silico* analysis, we identified 35 families of Copia retrotransposons and 46 families of Gypsy retrotransposons in 15 and 18 species of crustaceans (mainly malacostraca: crabs, shrimp, krill ...). These elements have different distribution and diversity in crustaceans. Gypsy elements appear relatively commonly and diverse in all species. Conversely, the Copia elements seem rare, and consequently more difficult to detect, and are largely dominated by the elements of the clade GalEa. These results suggest two different dynamic strategies for retrotransposons Gypsy (the Red Queen theory) and retrotransposons GalEa (‘domino days spreading’ branching process). In addition, GalEa elements present a great evolutionary success being widely distributed in many branches of metazoans. They are also present in certain red algae and we have also detected them in fungi. Taking advantage of the large amount of available genomic data, we have studied the distribution of GalEa elements of fungi, in order to compare it with the results obtained in crustaceans. In fact, they appear only in a large phylum of ascomycetes, in pezizomycotina, and they form a monophyletic group within the GalEa. Finally, in the fungi, the GalEa elements are not majority among Copia retrotransposons. We have therefore initiated a new study in molluscs, to define if the results obtained in crustaceans are a feature of GalEa elements, malacostraca or metazoans.

Key words: [LTR Retrotransposons, Gypsy, Copia, GalEa, Crustaceans, Rhodophyta, Fungi, Genome Analysis]

INTRODUCTION

Chapitre I. Introduction générale

Barbara McClintock mit en évidence des évènements d'insertions, de délétions et de translocations causés par les éléments Ac/Ds lors de son étude sur la mosaïque des patrons de couleurs des semences de maïs (*Zea maïs*) en 1948. Ces éléments ont depuis été identifiés comme des éléments transposables (ETs) et des études ont montré que plus de 75 % du génome du maïs est constitué d'ET (SanMiguel et al., 1996). Elle reçut pour cette découverte le prix Nobel de médecine en 1983. Les ETs sont des constituants majeurs des génomes, ce sont des séquences d'ADN mobiles, ou ayant été mobiles, capables de se déplacer et de se multiplier de manière autonome (ou non) au sein des génomes par un mécanisme appelé transposition. Présents chez presque tous les organismes vivants où ils ont été recherchés (eucaryotes et procaryotes), les ETs sont une part importante et variable des génomes eucaryotes: de 3 à près de 20% dans les génomes de champignons (eumycète) (Dhillon et al., 2014; Kim et al., 1998) **,** de 3 à 52% des génomes de métazoaires (près de 25% au sein du génome de la drosophile, environ 45% chez l'Homme (Jordan et al., 2003)) et jusqu'à plus de 90% dans les génomes de plantes (Leitch and Leitch, 2008; Mikkelsen et al., 2007; Wicker et al., 2007).

1.1 Tailles des génomes et paradoxe de la valeur C

La taille du génome correspond à la quantité d'ADN présente dans une copie d'un génome donné. Elle est mesurée en paires de bases (pb) ou en picogrammes (pg), elle est également appelée la valeur C. La variation de la taille des génomes peut être très forte au sein des eucaryotes (Figure 1), entre 2,8 Mb pour la microsporidie *Encephalitozoon cuniculi* (Peyretaillade et al., 1998) à plus de 690 000 Mb pour la diatomée *Navicola pelliculosa*. La variation de taille de génome peut être faible au sein d'un même taxon d'espèces, par exemple les amniotes (mammifères, reptiles, oiseaux). Dans d'autres taxons comme les mollusques, les crustacés ou encore les champignons, la taille des génomes varie de façon plus importante. Chez les angiospermes, le génome de certaines espèces peut être jusqu'à 1000 fois plus grand que celui des espèces possédant le plus petit génome. Mais la plus grande variation observable se fait au sein des protozoaires, des eucaryotes unicellulaires hétérotrophes qui ingèrent leur nourriture par phagocytose (ce groupe est paraphylétique, au sein des protistes).

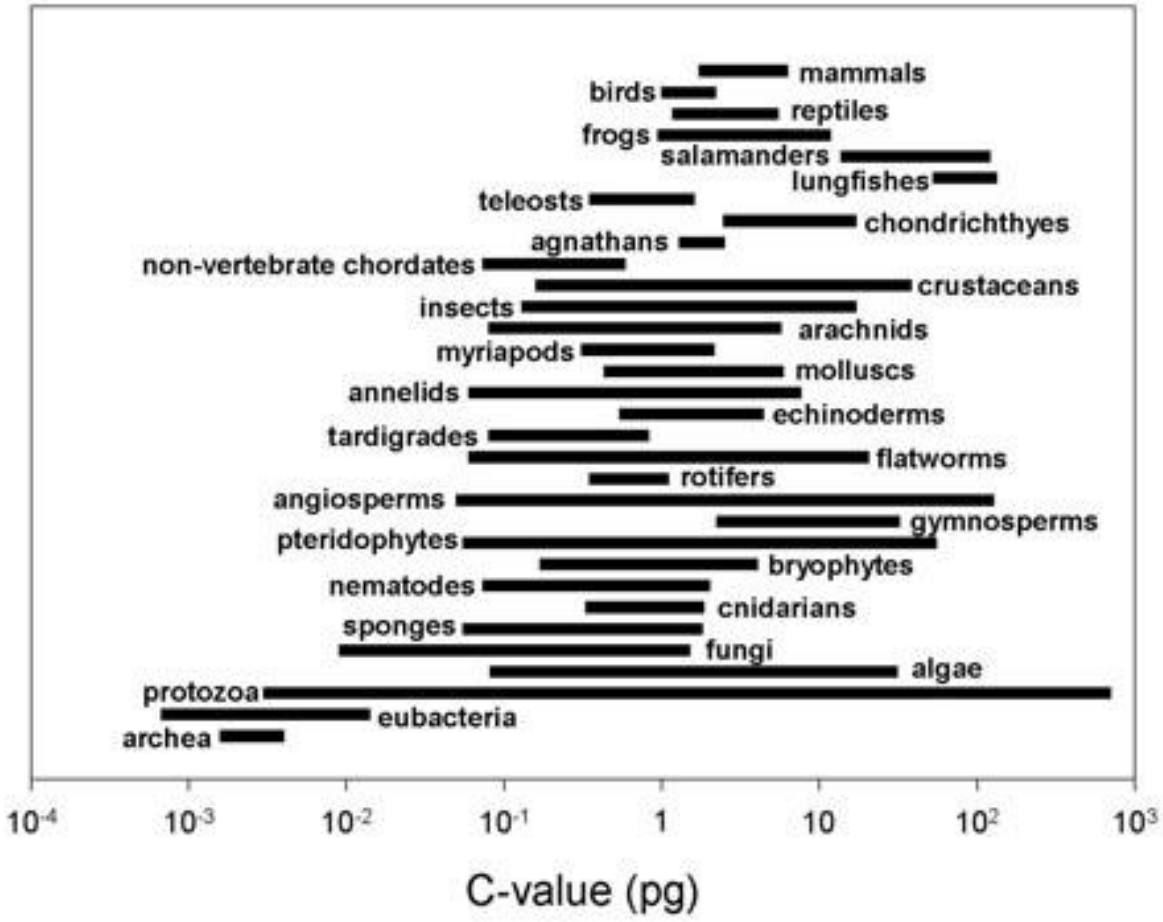


Figure 1: Variation de la taille des génomes eucaryotes (Gregory, 2001, www.genomesize.com)

Dans les années 1950, il était admis que plus un organisme semblait complexe, plus il présentait une grande quantité d'ADN et un plus grand nombre de gènes également. On sait aujourd'hui que ce n'est pas le cas. En effet, au sein des eucaryotes, les amibes sont considérées comme des organismes simples, car unicellulaires, par rapport aux mammifères qui seraient des organismes plus complexes. Cependant, certaines amibes (*Amoeba dubia*, 670 Gb, (http://www.genomenewsnetwork.org/articles/02_01/Sizing_genomes.shtml)) présentent un génome 200 fois supérieur au génome humain (3Gb). Il n'existe donc pas de lien entre la taille du génome et la complexité d'un organisme, c'est ce qu'on appelle le paradoxe de la valeur C (Thomas, 1971).

L'augmentation de la taille d'un génome peut avoir différentes origines. Un génome peut avoir

sa taille augmentée grâce à des duplications complètes (polyploidisation comme chez les plantes) ou segmentales, le transfert d'ADN exogène, l'augmentation du nombre ou de la taille des introns. Ce paradoxe peut aussi s'expliquer, au sein de certains génomes, par la quantité de séquences répétées, notamment d'ETs, qui peuvent connaitre des «bursts» de transposition (amplifications brusques et importantes du nombre de copies) et de ce fait se retrouver en grande quantité au sein du génome et agrandir celui-ci. On sait que chez les plantes, et plus particulièrement les angiospermes, l'amplification de rétrotransposons à LTR contribue à une forte augmentation de la taille des génomes (SanMiguel et al., 1998). Par exemple chez le maïs, cette augmentation du nombre de rétrotransposons à LTR a conduit au doublement de la taille du génome (SanMiguel al., 1998). Ainsi, les ETs participent à l'évolution de la taille des génomes et une corrélation entre la taille des génomes et la quantité d'ETs au sein de ces génomes a été soulignée (Gregory, 2005).

1.2 Impact des ETs sur les génomes

Les éléments transposables sont considérés comme des moteurs puissants de l'évolution et de la biodiversité. En effet, le fait qu'il s'agisse de séquences répétées capables de transposer en fait des facteurs de plasticité ayant un fort impact sur leurs génomes hôtes. Comme nous l'avons évoqué plus haut, ils participent à l'évolution de la taille des génomes. En tant que séquences répétées au sein des génomes, ils peuvent induire des réarrangements chromosomiques (Lim and Simmons, 1994) comme des duplications segmentales et des délétions par recombinaisons illégitimes (Bonnivard et al., 2009; Capy et al., 1997), constituant ainsi l'un des processus majeurs permettant de contrebalancer l'expansion des génomes (Devos et al., 2002). De plus, les duplications segmentales peuvent participer à l'apparition de nouvelles fonctions chez l'espèce hôte. En tant que séquences mobiles, ils induisent des mutations lors de leurs insertions. Ils peuvent, par exemple, modifier l'expression de certains gènes en s'insérant dans leurs régions régulatrices. Chez l'Homme, on retrouve des traces d'éléments insérés dans 18% des régions transcrites non traduites et dans 25% des promoteurs (Jordan et al., 2003). Cet effet est renforcé par le fait que les ETs autonomes possèdent leurs propres promoteurs et peuvent ainsi avoir un effet sur l'expression des gènes proches de la région où ils se sont insérés. Ceci est plus particulièrement le cas des rétrotransposons à LTR car il y a des séquences promotrices à la fois en 5' et 3' de l'élément.

Les ETs peuvent aussi participer à l'évolution des espèces de par leur capacité à influencer l'expression des gènes. Ainsi, lors de l'insertion d'un ET au sein d'une partie codante d'un gène, cela peut induire une perte de fonction de celui-ci. Les ETs ne sont pas présents au hasard au sein des génomes. En effet, on les retrouve le plus souvent au sein des télomères ou des centromères, régions pauvres en gènes, et induisent l'hétérochromatinisation de l'ADN nucléaire. L'hétérochromatinisation de l'ADN constitue l'un des principaux mécanismes impliqués dans la régulation de l'expression de l'ensemble des gènes et des ETs.

Le génome hôte peut aussi recruter une ou plusieurs des fonctions des ETs dans le cadre de leur domestication moléculaire (Bonnivard et al., 2009). Cette domestication semble être impliquée dans de nombreuses fonctions au sein des génomes eucaryotes, comme la régulation de la structure de la chromatine, le maintien de l'intégrité des chromosomes, la régulation de l'apoptose, le contrôle du cycle cellulaire, la régulation transcriptionnelle ou la protection contre l'invasion d'ETs (Sinzelle et al., 2009). Par exemple, le remplacement de la fonction de la télomérase chez la drosophile est due à la transposition des éléments TART et TAHRE et de l'élément HeT-A (type LINE) (Villasante et al., 2007). Chez les vertébrés, la fonction d'endonucléase d'ETs a été recrutée dans le cadre de la recombinaison V(D)J (Variable Diversity Joining), permettant la reconnaissance d'une grande diversité de pathogènes par le système immunitaire (Kapitonov and Jurka, 2005). De même, chez les métazoaires, les protéines de la famille THAP (« Thanatos-associated protein »), jouent des rôles importants dans la prolifération cellulaire, le contrôle du cycle cellulaire et l'apoptose (Chesney et al., 2006; Clouaire et al., 2005; Macfarlan et al., 2005). Ces protéines résultent de chimères entre des gènes cellulaires et le domaine de liaison à l'ADN de la transposase de l'élément P (Quesneville et al., 2005; Roussigne et al., 2003).

1.3 Les différentes échelles d'études des ETs

Les études des ETs portent sur de nombreux points. Certains s'intéressent par exemple à leur mobilité, aux systèmes de régulation de leur expression au sein des génomes (e.g. réponse au stress), aux étapes de leur mécanisme de transposition. On peut également étudier leur impact sur les génomes et la part des ETs au sein de ceux-ci. Il est intéressant de rechercher par phylogénie les relations entre les éléments et d'établir la classification de ces derniers. Enfin, rechercher la distribution des différents types d'éléments au sein des espèces permet

l'étude de leurs dynamiques (e.g. éventuel transfert entre espèces). Ces différentes études apportent des résultats essentiels à la compréhension des ETs, et sont réalisables grâce aux diverses techniques de génétique et génomique développées au cours de ces dernières années.

L'étude de la distribution des ETs peut se faire à différentes échelles. Du point de vue de l'hôte, on peut se placer au sein d'une espèce (*Drosophila melanogaster*), dans un groupe d'espèces (par exemple l'ordre des décapodes), un embranchement (comme les mollusques), ou encore chez l'ensemble des eucaryotes. De la même manière, du point de vue des ETs, on peut s'intéresser à un élément en particulier (l'élément *P* a ainsi été particulièrement bien étudié), à une famille d'éléments (un ensemble d'éléments assez proche, e.g. la famille des éléments Alvi1) ou même une superfamille (éléments assez éloignés mais qui partagent des critères communs, par exemple structuraux ; e.g. la superfamille des éléments DIRS). On peut par la suite envisager toutes les combinaisons possibles, le niveau de précision dépendant alors du choix du chercheur (Figure 2).

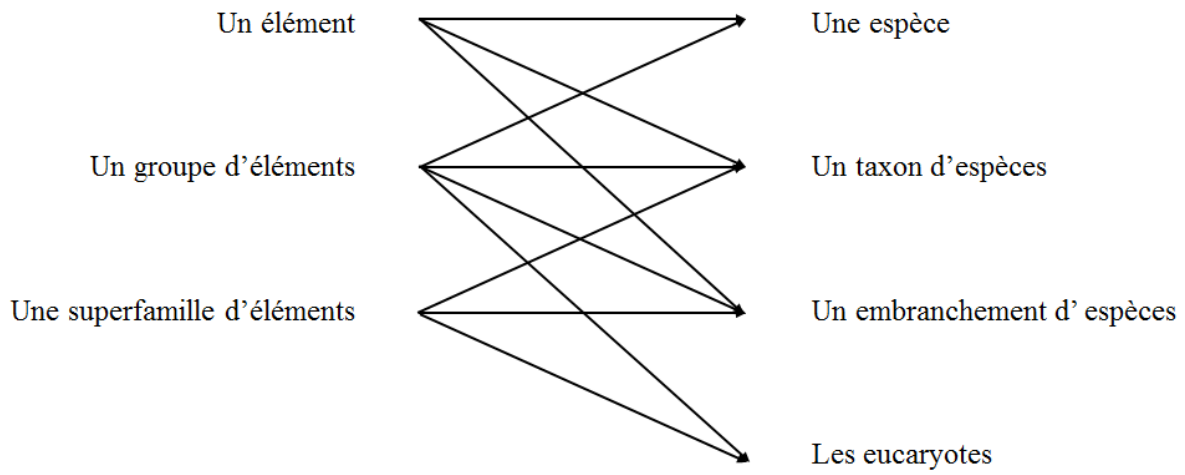


Figure 2: Représentation de l'étude de la distribution des ETs au sein des espèces hôtes.

De nos jours, de nombreuses études portent sur l'ensemble des ETs d'un génome (mobilome). Ceci notamment en lien avec l'annotation chez les espèces dont le génome est complètement séquencé, comme les espèces modèles. Cela permet également la découverte de nouveaux éléments et le positionnement de ces nouveaux éléments grâce, entre autres, à la phylogénie. De plus en plus de génomes complètement séquencés sont disponibles aujourd'hui grâce à l'avancée des nouvelles technologies. Le nombre de programmes de séquençage de génomes augmentant sans cesse que ce soit au niveau d'une espèce (projet 1000 génomes humains),

d’espèces proches (projets de séquençage de génomes de drosophiles), ou d’un phylum (programme 1000 génomes de champignons). Ces études ont permis d’enrichir considérablement les données sur les éléments annotés, sur leur abondance et leur très grande diversité. On peut ainsi mieux comprendre l’évolution des génomes et l’étude de l’évolution des tailles de génomes au sein d’un taxon.

L’étude de la distribution d’éléments chez des espèces non modèles (par exemple l’étude des éléments DIRS au sein des crustacés) permet de comprendre la dynamique d’un élément ou d’un groupe d’éléments au sein d’un taxon. On peut également faire des études comparatives de distributions de plusieurs superfamilles d’éléments (les Copia et les Gypsy) au sein d’une même espèce (*Rimicaris exoculata*) ou d‘un taxon d’espèces (les crustacés). Ces études peuvent nous donner des éléments de réponses pour comprendre l’origine de certains ET et nous permettre d’appréhender leur dynamique à plus large échelle. Très peu d’études comparatives de plusieurs types de rétrotransposons ont été réalisées sur leur distribution au niveau d’un embranchement d’espèces.

Au cours de ma thèse nous avons étudié les deux superfamilles Gypsy et Copia au sein des crustacés (Figure 3) ; en nous intéressant plus particulièrement à un clade d’éléments Copia: le clade des GalEa. Nous avons recherché ces éléments au sein de plusieurs taxons tels que les crustacés, les mollusques, mais également en dehors des métazoaires, chez des rhodophytes et des champignons, pour réaliser une étude comparative de ces éléments et mieux appréhender leur dynamique. En effet, nous voulions savoir si ces éléments GalEa présentaient la même dynamique au sein de différents taxons d’eucaryotes.

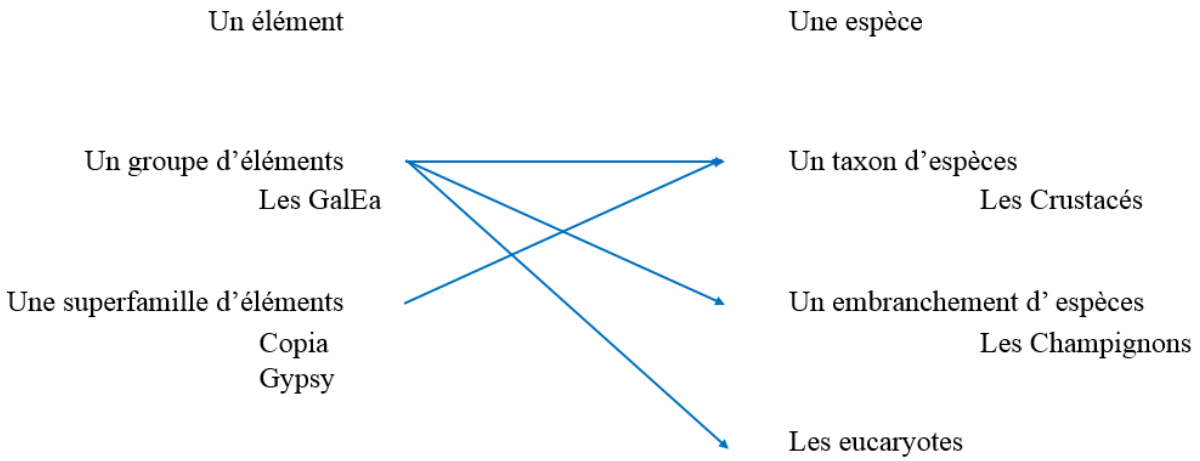


Figure 3: Représentation de l’étude de la distribution des éléments GalEa réalisé durant ma thèse

Chapitre II. Introduction

2.1 Description, diversité et classification des éléments

Les éléments transposables sont des acteurs majeurs de la diversité génétique. En effet, ils s'insèrent facilement au sein des génomes et souvent de façon répétée. Très diversifiés de par leur nature et leur nombre, la classification de ces éléments est donc nécessaire pour mieux les étudier. La classification des ETs se fait selon trois critères : leur mode de transposition, leur structure (présence et organisation des différents domaines codants, nature des terminaisons) et leur séquence nucléotidique (relations phylogénétiques). Le mode de classification est encore largement débattu, cependant beaucoup de chercheurs utilisent la classification proposée par Wicker et al. (2007) (Figure 4) qui les classe comme les êtres vivants, suivant des classes, ordres, superfamilles et familles. Nous ajouterons un niveau entre superfamilles et familles que nous appelons clades.

Suivant leur mode de transposition, on distingue deux classes : les transposons transposent *via* un intermédiaire ADN, sur le mode du « couper-coller ». Ils en existent 4 ordres dont un nettement majoritaire : les TIR qui présentent des Terminaisons Inverses Répétées, avec 9 superfamilles différentes, dont les plus connues sont celles des éléments *P* ou *Tc1/Mariner*. Les rétrotransposons transposent *via* un intermédiaire ARN sur le principe du « copier-coller » et ne sont présents que chez les eucaryotes. L'ADN est transcrit en ARN qui est rétrotranscrit en ADNc qui va lui-même être intégré au sein du génome. Parmi eux, on distingue les LINEs et les SINEs (Long et Short Interspersed Nuclear Elements). Ils se particularisent par une absence de terminaison particulière, même si les LINEs présentent une queue polyA. Les LINEs possèdent une endonucléase permettant l'intégration au sein des génomes et une transcriptase inverse permettant la transcription inverse directement au site d'insertion. Seuls les LINEs sont potentiellement autonomes. On en distingue au moins 5 superfamilles (Figure 4), avec différentes caractéristiques structurales. Les SINEs, eux, sont non autonomes et utilisent les LINEs pour leur transposition (Dewannieux et al., 2003; Kajikawa and Okada, 2002; Kramerov and Vassetzky, 2005). Ce sont de courtes séquences d'ADN, comprises entre 80 et 500 pb, non codants et sont transcrits par la RNA polymérase III. Trois superfamilles de SINEs peuvent être définies selon leur origine moléculaire, suivant qu'ils sont issus de la dimérisation et/ou trimérisation d'éléments SINE.

Classification		Structure	TSD	Code	Occurrence
Order	Superfamily				
Class I (retrotransposons)					
LTR	*Copia*	GAG AP INT RT RH	4–6	RLC	P, M, F, O
	Gypsy	GAG AP RT RH INT	4–6	RLG	P, M, F, O
	Bel–Pao	GAG AP RT RH INT	4–6	RLB	M
	Retrovirus	GAG AP RT RH INT ENV	4–6	RLR	M
	ERV	GAG AP RT RH INT ENV	4–6	RLE	M
DIRS	*DIRS*	GAG AP RT RH YR	0	RYD	P, M, F, O
	Ngaro	GAG AP RT RH YR	0	RYN	M, F
	VIPER	GAG AP RT RH YR	0	RYV	O
PLE	*Penelope*	RT EN	Variable	RPP	P, M, F, O
LINE	*R2*	RT EN	Variable	RIR	M
	RTE	APE RT	Variable	RIT	M
	Jockey	ORF1 APE RT	Variable	RIJ	M
	L1	ORF1 APE RT	Variable	RIL	P, M, F, O
	I	ORF1 APE RT RH	Variable	RII	P, M, F
SINE	*tRNA*		Variable	RST	P, M, F
	7SL		Variable	RSL	P, M, F
	5S		Variable	RSS	M, O
Class II (DNA transposons) - Subclass 1					
TIR	*Tc1–Mariner*	Tase*	TA	DTT	P, M, F, O
	hAT	Tase*	8	DTA	P, M, F, O
	Mutator	Tase*	9–11	DTM	P, M, F, O
	Merlin	Tase*	8–9	DTE	M, O
	Transib	Tase*	5	DTR	M, F
	P	Tase	8	DTP	P, M
	PiggyBac	Tase	TTAA	DTB	M, O
	PIF–Harbinger	Tase* ORF2	3	DTH	P, M, F, O
	CACTA	Tase ORF2	2–3	DTC	P, M, F
Crypton	*Crypton*	YR	0	DYC	F
Class II (DNA transposons) - Subclass 2					
Helitron	*Helitron*	RPA Y2 HEL	0	DHH	P, M, F
Maverick	*Maverick*	C-INT ATP CYP POL B	6	DMM	M, F, O

Structural features

Long terminal repeats — Terminal inverted repeats — Coding region — Non-coding region

Diagnostic feature in non-coding region — Region that can contain one or more additional ORFs

Protein coding domains

AP, Aspartic proteinase; APE, Apurinic endonuclease; ATP, Packaging ATPase; C-INT, C-integrase; CYP, Cysteine protease; EN, Endonuclease; ENV, Envelope protein; GAG, Capsid protein; HEL, Helicase; INT, Integrase; ORF, Open reading frame of unknown function; POL B, DNA polymerase B; RH, RNase H; RPA, Replication protein A (found only in plants); RT, Reverse transcriptase; Tase, Transposase (* with DDE motif); YR, Tyrosine recombinase; Y2, YR with YY motif

Species groups

P, Plants; M, Metazoans; F, Fungi; O, Others

Figure 4 : Classification des éléments transposables avec les différents domaines et structures selon Wicker et al., (2007)

Les rétrotransposons *sensus stricto*, se distinguent par la présence de terminaisons particulières. Les Pénélope (PLE) se différencient par des terminaisons répétées, et s'intègrent grâce à une endonucléase (Eickbush and Jamburuthugoda, 2008). Ils ne présentent qu'une seule phase ouverte de lecture, qui code une *pol* (polymérase) composée de deux domaines protéiques, une transcriptase inverse et une endonucléase (Evgen'ev and Arkhipova, 2005). Ils sont assez peu étudiés jusqu'à aujourd'hui, et se distinguent des autres rétrotransposons par leur capacité à maintenir des introns chez certains éléments (Arkhipova et al., 2003). Les éléments à Tyrosine Recombinase se différencient des autres rétrotransposons par des terminaisons répétées en sens direct ou indirect et un domaine Tyrosine Recombinase (YR) permettant l'intégration de l'élément (Goodwin and Poulter, 2001). La YR forme une troisième phase ouverte de lecture, en plus d'une *gag* classique et d'une *pol* présentant au moins deux domaines protéiques (transcriptase inverse et RNase H). A l'inverse des endonucléases ou intégrases, la YR n'entraine pas de duplication de la séquence au site cible où l'élément s'insère. Cette YR est proche de celles portées par les bactériophages lambda et les transposons Cryptons (Goodwin and Poulter, 2001; Goodwin et al., 2003). On distingue 4 différentes superfamilles d'éléments à YR, tels que les Ngaro, les Viper, les PAT et les DIRS1 qui se différencient principalement par la nature de leurs structures et de leurs terminaisons (Goodwin and Poulter, 2004). Les Ngaro (Figure 5), Viper et PAT présentent des terminaisons répétées en sens direct (SDR, Split Direct Repeats). Les Ngaro se distinguent principalement par l'absence de domaine méthyl-transférase (MT, dont le rôle est encore inconnu), au sein de leur *pol* (Poulter and Goodwin, 2005). Les rétrotransposons de type Viper, uniquement identifiés chez des espèces de trypanosomes (Lorenzi et al., 2006), ne présentent pas de MT et sont les seuls éléments à présenter la YR en 5' de la *pol*. Enfin, pour les éléments de type PAT, différentes analyses phylogénétiques ont montré qu'ils constituent le groupe frère (deux taxons pouvant être regroupés dans un même groupe monophylétique plus large) des éléments DIRS1 (Lorenzi et al., 2006; Poulter and Goodwin, 2005). Ils sont d'ailleurs régulièrement regroupés en un seul groupe d'éléments appelé 'DIRS'. Cependant, lorsque l'on regarde leurs différences structurales (notamment leurs terminaisons) et leurs relations phylogénétiques, ces deux groupes apparaissent bien distincts. En effet, une particularité de structure des rétrotransposons de type DIRS1 est la nature de leurs séquences répétées. Ils sont bordés par des Terminaisons Répétées en sens Inverse (ITR : «Inverted Terminal Repeats», Cappello et al., 1985; Zuker et al., 1984), (Figure 5). Ils possèdent également en amont de la terminaison droite des Régions

Complémentaires Internes (ICR : Internal Complementary Region), qui sont respectivement inversement complémentaires à l'un des deux ITRs. Toutes ces répétitions sont impliquées dans la formation de l'ADN circulaire double brin.

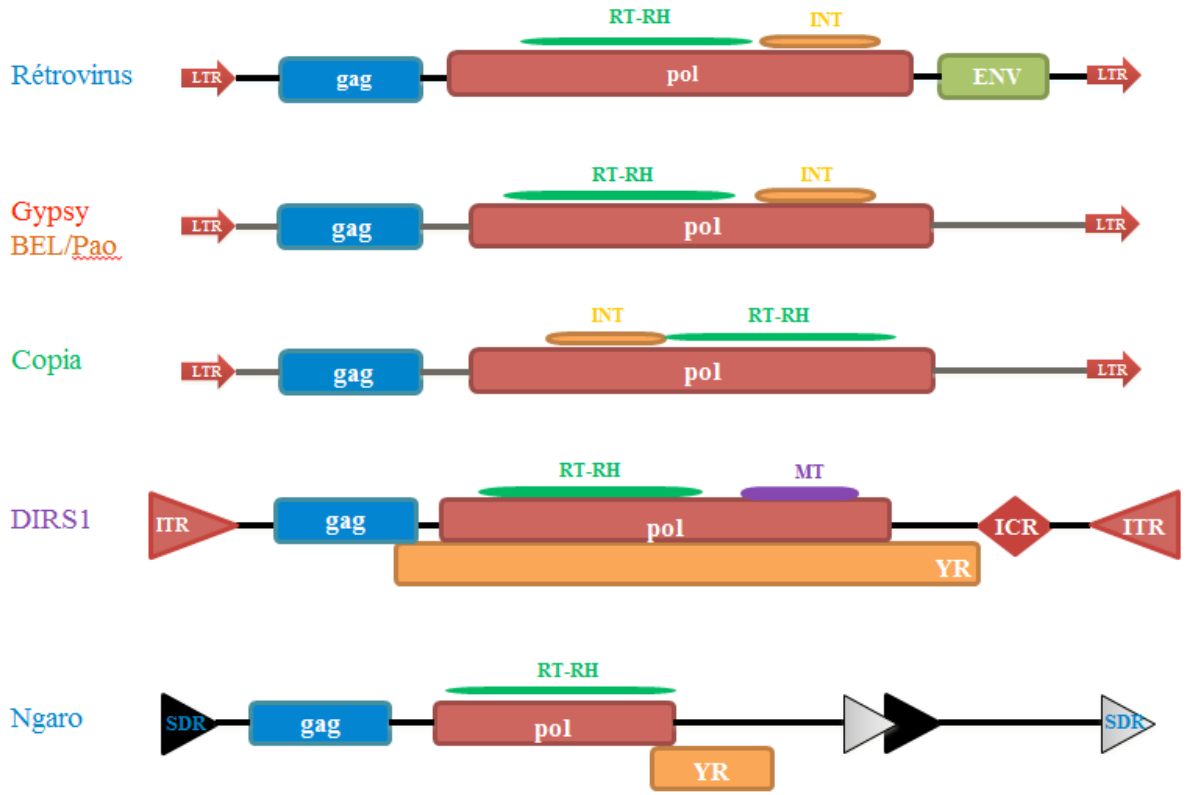


Figure 5: Structure des éléments à LTR et à YR. *Gag* : group specific antigens, *Pol* : polymérase, RT : Reverse Transcriptase, RH : RNaseH, MT : Methyl-transférase, INT : Intégrase, YR : Tyrosine Recombinase, LTR : Long Terminaison Repeats, ITR : Inverted Terminal Repeats, ICR : Internal Complementary Region, ENV : Enveloppe, SDR : Split Direct Repeat

Les rétrotransposons à LTR se différencient par de Longues Terminaisons Répétées (LTR) en orientation directe (de 100 pb à plus de 5 kb) qui jouent un rôle important dans la transposition des éléments car elles contiennent le promoteur, le terminateur de la transcription et des séquences régulatrices. Les rétrotransposons à LTR codent deux principales protéines *gag* (group specific antigens) et *pol*. La protéine *gag* code la polyprotéine à l'origine de la Virus Like Particule (VLP), siège de la reverse transcription. On retrouve au sein de la protéine *pol* au moins quatre domaines protéiques : une protéase aspartique (PR) responsable de la maturation des protéines nécessaires au mécanisme de réplication de l'élément, une transcriptase inverse (RT) responsable de la synthèse de l'ADNc à partir d'un des deux intermédiaires ARNs coencapsidés dans la VLP, une Rnase H (RH), responsable de la dégradation de l'intermédiaire

ARN lors de la synthèse de l'ADNc, ainsi que l'intégrase (INT) responsable de l'intégration de l'ADNc au sein du génome. Les éléments à LTR sont proches des rétrovirus et on distingue 3 principales superfamilles, les Copia, Gypsy et les Bel/Pao, qui se différencient principalement par la nature de leur structure (Figure 5). Chez les rétrovirus et quelques éléments Gypsy (plus rarement des éléments Copia et Bel/Pao) (Llorens et al., 2009), on retrouve un domaine supplémentaire *env* (enveloppe), pouvant conférer un caractère infectieux. Cette protéine d'enveloppe permet l'adhésion et la pénétration de l'élément au sein d'une nouvelle cellule cible (Varmus and Brown, 1989). Les éléments Copia se différencient des éléments Gypsy et Bel/Pao par la position de l'intégrase au sein de la *pol*, qui se situe en 5' pour les Copia et en 3' pour les Gypsy et Bel/Pao. La structure des éléments ne permet pas toujours de différencier toutes les superfamilles. Les Gypsy et les Bel/Pao présentent la même structure, ils se différencient alors par leur séquence. De plus, au sein d'une même superfamille, certains éléments peuvent présenter des structures divergentes. Par exemple, les éléments de type Gmr1 chez les Gypsy qui présentent une structure typique d'éléments Copia (Malik and Eickbush, 1999; Goodwin and Poulter, 2002). Pour classifier ces éléments, une étude phylogénétique des éléments et l'existence de groupes monophylétiques permettent le rattachement à une superfamille.

Au sein des superfamilles, on peut distinguer différentes familles de rétrotransposons par la divergence de leur séquence codante. Wicker et al., (2007) proposent de définir ces familles sur la base d'une similitude de séquences. Deux éléments de la même famille présentent une similarité de séquences d'au moins 80% d'identité, sur au moins 80% de la séquence alignée. Par la suite, les relations entre ces différentes familles peuvent être étudiées grâce à des analyses phylogénétiques, dont la plupart sont basées sur la séquence de la RT (Xiong and Eickbush, 1990).

En plus des ETs capables de transposer, il existe des éléments non autonomes, que ce soit au niveau d'une superfamille (aucun élément de la superfamille n'est autonome, par exemple les SINEs) ou au sein d'une famille (seules certaines copies ne sont pas autonomes). Chez les transposons, certains éléments sont non autonomes et doivent alors utiliser la transposase d'éléments autonomes. Par exemple, les MITEs (Miniature Inverted repeat Transposable Elements) qui sont de courtes séquences non codantes d'environ 400 pb, bordées par des séquences inversées répétées (Bureau and Wessler, 1992).

Compte tenu de la diversité des ETs décrits, nous ne pouvons pas tous les étudier. Nous avons

décidé de nous intéresser aux rétrotransposons à LTR et plus particulièrement aux éléments de la superfamille des Copia. Les rétrotransposons à LTR sont ubiquitaires au sein des eucaryotes, mais les différentes superfamilles peuvent se répartir différemment. On retrouve des éléments Gypsy et Copia de manière ubiquitaire au sein des eucaryotes et les Bel/Pao ont quant à eux été retrouvés uniquement chez les métazoaires avec une distribution assez inégale (de la Chaux et Wagner, 2011). La superfamille Ty1/Copia doit son nom aux premiers éléments décrits, chez *Drosophila melanogaster* (Copia) et au sein du génome de *Saccharomyces cerevisiae* (Ty1). Pour plus de simplicité nous appellerons cette superfamille, la superfamille Copia. Les Copia ont été retrouvés au sein des génomes eucaryotes dans des génomes de plantes et de champignons, au sein des straménopiles chez des diatomées (Maumus et al., 2011) et chez de nombreuses espèces de métazoaires comme des cnidaires, des nématodes, des crustacés et des hexapodes. On les retrouve également de manière plus parcellaire chez des téléostéens, des amphibiens et des squamates, mais aucun élément n'a été découvert chez les mammifères et les oiseaux (De la Chaux et Wagner, 2011). Les rétrotransposons sont généralement présents en plus grand nombre que les transposons au sein des génomes et peuvent atteindre un très grand nombre de copies. Les éléments Gypsy sont les éléments à LTR les plus abondants au sein des métazoaires, tandis que les éléments Bel/Pao apparaissent intermédiaires et les Copia moins fréquents (De la chaux et Wagner, 2011). Chez les champignons, les éléments Gypsy sont également plus présents au sein des génomes que les Copia, qui ont un nombre de copies assez faible au sein des génomes (entre 0 et 274 contre plus de 2500 pour les Gypsy dans certains génomes) (Muszewska et al., 2011). Les Copia sont par contre bien représentés au sein d'espèces de plantes (Navarro-Quezada and Schoen, 2002). On peut retrouver au sein des génomes, un nombre de copies variables pouvant aller de plusieurs millions pour la fève (*Vicia faba*), à environ 196 000 chez l'orge et 50 000 chez le maïs, pour arriver à une centaine de copies chez le riz (Wicker and Keller, 2007). Ils ont été retrouvés en faible nombre de copies au sein des diatomées, (53 pour *P. tricornutum* ; Maumus et al., 2009).

Les analyses phylogénétiques (très souvent réalisées sur le domaine *pol*, et plus particulièrement sur la RT-RH, mais également sur l'intégrase, domaines les plus conservés au sein des éléments) montrent que les éléments Copia sont divisés en plusieurs clades et révèlent une distribution hôte dépendante (Tableau 1). C'est-à-dire qu'un clade de Copia ne se retrouve que dans un taxon particulier. Par exemple on ne retrouve les éléments de type Osser, Tork, Retrofit, Sire et Oryco que chez des plantes et algues ; les éléments de type Pseudovirus et

pCreto chez les champignons, et les Hydra, GalEa, Copia, 1731, Tricopia, Mtanga et Humnum que chez des métazoaires.

Branch	Host Phyla	Genus	Clade	Env	Chromodomain
Branch 1	Fungi	Pseudovirus *(Pseudoviridae)*	*Ty (Pseudovirus)*	no	no
Branch 1	Diatoms *(Heterokontophyta)*		*CoDi-I* or *CoDi-A*	no	yes
Branch 1	Diatoms *(Heterokontophyta)*		*CoDi-II* or *CoDi-B*	no	no
Branch 1	Diatoms *(Heterokontophyta)*		*CoDi-C*	no	no
Branch 1	Diatoms *(Heterokontophyta)*		*CoDi-D*	no	no
Branch 1	Marine Arthropoda		*GalEA*	no	no
Branch 2	Fungi		*p-Cretro*	no	no
Branch 2	Land plants *(Viridiplantae)*	Sirevirus *(Pseudoviridae)*	*Sire*	yes	no
Branch 2	Land plants *(Viridiplantae)*	Sirevirus *(Pseudoviridae)*	*Oryco*	no	no
Branch 2	Land plants *(Viridiplantae)*		*Retrofit*	no	no
Branch 2	Land plants *(Viridiplantae)*		*Tork*	no	no
Branch 2	Green algae *(Viridiplantae)*		*Osser*	no	no
Branch 2	Red algae *(Rhodophyta)*		*PyRE1G1*	no	no
Branch 2	Cnidaria *(Metazoa)*		*Hydra*	no	no
Branch 2	Arthropoda	Hemivirus *(Pseudoviridae)*	*Copia*	no	no
Branch 2	Arthropoda		*1731*	no	no
Branch 2	Arthropoda		*Tricopia*	no	no
Branch 2	Arthropoda		*Mtanga*	no	no
Branch 2	Arthropoda		*Humnum*	no	no

Tableau 1 : Les éléments Copia (Llorens, 2009)

2.2 Les éléments GalEa

Un des tout dernier clade de Copia défini est celui des GalEa, dans un groupe particulier de crustacés décapodes (Terrat et al., 2008). La description des premiers éléments GalEa a été faite chez des galathées et notamment chez *Eumunida annulosa.* D'après l'analyse de la structure et des relations phylogénétiques, les GalEa sont bien des rétrotransposons de type Copia. L'élément chimérique *GalEa1* de *E. annulosa* a été décrit entièrement (Figure 6a). L'élément *GalEa1* mesure 4669 pb avec une partie interne de 4421 pb. Les LTRs font 124 pb, commencent en 5' par TG et finissent en 3' par CA (comme observé dans de nombreux rétrotransposons). A la position 126, la partie interne porte un Primer Binding Site (PBS) dont la séquence (TGGTAGCAGAGC) est complémentaire de la région 3' terminale du gène ARNt^{Met} de *D. melanogaster*. Cette séquence est bien conservée entre les différents GalEa alors

décrits. Il a également été décrit un signal polypurine putatif (PPT: GAAGAAATGGA) à la position 4522. La partie centrale de *GalEa1* comprend une seule grande phase ouverte de lecture qui présente les 5 domaines typiques ordonnées des rétrotransposons Copia : la *gag* et les domaines de la région *pol* (dont les motifs conservés sont représentés sur la Figure 6). Le 1ier domaine contient un motif zinc finger (CX2CX4HX4C) que l'on retrouve dans de nombreux gènes *gag* rétroviraux. Le deuxième domaine est le domaine PR dont le motif typique DSGA des rétrotransposons Copia est substitué par un motif DSGC. Le troisième domaine est le domaine INT avec le motif HX4HX30CX2C et les signatures DD35E. Le quatrième domaine est le domaine RT contenant sept sous-domaines conservés dans toutes les séquences RT (Capy et al., 1997; Xiong and Eickbush, 1990). Le cinquième domaine correspond à la RH avec le motif TRPDI hautement conservé.

D'autres éléments *GalEa1* ont été retrouvés au sein des génomes d'autres galathées, *Agononida laurentae* par exemple. Des recherches ont été réalisées dans d'autres génomes, afin de préciser la répartition des éléments GalEa, en utilisant les domaines protéiques de la *pol* de *GalEa1* comme requête pour un TBLASTN. Cela a permis de caractériser quatre nouveaux rétrotransposons: *Cico1* (DQ913003) et *Cico2* (DQ913004) chez un urochordé *Ciona intestinalis*, *Zeco1* (DQ913001) chez le *Danio rerio* et *Olco1* (DQ913000) dans un autre téléostéen *Oryzias latipes*. Toutes les séquences codantes obtenues sont remaniées par des changements de cadres de lecture et des codons stop, ce qui suggère que les copies décrites ne sont plus actives au sein des génomes hôtes. Les principales caractéristiques de ces rétrotransposons GalEa sont présentées dans la Figure 6b.

Les longueurs des éléments *Zeco1, Cico1* et *2* et *Olco1* (de 4500 à 4800 pb) sont semblables à celles de *GalEa1* d'*E. annulosa*, même si leurs LTR sont plus longs (187 à 323 pb). De par la méthode utilisée pour les rechercher, ces éléments partagent de nombreuses caractéristiques avec *GalEa1*, comme les LTR bordés par 5'-TG et CA-3 '; un homologue du PBS ARNtMet; un grand ORF unique contenant un motif zinc-finger (CX2CX4HX4C) dans la région *gag*, les signatures HHCC et DD35E de l'intégrase et le motif KARLVA de la RT. Cependant, chaque élément présente quelques particularités. Tous les quatre ont un motif DTAC dans la région de codage de la PR, les éléments *Zeco1* et *Olco1* ont un motif HVDD (au lieu de YVDD) au sein de la RT et Olco1 présente un motif SRPDV (au lieu de TRPDI) au sein de la RH. La comparaison des séquences LTR 5 'et 3' a révélé 100% d'identité pour *Cico1* et *Zeco1*, ce qui suggère que ces éléments peuvent avoir été récemment actifs. Cette hypothèse est étayée par le

fait que les transcrits de *Cico1* ont été également détectés dans des bases de données. Enfin, l'analyse de la séquence nucléotidique de sept copies de *Zeco1* a révélé qu'ils sont flanqués par une duplication au site cible de 5 pb ce qui est commun pour les éléments Copia.

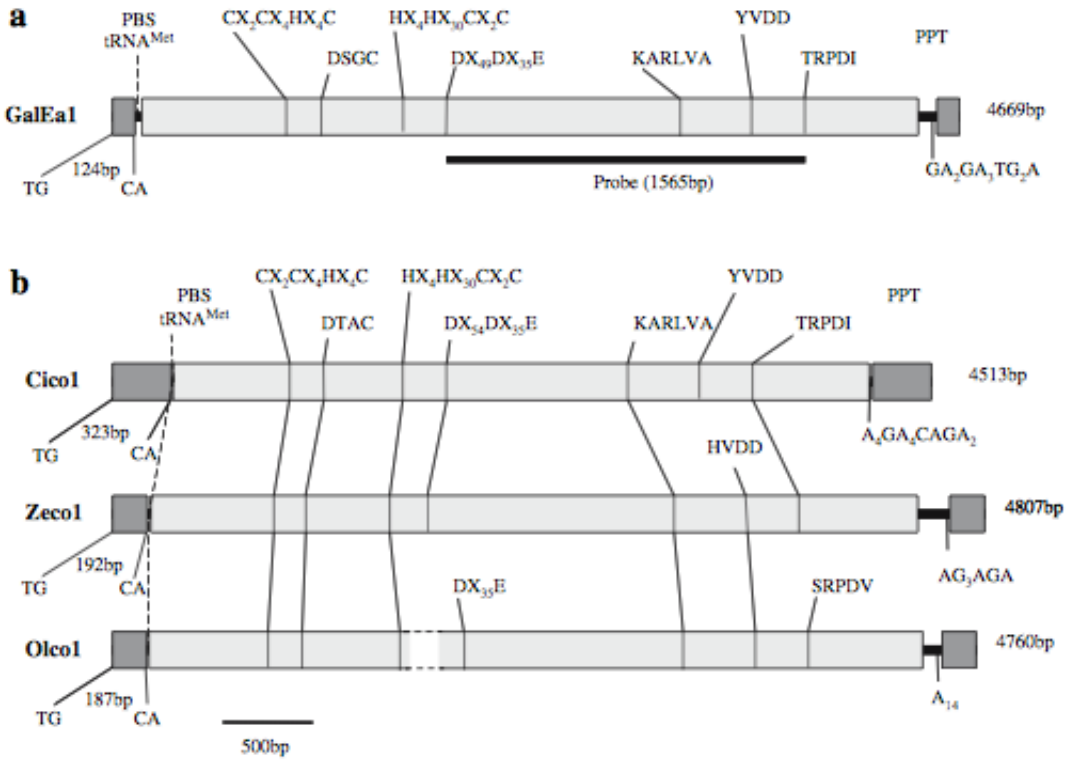


Figure 6 : Organisation structurelle du rétrotransposon GalEa1 de *E. annulosa* (EU097705) et des autres rétrotransposons GalEa. Les boîtes gris foncé indiquent de longues répétitions terminales (LTR), les rectangles gris clair indiquent l'ORF. Les séquences d'acides aminés codant pour les motifs ou les signatures de rétrotransposons Copia sont présentés au-dessus des éléments ainsi que le PBS, site de liaison de l'amorce et le signal PPT polypurine. (Terrat et al., 2008)

Pour évaluer la relation entre différents éléments GalEa, une analyse phylogénétique basée sur les séquences RT-RH en utilisant les copies qui ont été caractérisées au niveau moléculaire chez les galathées et les copies extraites de recherche BLAST a été réalisée (Terrat et al., 2008). Il a été observé deux groupes monophylétiques; le premier contient les éléments de galathées et le second ceux des chordés (de la cione et des téléostéens). Pour analyser la parenté des éléments GalEa avec d'autres rétrotransposons, une analyse phylogénétique de séquences protéiques a été menée (Figure 7). Elle confirme que les éléments GalEa se groupent avec les rétrotransposons Copia. Cependant, la topologie observée révèle deux groupes bien distincts: les GalEa sont séparés de tous les autres éléments Copia et définissent un nouveau clade à part

d'éléments Copia.

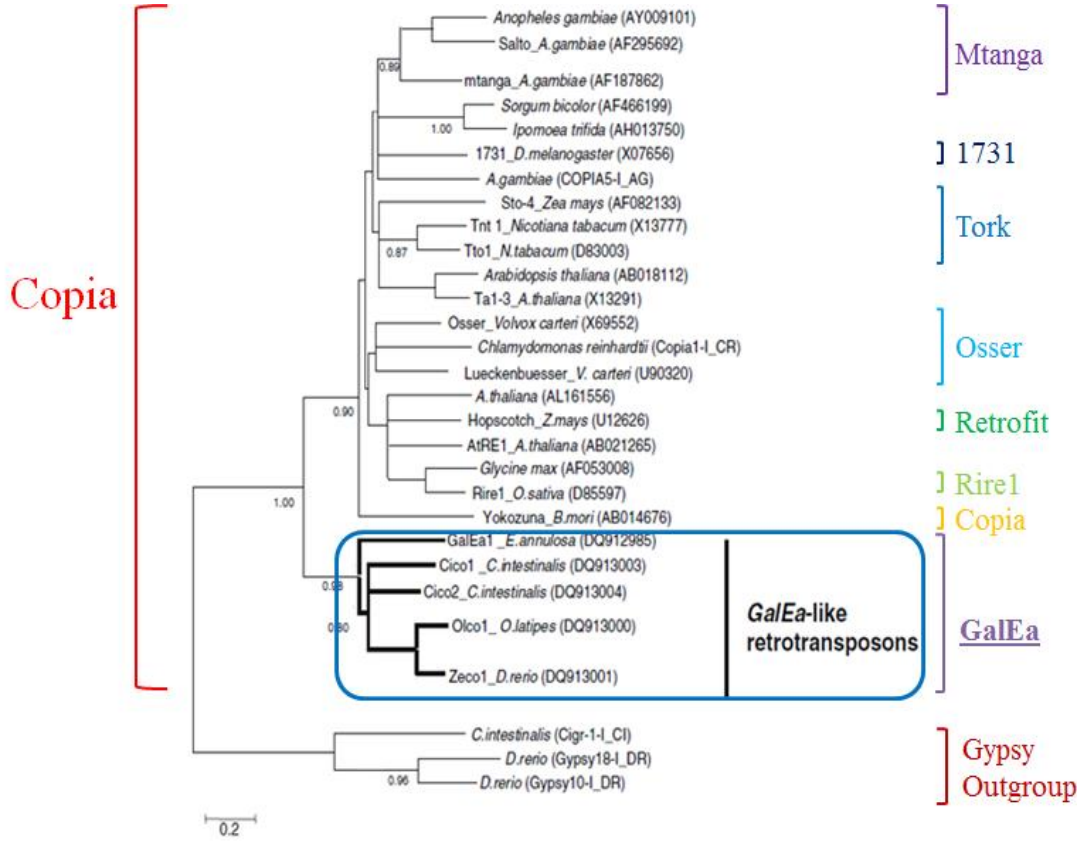


Figure 7 : Phylogénie des éléments Copia basée sur la séquence en acides aminés de la RTRH des éléments. Le nom du clade auquel appartiennent les éléments est indiqué à droite. Arbre enraciné grâce aux éléments Gypsy en outgroup. (D'après Terrat et al, 2008)

Parmi toutes les caractéristiques des éléments GalEa, l'une des plus marquantes est leur distribution complexe qui présente trois particularités: ils sont présents dans des espèces phylogénétiquement éloignées, d'une manière discontinue et avec apparemment une restriction à des espèces aquatiques. La répartition des éléments GalEa est très discontinue car limitée à certains taxons éloignés de bilateriens tel que les urochordés et les téléostéens (poissons). Mais cette discontinuité apparaît également au sein des téléostéens, car les GalEa sont introuvables chez *Takifugu rubipes* et *Tetraodon nigroviridis*, deux espèces proches d'*O. latipes.* Deux hypothèses peuvent expliquer cette distribution morcelée: les transferts horizontaux et/ou les pertes stochastiques. Volff et al., (2000) suggèrent que les rétrotransposons sont capables de transferts horizontaux entre taxons apparentés. Cela pourrait également être le cas pour les rétrotransposons GalEa de façon plus fréquente qu'on ne le pensait. La perte différentielle peut aussi être considérée comme un facteur compte tenu du faible nombre de copies (une dizaine

chez *D. rerio*). Ce faible nombre de copies pourrait expliquer pourquoi les GalEa n'étaient auparavant pas décrits en dépit des recherches intensives sur ces organismes modèles. Quelle que soit l'hypothèse envisagée, la vision actuelle est fortement influencée par les phyla pour lesquels il y a des génomes disponibles. Par conséquent, pour une meilleure compréhension des mécanismes sous-jacents de l'évolution, il est encore nécessaire d'affiner la répartition des éléments GalEa; ce qui permettra en outre de confirmer la «nature aquatique» des GalEa.

2.3 Les espèces hôtes et leurs génomes

L'étude de la distribution des ETs suppose de bien connaître les relations phylogénétiques des espèces hôtes. Au sein des êtres vivants, on distingue 3 grands domaines: les eubactéria, les archaea et les eucaryota. Les eucaryotes présentent des caractères dérivés propres tels que la présence de mitochondries au sein des cellules et l'ADN contenu dans un noyau délimité par une enveloppe nucléaire. Lors de la division cellulaire, cet ADN est divisé et compacté en chromosomes. On dénombre plus de 1,8 10^6 espèces d'eucaryotes divisées en 2 clades principaux: les unikontes et les bikontes (Figure 8). Au sein des unikontes on retrouve les amibes et les opisthokontes: champignons et métazoaires, bien connus et bien étudiés en biologie. Les bikontes sont composés des archaeplastida ou plantes et algues également bien connus, mais aussi d'autres groupes assez peu connus tel que les excavates (protistes hétérotrophes et généralement flagellés, par exemple les trypanosomes et les euglénoides.) et les chromoalveolates regroupant entre autres les ciliées et les straménopiles (algues brunes, diatomées et oomycètes).

Depuis le séquençage du premier génome eucaryote en 1996 (*Saccharomyces cerevisiae*, une levure), le nombre d'espèces dont le génome est complètement séquencé est en croissance constante et la liste de projets de séquençage ne fait qu'augmenter. La Figure 8 présente un état des lieux en 2009. Par contre, les études portant sur les génomes complètement séquencés sont en partie biaisées par l'échantillonnage avec des taxons hyper représentés et d'autres sous représentés. En effet, certains groupes d'espèces sont très étudiés en biologie car ils ont un intérêt direct pour des applications médicales, par exemple le groupe des métazoaires qui regroupent un bon nombre d'espèces modèles, comme les mammifères. D'autres ont un intérêt agronomique ou économique comme le groupe des plantes ou bien celui des champignons, qui regroupent des espèces pathogènes des plantes ou bien des espèces qui participent à

l'élaboration du pain, vin, bière ou encore du fromage. Les mammifères (environ 4 500 espèces) présentent en 2015 une soixantaine de génomes complètement séquencés. Alors que pour certains taxons comme celui des chondrichtyens (846 espèces), des chélicérates (74 450 espèces) ou encore des oiseaux (9 672 espèces), 1 seul génome complètement séquencé est pour le moment disponible. Les taxons des mammifères et des ascomycètes (grâce au projet 1000 génomes de champignons) présentent, à eux seuls, près de la moitié des génomes eucaryotes complètement séquencés. Les espèces ayant leur génome complètement séquencé sont aussi des espèces choisies sur des critères comme la taille de génome ou le fait que ce soit des espèces modèles bien étudiées en génétique. Ainsi sur les 27 espèces d'hexapodes la moitié appartient au seul genre *Drosophila*.
Au cours de ma thèse nous avons travaillé sur des grands groupes d'eucaryotes importants en nombre d'espèces mais dont on a très peu de génomes complètement séquencés tels que crustacés, mollusques ou rhodophytes, et qui sont également peu étudiés pour les ETs.

Pour plus de compréhension, j'appelle phylum, le niveau le plus large que nous étudions après les eucaryotes, c'est à dire les métazoaires et les champignons par exemple. Puis viennent au niveau intermédiaire les embranchements : les crustacés, les mollusques ou les ascomycètes. Enfin j'utilise le terme de taxon pour désigner un groupe d'espèces, quel que soit son niveau phylogénétique.

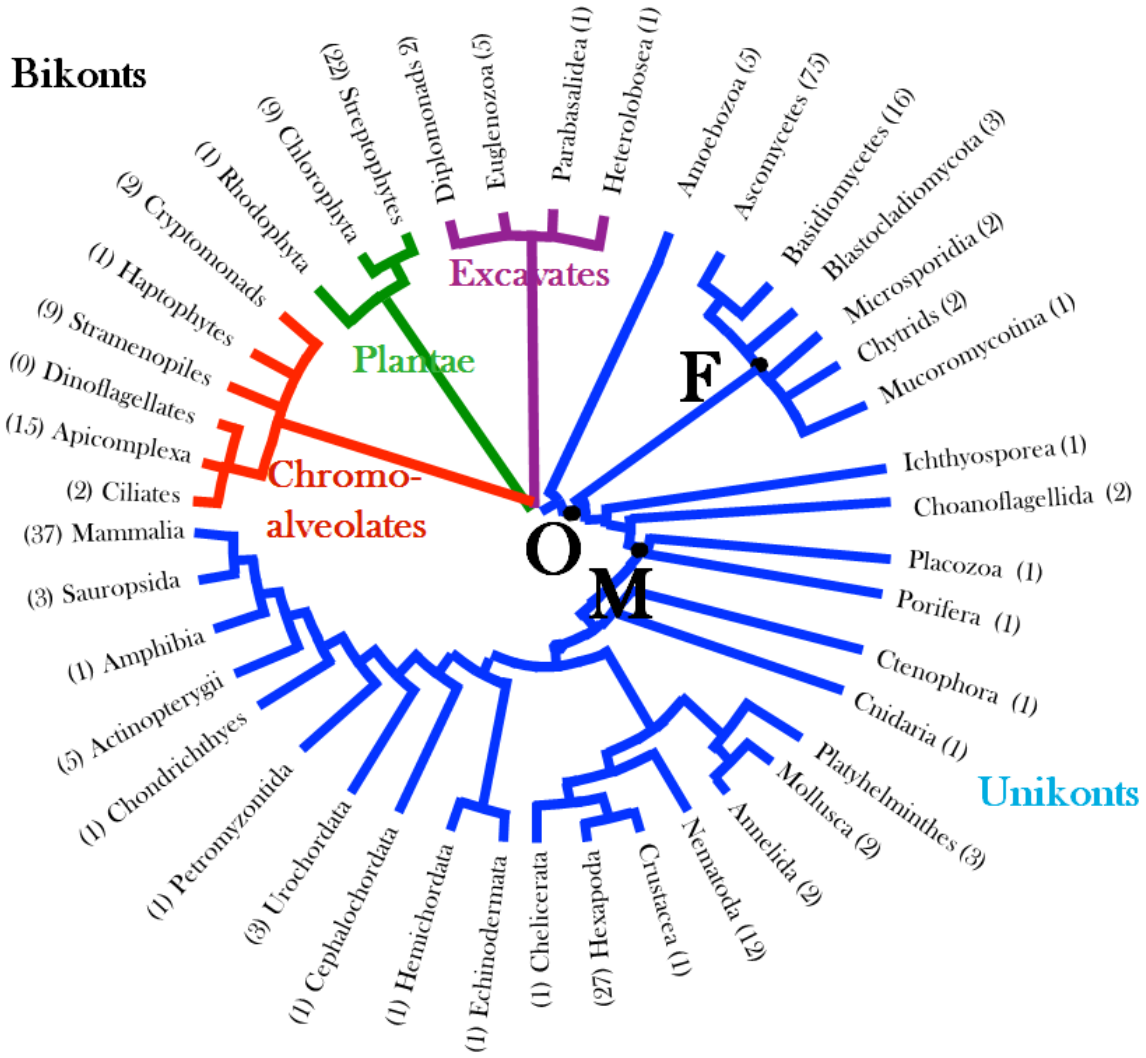


Figure 8: Arbre des grands phyla d'eucaryotes. La phylogénie des espèce a été redessiné à partir de [Hibbett DS et al., 2007; Keeling PJ et al., 2005; Dunn CW, et al., 2008; Philippe H, et al., 2009]. F : Champignon, O : Opisthokontes, M : Métazoaires. Dans chaque groupe, nous incluons entre parenthèses le nombre d'espèces avec génome séquencés en 2009

2.4 Comment rechercher des ETs au sein des espèces

Lors de la recherche d'ETs au sein des espèces d'intérêt, on doit tout d'abord savoir si on s'intéresse à une ou plusieurs espèces et à la recherche d'un ou plusieurs type d'élément car il existe 2 grands types d'approches, l'approche "humide" et l'approche *in silico*.

- L'approche "humide"

Cette méthode suppose que l'on ne possède pas ou peu de données de séquençage pour la ou les espèces d'intérêt. Lors de la recherche d'une famille particulière de rétrotransposon, on

procède à une PCR avec des amorces dégénérées. Les amorces dégénérées sont définies, le plus souvent afin de couvrir une partie assez grande de la RT/RH, et leur efficacité testée. Elles sont obtenues à partir d'un alignement protéique de séquences d'éléments déjà décrits. A partir de ces alignements protéiques, les motifs protéiques les mieux conservés sont sélectionnés pour définir la séquence des oligonucléotides dégénérés (plusieurs oligonucléotides au sein d'une même synthèse). L'intérêt de faire varier le niveau de dégénérescence pour un motif donné résulte dans l'équilibre entre une meilleure efficacité (niveau de dégénérescence plus bas) et la capacité de détecter des éléments qui sont éloignés des éléments utilisés pour l'alignement (niveau de dégénérescence plus fort). On module le niveau de dégénérescence suivant que l'on recherche des éléments de la même superfamille (forte dégénérescence) ou de la même famille (faible dégénérescence). Des amorces non dégénérées peuvent être définies à partir de motifs très conservés, qui nous servent lors de la recherche d'un même élément. Plusieurs couples d'amorces différents sont utilisés afin d'optimiser la détection des éléments au sein de nouvelles espèces. Si l'on souhaite caractériser des éléments entiers et pouvoir réaliser une analyse phylogénétique, une approche de marche par PCR permet d'étendre les séquences à partir de chaque fragment initial. Pour cela, nous utilisons la méthode du TE Walking (marche sur l'élément, figure 9),(Piednoël and Bonnivard, 2009). Celle-ci consiste à utiliser une amorce spécifique définie sur la séquence nouvellement caractérisée, qui sert de point d'ancrage, et de l'associer à une nouvelle amorce définie sur un motif conservé éloigné du précédent (en 5' ou en 3'). Ces amorces spécifiques sont généralement définies dans les régions les plus terminales du fragment afin d'optimiser la taille des produits d'amplification à séquencer. Cependant, elles doivent aussi être éloignées des extrémités d'une centaine de pb pour que l'on puisse observer un chevauchement d'au moins 50 pb entre la séquence connue et la nouvelle séquence caractérisée par la marche. Le pourcentage d'identité au niveau du chevauchement permettra de confirmer ou non que l'on progresse bien sur la séquence voulue (seuil supérieur à 95% d'identité). La séquence est ensuite reconstruite à l'aide du programme Cap-contig inclus dans le logiciel BioEdit. On récupère grâce à cette technique des éléments qui sont bien souvent des chimères. On appelle «élément chimère», un élément obtenu à partir de différentes séquences suite à des PCR, reconstruit grâce à des copies différentes d'un même élément. L'intérêt de cette méthode est de pouvoir caractériser des éléments dans des espèces non modèles, par exemple les espèces pour lesquelles il n'existe aucune donnée génomique dans les banques de données. Toutefois cette méthode limite le nombre d'espèces étudiables.

On peut analyser une trentaine d'espèces tout au plus, car cela prend énormément de temps. Une autre limite est la capacité de détection car si on n'obtient pas de résultats plusieurs questions se posent alors à nous: l'élément est-il présent? Les éléments sont-ils trop dégénérés par rapport à ceux avec lesquels on a définis nos amorces? etc. En effet, la PCR est une méthode sensible. Enfin, cette méthode nous permet de récupérer des éléments chimères et non une même copie.

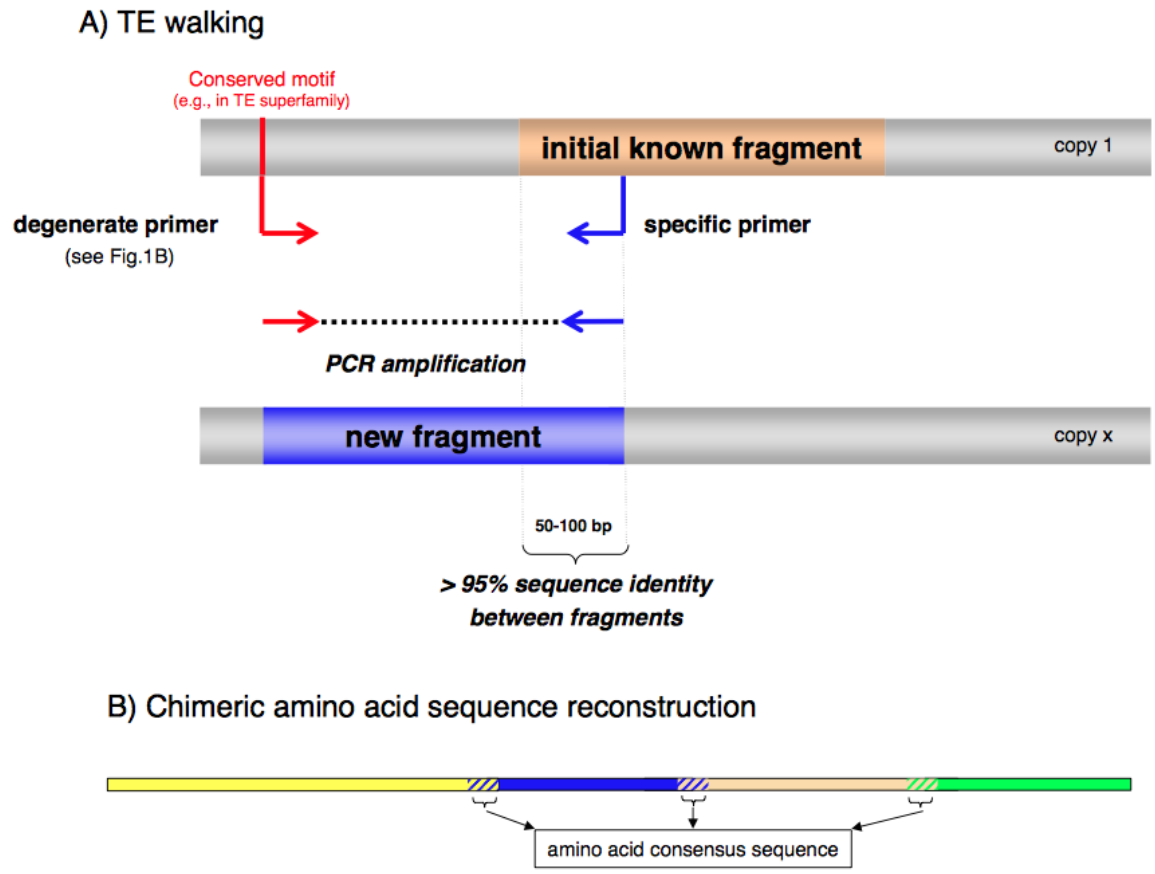


Figure 9: Méthodologie du TE Walking. Résumé de la méthodologie utilisée pour allonger la séquence de l'élément transposable.

- L'approche *in silico*

L'approche *in silico* à l'avantage de permettre d'utiliser un bon nombre de données génomiques disponibles dans les bases de données. Cependant, on ne choisit pas les espèces pour lesquelles ces données existent, à moins de faire séquencer soi-même son espèce d'intérêt. Il existe une grande diversité de données dans les bases de données qui implique que les résultats obtenus donneront des informations différentes.

Tout d'abord il existe des bases de données avec des données de transcriptomique (ESTs) ou de génomiques partielles (BACs ou séquençages partiels de génomes). Ces bases de données permettent d'avoir des informations sur la présence de l'élément au sein des génomes, mais ne permettent pas de conclure sur l'absence. On obtient les mêmes types de résultats que lors de l'utilisation de l'approche "humide". Néanmoins, les données transcriptomiques permettent d'obtenir des informations complémentaires telle que l'activité de l'élément car s'il est transcrit il est toujours potentiellement actif. De plus, contrairement à l'approche "humide", nous pouvons potentiellement récupérer des éléments entiers, en tout cas non chimérique. Et par extrapolation, il peut être possible d'estimer grossièrement le nombre d'éléments à partir de séquençage même partiel. Il existe également des bases de données regroupant des génomes complets tel que NCBI (http://www.ncbi.nlm.nih.gov/genome/), Broad institute (https://www.broadinstitute.org/scientific-community/data) ou encore JGI (http://genome.jgi.doe.gov/). Ce type de données permet de récupérer bien plus d'informations. En effet, on peut conclure à une réelle présence ou absence des éléments au sein des génomes d'intérêt ; avoir une estimation précise du nombre de copies, leur part au sein des génomes, leur position au sein des chromosomes etc. Avec ces données *in silico*, nous pouvons grâce à différentes méthodes de recherche, par différents logiciels, récupérer les informations importantes pour l'étude. Des logiciels comme REPET (Flutre et al., 2011) servent à rechercher de nouveaux éléments non encore décrit grâce à la répétition de ces éléments au sein de génomes complètement séquencés. Lorsque nous possédons des éléments de références, Blast permet de rechercher un élément par similarité de séquence au sein de bases de données génomiques ou transcriptomiques. Il est également possible d'utiliser des logiciels tel que RepeatMasker (RM) ((Tarailo-Graovac and Chen, 2009; http://www.repeatmasker.org/), qui permet de repérer dans un génome complet, ou partiel, les séquences d'éléments préalablement répertoriés dans une base de données, comme celles déjà établie par RepBase (http://www.girinst.org/repbase/) ou une base « maison ». On peut ainsi utiliser une base de données d'ETs réalisée par nous même avec des éléments choisis (par exemple des éléments nouvellement décrits et non encore déposés dans RepBase). Basé sur une recherche par similarité entre séquences, RM s'affranchit de toute notion de structure. En fonction des filtres utilisés ultérieurement, RM permet donc de dénombrer des copies d'ETs (en fixant par exemple une limite de taille), de rechercher des éléments délétés, voire même de petits fragments ou traces. D'autres logiciels utilisent des données structurales. Par exemple,

LTR Harvest (Ellinghaus et al., 2008) permet de détecter des séquences directement répétées, comme des LTR, au sein d'un génome et donc de rechercher des éléments correspondant, des rétrotransposons à LTR. Les séquences récupérées peuvent par la suite être annotées par Blast pour les classer, par exemple au niveau des différentes superfamilles de rétrotransposons à LTR : Copia, Gypsy ou Bel/Pao. Cet outil permet de récupérer des éléments avec une structure particulière (ici des LTR) et d'estimer le nombre de copies, *a priori* complètes, d'un rétrotransposon. Bien sur les éléments dont les LTRs ont trop divergées ou ont disparues ne pourrons pas être récupérer par LTR Harvest. Enfin, des logiciels permettent de retrouver des éléments avec une structure et des séquences très bien conservées, comme ReDoSt (Piednoël et al., 2011). Des profils sont créés à partir de séquences de DIRS1 connues, et l'ordre des différents domaines RT, RH, YR et MT les uns par rapport aux autres est recherchés. ReDoSt permet de récupérer des copies entières d'éléments et de dénombrer les copies au sein des génomes. Les limites de l'approche *in silico* résident dans les limites de chacun des logiciels utilisés pour la recherche des éléments. Mais un bon nombre des limites peuvent être contrées grâce à la combinaison de différents logiciels. De plus, ces logiciels nous permettent de rechercher des éléments au sein d'un grand nombre d'espèces ce qui est un avantage lors d'études comparatives.

2.5 Exemples d'utilisation de génomes complets

Grâce aux génomes complètement séquencés, la recherche des éléments transposables et surtout les résultats que l'on peut en tirer ont bien évolué. En effet, lorsque l'on possède le génome d'une espèce d'intérêt, nous pouvons aujourd'hui penser récupérer tous les ETs connus de ce génome, c'est ce qu'on appelle le «mobilome», et estimer la part du génome de ces éléments. C'est ce qui a été fait, par plusieurs équipes indépendamment, lors de la sortie du génome d'une espèce de crustacés modèle: la daphnie (*Daphnia pulex*) en 2010. Schaack et al., (2010) ont étudié les transposons et ont trouvé 56 familles appartenant à 10 des superfamilles connues. Ils ont chiffré à 1 466 236 pb, la portion du génome correspondant à des transposons soit 0,7%. Quant à eux, Rho et al., (2010) ont recherché les rétrotransposons à LTR. Ils ont mis en évidence la présence d'éléments Bel/Pao, Copia, Gypsy et DIRS1 au sein du génome, qui représentent 7,9% du génome. Ils ont également fait une étude phylogénétique de ces rétrotransposons, ce qui a permis de confirmer que, malgré leur structure identique, les

éléments Gypsy et Bel/Pao sont plus éloignés que ne le sont les éléments Gypsy et DIRS1 lorsque l'on regarde leur *pol* (Figure 10).

Un génome complètement séquencé permet aussi de connaitre la diversité des éléments présents. On peut savoir à quelles superfamilles et à quels clades ils appartiennent ; s'ils semblent intègres et donc potentiellement capables de transposer ; s'ils sont tronqués. On estime alors le nombre de copies complètes et incomplètes de chaque élément. Par exemple, une étude complète des éléments euchromatiques de *D. melanogaster* (20% du génome) résume les différentes superfamilles d'éléments présents, le nombre total d'éléments et le nombre d'éléments pleine taille (Tableau 2) (Kaminker et al., 2002). Elle présente aussi la taille des éléments décrits et leur répartition sur chaque chromosome. Lorsque l'on possède un génome très bien assemblé, avec les différents chromosomes définis, on peut étudier le positionnement et la part des éléments au sein des chromosomes. C'est ce qui a été fait chez la drosophile lors de la même étude que précédemment (Tableau 3).

Les génomes d'individus de la même population, de la même espèce ou du même genre, rendent possible des études comparatives de mobilomes à différentes échelles. Chez *D. melanogaster*, il existe de nombreux individus de populations différentes, dont le génome a été séquencé pour réaliser des études comparatives (The *Drosophila melanogaster* Genetic Reference Panel ; Mackay et al., 2012). On peut aussi faire des études de génomique comparative d'ETs au sein d'un genre. Il existe de nombreuses données pour le genre *Drosophila* avec au moins une vingtaine d'espèces différentes séquencées. Ces études permettent de connaitre entre autres, la distribution des ETs et le variation du nombre de copies, de remarquer les éléments ubiquitaires, de percevoir les transferts horizontaux possibles, et de comprendre la dynamique d'ETs à petite échelle.

Arabidopsis thaliana et *Arabidopsis lyrata* sont deux plantes modèles dont les génomes sont bien étudiés, notamment vis à vis des ETs. Cela permet d'appréhender les évènements intervenus lors de la spéciation à partir des variations entre les 2 génomes: duplications de gènes et réarrangements, régulations épigénétiques, réponses aux stress, etc. D'après Joly-Lopez and Bureau, (2014), 15% à 24% du génome d'*A. thaliana* est constitué d'ETs contre 25% à 30% chez *A. lyrata*. Les transposons constitueraient 50% de la part des ETs. Ces espèces présentent une très grande plasticité de génome, où l'on a mis en évidence de nombreux évènements de «burst» de transposition d'ETs. Les études comparatives peuvent également se faire au niveau de la famille des espèces. Dans une étude réalisée chez *Capsella*

rubella, A. thaliana et *A. lyrata,* les auteurs ont recherché la part des génomes constitué par de nombreux éléments tels que les Copia, Gypsy, LINEs, SINEs et des transposons (Figure 11). D'après cette étude, *A. lyrata* a connu une amplification des ETs au sein de son génome par rapport aux 2 autres espèces. En effet, la quantité d'éléments varie d'un génome à l'autre même chez des espèces proches comme *A. thaliana* et *A. lyrata*, et cela pour toutes les catégories d'ETs.

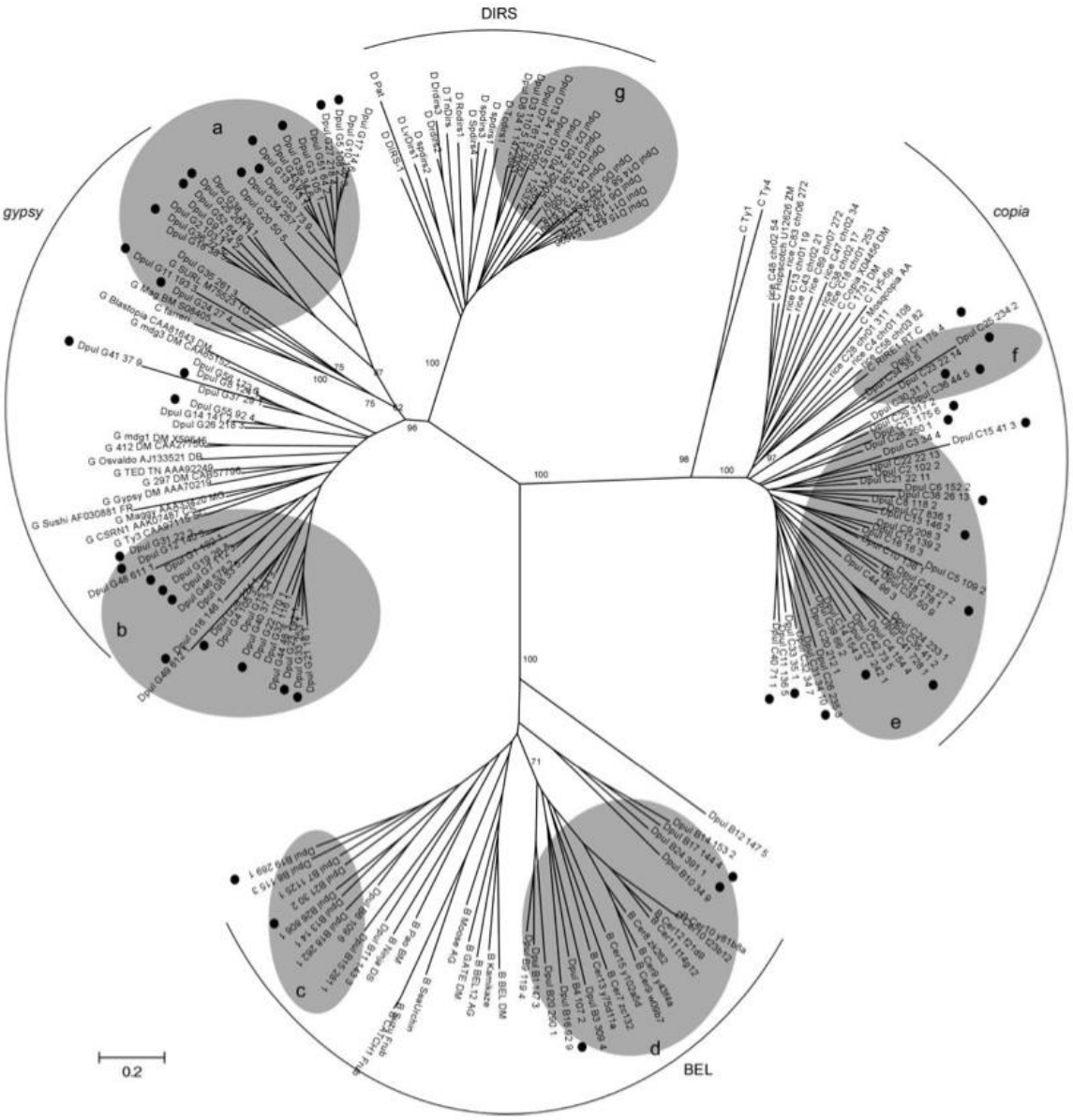


Figure 10 : Arbre phylogénétique des rétroéléments LTR dans le génome de *D. pulex* basé sur la RT des éléments. Eléments Gypsy, DIRS1, Copia et Bel/Pao. Les familles qui ont des éléments actifs au niveau de la transcription sont marquées par des cercles noirs. Les éléments identifiés *D. pulex* sont marqués en cercles gris. (Rho et al., 2010)

Class	Family	Canonical length	X	2L	2R	3L	3R	4	Total number	Number full length	Number partial	Number in proximal 2 Mb	Average pairwise distance
LTR	*17.6*	7439	2	0	3	5	2	0	12	7	5	4	0.006
	1731	4648	1	0	0	0	1	0	2	1	1	1	0.000
	297	6995	22	12	6	7	10	0	57	18	39	12	0.032
	3S18	6126	4	0	1	1	0	0	6	4	2	3	0.075
	412	7566	8	0	7	11	5	0	31	24	7	6	0.024
	accord	7404	0	0	1	0	0	0	1	0	1	0	-
	aurora	4263	0	0	2	1	0	0	3	1	2	3	0.074
	blastopia	5034	5	2	7	1	2	0	17	13	4	4	0.016
	blood	7410	1	11	2	3	5	0	22	22	0	6	0.001
	Burdock	6411	2	4	4	0	3	0	13	7	6	4	0.002
	Circe	6356	0	0	2	0	0	0	2	0	2	2	0.057
	copia	5143	4	13	4	5	4	0	30	26	4	3	0.002
	diver	6112	1	1	3	1	3	0	9	9	0	1	0.002
	diver2	4917	0	4	3	2	0	0	9	0	9	9	0.032
	Dm88	4558	0	0	2	0	30	0	32	0	32	2	0.015
	frogger	2483	0	1	0	0	0	0	1	1	0	1	-

Tableau 2 : Les éléments transposables de *D. melanogaster*

La longueur canonique de chaque élément (en pb) est indiquée dans la colonne 3, le nombre total de chaque famille sur chaque bras de chromosome dans les colonnes 4-9, les totaux pour chaque famille dans la colonne 10, et le nombre de ceux qui sont pleine taille, et partielle et dans les 2 Mb plus proximaux des principaux bras chromosomiques, dans les colonnes 11-13. Les éléments partiels sont définis comme ceux dont la longueur est inférieure à 97% de l'élément canonique. La distance moyenne par paires au sein de chaque famille est montrée dans la colonne 14. (extrait Kaminker et al., 2002)

Class	Arm	Total transposable element sequence (in bp)	% of arm	Total number of transposable elements	Number full length	% Full length	Number of transposable elements per Mb in genome	Number of transposable elements per Mb in proximal 2 Mb
All families	X	828,370	3.80	276	83	30.43	12.67	50
	2L	878,471	3.95	305	100	32.79	13.73	58.5
	2R	870,914	4.29	313	84	26.84	15.42	89
	3L	938,947	4.02	288	100	34.72	12.33	66.5
	3R	866,971	3.11	288	102	35.76	10.33	24.5
	4	127,874	10.33	102	9	8.82	82.40	-
	Total	4,511,547		1,572	478			
	Average		3.86			30.53	13.46	57.70
LTR	X	628,924	2.89	134	54	41.04	6.15	19.00
	2L	603,536	2.72	127	67	52.76	5.72	20.00
	2R	573,034	2.82	140	54	38.57	6.90	30.50
	3L	618,441	2.65	117	58	49.57	5.01	24.00
	3R	621,272	2.23	154	67	44.16	5.52	7.50
	4	44,121	3.56	10	4	40.00	8.08	-
	Total	3,089,328		682	304			
	Average		2.65			44.87	5.84	20.20
LINE-like	X	136,348	0.63	71	18	25.35	3.26	14.50
	2L	185,499	0.83	98	20	20.41	4.41	18.50
	2R	225,984	1.11	109	18	16.51	5.37	37.50
	3L	251,077	1.08	106	27	25.47	4.54	24.00
	3R	176,355	0.63	70	19	27.14	2.51	4.50
	4	37,399	3.02	32	1	3.12	25.85	-
	Total	1,012,662		486	103			
	Average		0.87			21.19	4.16	19.80
TIR	X	45,324	0.21	59	7	11.86	2.71	14.50
	2L	82,761	0.37	76	11	14.47	3.42	18.00
	2R	69,291	0.34	62	11	17.74	3.05	20.50
	3L	52,743	0.23	57	12	21.05	2.44	16.50
	3R	63,359	0.23	60	14	23.33	2.15	12.50
	4	44,195	3.57	58	3	5.17	46.85	-
	Total	357,673		372	58			
	Average		0.31			15.59	3.19	16.40
FB	X	17,774	0.08	12	4	33.33	0.55	2.00
	2L	6,675	0.03	4	2	50.00	0.18	2.00
	2R	2,605	0.01	2	1	50.00	0.1	0.50
	3L	16,686	0.07	8	3	37.50	0.34	2.00
	3R	5,985	0.02	4	2	50.00	0.14	0.00
	4	2,159	0.17	2	1	50.00	1.62	-
	Total	51,884		32	13			
	Average		0.04			40.62	0.27	1.30

Tableau 3 : Vue d'ensemble des éléments transposables au sein de l'euchromatine du génome de *D. melanogaster*.

Pour chaque classe, le nombre total de chaque famille de l'élément, ainsi que le pourcentage d'éléments qui sont de pleine taille est donnée pour chaque bras de chromosome. La colonne 3 donne les paires de bases totaux constitués par des éléments transposables, la colonne 4 le pourcentage de chaque bras chromosomique composé de séquences d'éléments transposables, la colonne 8 le nombre d'éléments par Mb, et la colonne 9 le nombre d'éléments dans le plus proximal 2 Mb de chacun des cinq grands bras chromosomiques. (Kaminker et al., 2002)

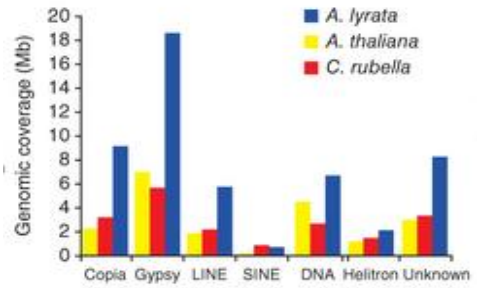


Figure 11 : Abondance des éléments transposable au sein du génome de *Arabidopsis thaliana*, *A. lyrata* et *C. rubella*. Couverture génomique dans les trois espèces des éléments transposables. LINE, SINE, Copia, Gypsy et autres transposons. (extrait ; Slotte et al., 2013)

2.6 Analyse des ETs au sein des espèces et/ou des génomes

La distribution des éléments transposables est un point important pour comprendre leur dynamique au sein de taxons d'espèces étudiées ou même au sein des eucaryotes. Elle traduit notamment leur capacité à s'amplifier et à se maintenir au sein des génomes. En effet, la distribution des éléments permet de connaitre l'importance de leur répartition au sein du vivant. Si les ETs sont trouvés dans tous les génomes des espèces eucaryotes étudiés jusqu'à aujourd'hui, les superfamilles d'éléments présentent des distributions différentes au sein des eucaryotes. L'étendue de la distribution des éléments dépend du niveau où l'on regarde. Elle sera différente si l'on regarde une superfamille, une famille ou un élément particulier. De plus, du point de vue des hôtes, l'échelle où l'on recherche la distribution des éléments est aussi importante. La distribution sera différente si l'on regarde au niveau des eucaryotes, au niveau d'un phylum ou d'un taxon inférieur comme un embranchement ou une classe, car la distribution des éléments traduits leur histoire évolutive.

Certaines superfamilles d'éléments présentent une distribution ubiquitaire au sein des eucaryotes, comme les rétrotransposons LINEs et SINEs, et les transposons *Tc1/Mariner*. D'autres ETs présentent une distribution parcellaire au sein des eucaryotes. Les éléments à Tyrosine Recombinase DIRS1 présentent une distribution inégale, taxon dépendante à l'échelle d'un ordre. Chez les décapodes, dans certains groupes, comme ceux des crevettes et des homards, on retrouve des éléments DIRS1 dans toutes les espèces étudiées. Dans d'autres groupes, les crabes et les galathées, on ne retrouve ces éléments que dans quelques espèces (Figure 12).

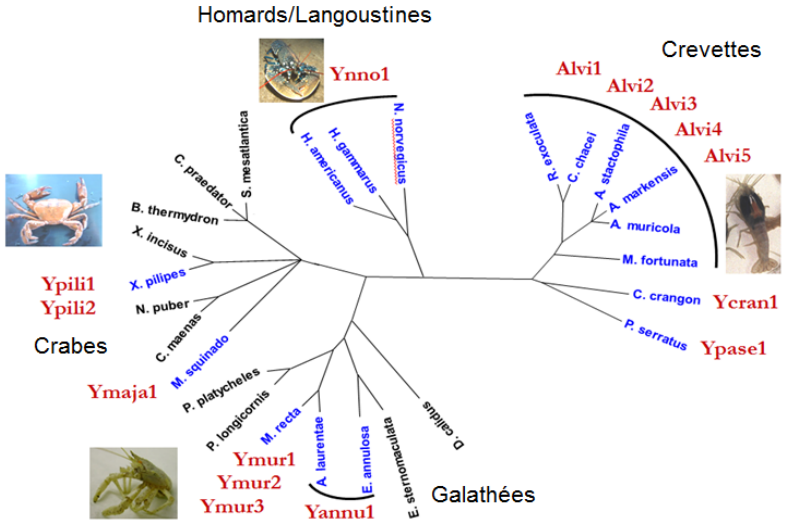


Figure 12 : Distribution des éléments DIRS1 dans 4 taxons de crustacés décapodes d'après une approche par PCR. Les espèces dont le nom est en bleu sont les espèces dans lesquelles la présence de DIRS1 a été établie, contrairement aux espèces dont le nom est en noir. Le nom des éléments retrouvés au sein des espèces est écrit en rouge. Les accolades noires définissent les espèces dans lesquelles on a retrouvé les éléments. (Piednoël et Bonnivard, 2009)

Le même type de distribution se retrouve à une échelle plus importante, celle des eucaryotes (Bui et al., 2007, 2008; Casse et al., 2006)(Piednoël et al., 2011). Une étude de 274 génomes complètement séquencés révèle que dans certains groupes, un grand nombre d'espèces présentent des DIRS1 (Figure 13). Ils sont relativement fréquents au sein des unikontes (notamment chez les métazoaires), groupe avec un grand nombre de génomes séquencés. On observe une réussite dans certains groupes comme celui des actinoptérygiens avec 5 espèces présentant des DIRS1 sur 5 espèces testées. Il existe à l'inverse des groupes dans lesquels on ne retrouve pas d'éléments DIRS1, comme par exemple chez les mammifères malgré 37 génomes complètement séquencés pour ce taxon. Enfin, on observe une distribution intermédiaire avec des espèces présentant des DIRS1 et d'autres non, comme par exemple, chez les hexapodes ou chez les amibes. Chez les bikontes (plantes, algues, excavates et chromoalveolates) les éléments DIRS1 sont beaucoup plus rares et leur présence reste à confirmer dans d'autres groupes (haptophytes).

Cette étude a permis de décrire les éléments DIRS1 dans de nombreuses espèces et de démontrer une grande présence de ces éléments au sein des eucaryotes. En effet, ces éléments étaient assez peu décrits jusqu'à présent au sein des espèces et donc considérés comme des éléments rares. Les DIRS1 ne sont effectivement pas présents chez de nombreuses espèces modèles comme *D. melanogaster*. L'étude de la répartition d'un type d'élément permet ainsi

d'appréhender sa dynamique au sein des eucaryotes et de formuler des hypothèses sur son origine. On retrouve des DIRS1 au sein de presque tous les groupes de métazoaires, donc l'hypothèse la plus parcimonieuse serait une origine ancienne associée à la perte des éléments dans des groupes tels que les mammifères. La perte des éléments est liée au nombre de copies. En effet, une espèce avec un faible nombre de copies d'un élément à plus de chance de perdre cet élément, qu'une espèce présentant un grand nombre de copies. Concernant la présence d'éléments DIRS1 dans certains groupes de champignons, une hypothèse serait la présence de DIRS1 au niveau de l'ancêtre des unikontes, puis la perte dans les taxons vides. Par contre, la distribution très morcelée des DIRS1 au sein des bikontes peut suggérer (en plus de l'hypothèse d'une origine très ancienne de ces ETs) une acquisition parallèle par transfert horizontal (acquisition par une espèce de matériel génétique, ici d'un ET, provenant d'une autre espèce, indépendamment de croisement).

On peut étudier la distribution des ETs à un niveau intermédiaire entre un embranchement (par exemple les crustacés) et les eucaryotes (très grande échelle). De la Chaux et Wagner (2011), ont étudié les éléments Bel/Pao au niveau des métazoaires, seul phylum dans lequel on retrouve ces éléments. L'étude au niveau d'un phylum permet par exemple la comparaison avec d'autres éléments. Ici, De la Chaux et Wagner ont étudié la distribution et le nombre de copies des éléments Bel/Pao par comparaison avec les autres rétrotransposons à LTR Copia et Gypsy. Ils ont mené cette étude dans 62 génomes complètement séquencés et n'ont pas retrouvé d'éléments Bel/Pao dans certains embranchements, comme par exemple les mammifères (Figure 14). Les Bel/Pao sont malgré tout bien présents au sein des métazoaires, avec des groupes où toutes les espèces (ou presque) en présentent, comme les nématodes, des groupes ou seulement certaines espèces en présentent, comme les arthropodes. Ce type d'analyse a permis de constater que contrairement à ce que l'on pensait, les Bel/Pao ne sont pas des éléments si rares au sein des métazoaires. Au contraire, ils apparaissent comme la seconde superfamille de rétrotransposons à LTR au sein des génomes étudiés, derrière les Gypsy et devant les Copia.

En conclusion, de la même manière que pour les DIRS1, ce genre d'étude permet de rendre compte de l'importance de certains types d'éléments que l'on pensait rare car encore peu étudié à grande échelle. En effet, un tel résultat n'est pas forcément visible à l'échelle d'un seul embranchement. Cependant, il faut tout de même se méfier des biais dans le choix des espèces séquencés, puisque sur les 62 génomes étudiés, 12 sont des génomes de drosophiles et 5 de

Caenorhabditis. A l'inverse, des groupes comme les mollusques, les porifères (éponges) ou les cnidaires sont fortement sous représentés. Toutefois, même si ces biais pondèrent un peu les résultats, les principales conclusions, à savoir l'importance des Bel/Pao au sein des métazoaires, reste solide. Une seconde information est que les mammifères ont l'air d'être hermétiques à certains ET comme les Bel/Pao et les DIRS1, c'est à dire que dans aucunes des espèces ayant le génome complètement séquencé nous ne retrouvons ces éléments.

Comme nous avons pu le voir, la dynamique des ET, et donc leur distribution, dépend du nombre de copies de ces éléments au sein des génomes. L'estimation du nombre de copies d'un élément ou d'une famille d'éléments particuliers au sein des génomes peut donc permettre de comprendre la distribution de l'ET. En effet, si dans une espèce il y a peu de copies d'un élément, la perte de l'élément au sein du taxon de l'espèce parait plus facile, tandis que si au sein d'une espèce il y a un grand nombre de copies d'un élément, le maintien de cet élément est rendu plus compréhensible. Une bonne estimation implique de disposer de génomes séquencés avec une bonne couverture et un assemblage correct. L'estimation du nombre de copies au sein d'un génome peut correspondre à des informations différentes suivant les études. On peut estimer les éléments « complets » (« pleine taille ») c'est à dire des éléments dont la structure et la séquence sont encore bien conservées, même s'ils ne sont pas forcément toujours actifs. Ces éléments sont des témoins d'une insertion relativement récente au sein du génome. On peut s'intéresser à des éléments qui présentent un nombre N de domaines encore bien conservés mais qui auraient par exemple des terminaisons remaniées.

On peut retrouver ces éléments grâce à des logiciels comme ReDoSt, qui recherche des domaines conservés au sein des éléments mais ne tient pas compte des terminaisons. Certains auteurs considèrent, en plus des copies « pleine taille », les éléments délétés. Ils vont alors rechercher au sein des génomes des domaines particuliers, par exemple la RT/RH ou l'intégrase. On peut aussi rechercher une structure particulière comme une LTR, qu'elle soit encore liée à un élément ou sous forme de solo-LTR. En effet, les LTR portent les séquences régulatrices et promotrices qui peuvent influencer la régulation de certains gènes physiquement proches. Si l'on s'intéresse à la part que peuvent représenter les ETs au sein d'un génome, alors on peut aussi considérer les fragments d'éléments ou encore des traces que l'on peut rechercher grâce à RepeatMasker par exemple. La caractérisation des différentes copies présentent dans un génome permet par la suite d'estimer la diversité de celles-ci. De la Chaux et Wagner (2011) ont recherché le nombre de copies de Bel/Pao, avec des LTR et au moins un domaine

fonctionnel. Le but de cette étude était de décrire tous les éléments et d'analyser leur relation phylogénétique. Les éléments trouvés ont été regroupés en plus de 1 725 familles dont 1 623 nouvelles, qui se répartissent en sept clades (Tableau 4).

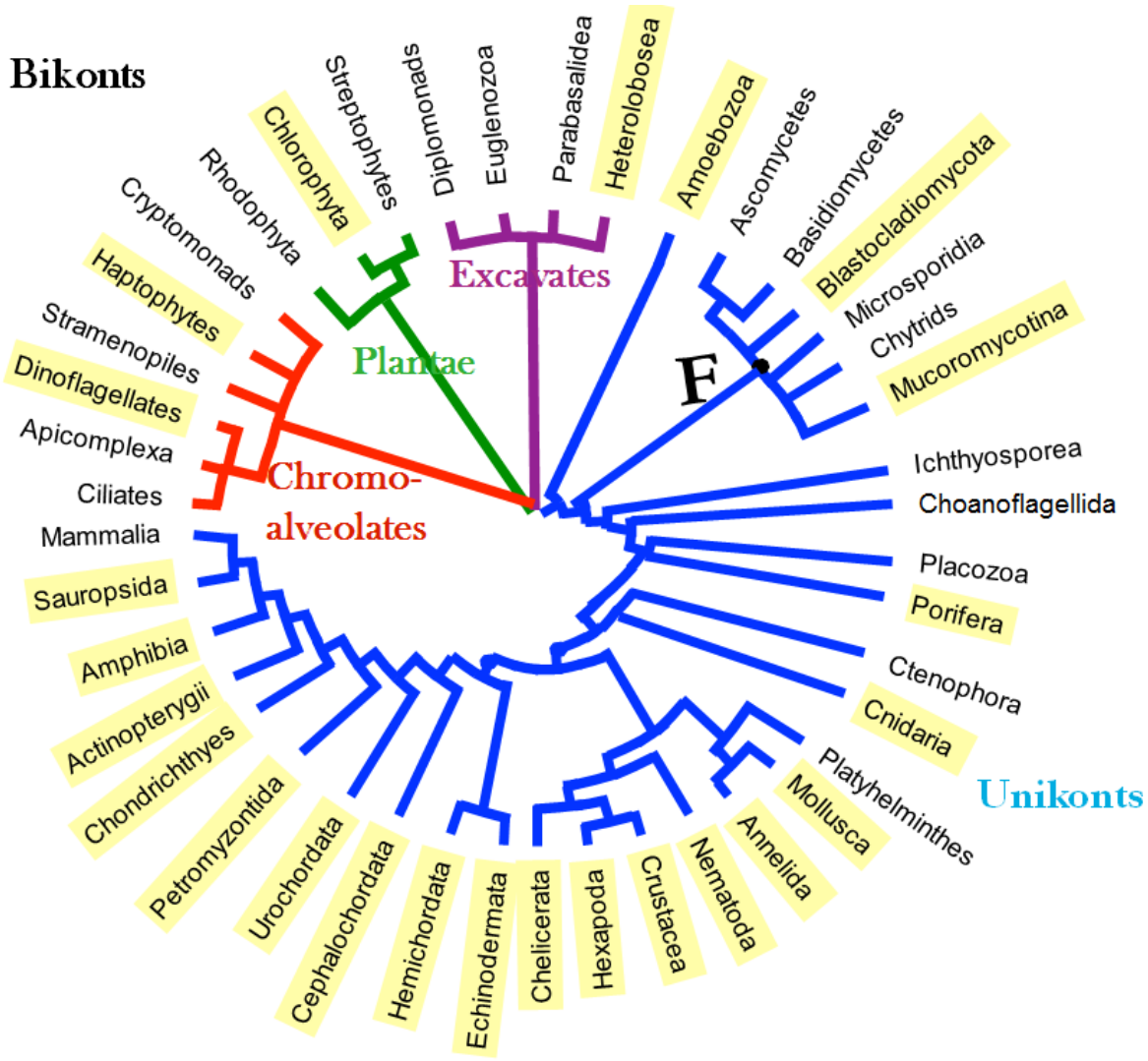


Figure 13: Distribution des éléments DIRS1 (en 2014) au sein des eucaryotes d'après une analyse bio-informatique de 276 génomes complètement séquencés, une analyse par approche par PCR sur les décapodes (crustacés) et la bibliographie sur les DIRS1 (Piednoël et al., 2011). Et complétée par une analyse par PCR sur des espèces de chondrichtyens et par la bibliographie des génomes complètements séquencés depuis 2009. Les taxons surlignés en jaune sont les taxons avec présence des DIRS1.

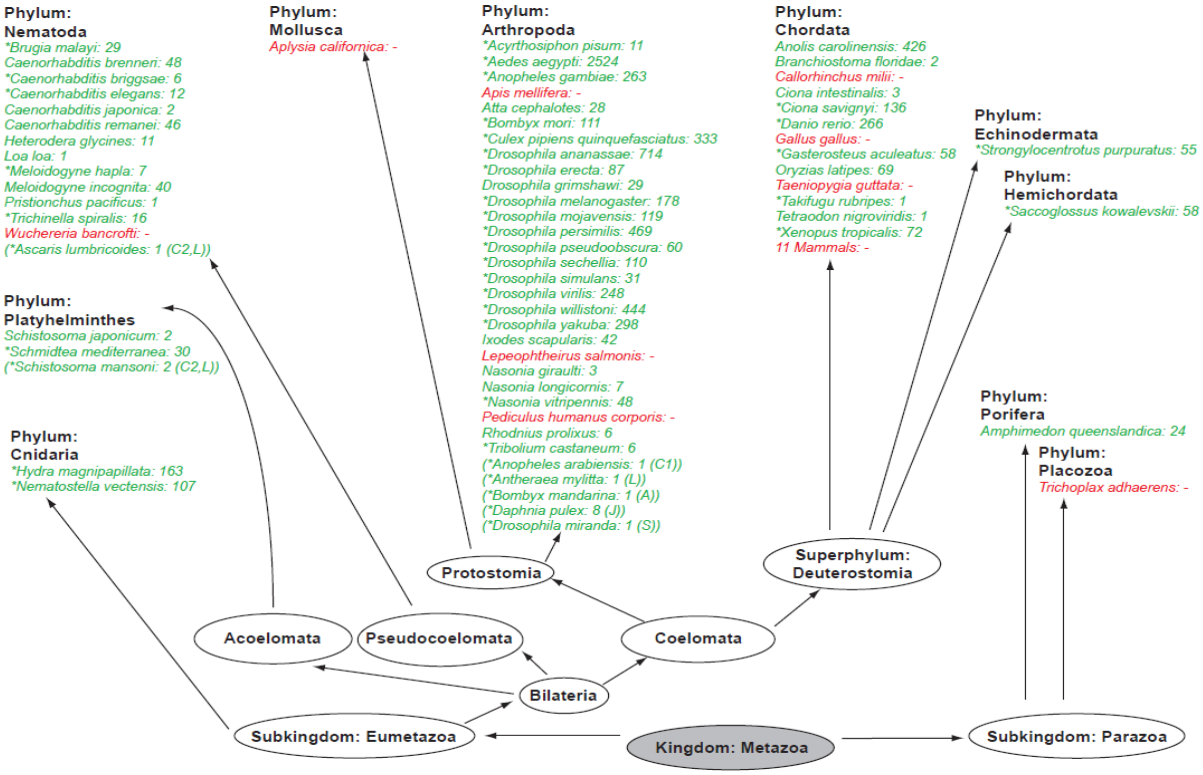


Figure 14 : Vue d'ensemble des séquences des génomes analysées et leur classification taxonomique. Les noms des 62 espèces non mammifères dont les génomes ont été analysés sont regroupés par taxon. 11 génomes de mammifères supplémentaires sont résumés comme "11" Mammifères. Pour chaque espèce, le nombre d'éléments Bel/Pao identifiés est affiché. Le nom du génome est représenté en rouge si les auteurs n'étaient pas en mesure d'identifier d'éléments dans le génome. Les génomes dans lesquels des éléments Bel/Pao avaient déjà été identifiés précédemment sont marqués par un astérisque (*). Sept autres espèces où aucune séquence complète du génome était disponible, mais où les éléments Bel/Pao avaient été identifiés précédemment sont indiquées entre parenthèses. (De la chaux et Wagner, 2011)

Superfamily	Elements	Families	Species	Phyla
Pao	1723	323	14	1
Sinbad	791	85	9	5
Dan	85	33	1	1
Flow	83	5	3	2
Tas	270	104	17	5
BEL	2955	333	21	2
Suzu	30	10	5	3

Tableau 4 : Nombre d'éléments, de familles, d'espèces et de phyla pour chaque superfamille.

(De la Chaux et Wagner, 2011)

Piednoël et al., (2011) ont relevé le nombre de copies presque entières des éléments DIRS1, c'est à dire avec les domaines conservés et en ordre. Cette étude a été réalisée afin de connaitre le nombre de copies au sein des génomes et le nombre de familles différents de DIRS1 (Tableau 5). Le but était de caractériser des éléments DIRS1 récemment actifs ou potentiellement encore actifs et d'analyser leur relation phylogénétique. Ils ont identifié plus de 4000 copies de DIRS1 répartis dans 30 espèces qui peuvent être regroupés dans environ 300 clusters/familles. Ils ont également recherché le nombre de familles par espèce ainsi que le nombre de copies minimales et maximales dans les familles. Alors que la diversité dans la plupart des espèces semble limitée à un faible nombre de copies comme chez *Oryzias latipes* (6 copies) ou *Emiliana huxleyi* (1 copie), quelques «bursts» de transposition ont probablement eu lieu dans certaines espèces, pour lesquelles un grand nombre de copies ne se regroupent que dans quelques gros clusters; un cluster à 1157 copies chez le poisson zèbre (*Danio rerio*) et à 319 copies chez le lézard (*Anolis carolinensis*).
Les relations phylogénétiques entre les différentes copies nouvellement décrites lors d'une étude, permettent de définir des familles d'éléments grâce à leurs regroupements dans des clades monophylétiques. Le regroupement dans un même clade témoigne d'une origine commune. Grâce à la phylogénie des éléments, on peut également suspecter les transferts horizontaux.

Concernant les éléments DIRS1, l'étude phylogénétique a permis de voir que la plupart des clades sont espèce ou taxon dépendants (Figure 15). Les éléments d'actinoptérygiens se groupent dans un seul et même clade : «fish group». La plupart des éléments DIRS1 de champignons se groupent dans un clade majoritaire «fungi1». Cependant, certains éléments d'une même espèce peuvent appartenir à plusieurs clades. Par exemple, les éléments de *Lottia gigantea* se regroupent dans 3 clades différents. Ceci peut s'expliquer par plusieurs évènements d'envahissement du génome par les éléments DIRS1. Cette phylogénie confirme également que les éléments PAT sont le groupe frère des éléments DIRS1.

Higher taxon	Species	Copy number	Family number	Min	Max
Actinopterygii	*Danio rerio* *	2091	14	1	1157
	Gasterosteus aculeatus	21	4	1	12
	Oryzias latipes	6	1	-	-
	Takifugu rubripes *	7	1	-	-
	Tetraodon nigroviridis *	8	2	1	7
Amoebozoa	*Dictyostelium discoideum* *	16	1	-	-
	Acanthamoeba sp.	1	1	-	-
Amphibia	*Xenopus tropicalis* *	692	81	1	38
Annelida	*Capitella sp. I*	5	2	1	4
Blastocladiomycota	*Allomyces macrogynus*	21	6	1	10
Cephalochordata	*Branchiostoma floridae*	15	11	1	3
Chlorophyta	*Chlamydomonas reinhardtii*	11	5 (3)	1	4
	Volvox carteri	36	6 (4)	2	13
Cnidaria	*Nematostella vectensis* *	60	21 (1)	1	7
Crustacea	*Daphnia pulex* *	100	39	1	5
Echinodermata	*Strongylocentrotus purpuratus* *	4	4	-	-
Haptophytes	*Emiliana huxleyi*	1	1	-	-
Hemichordata	*Saccoglossus kowalevskii* *	240	8 (1)	1	175
Heterolobosea	*Naegleria gruberi*	7	6	1	2
Hexapoda	*Bombyx mori*	6	2	3	3
	Nasonia vitripennis *	37	18	1	4
	Tribolium castaneum *	1	1	-	-
Mucoromycotina	*Mucor circinelloides*	3	2	1	2
	Phycomyces blakesleeanus *	28	13	1	5
	Rhizopus oryzae *	24	11	1	4
Mollusca	*Aplysia californica*	39	7	2	10
	Lottia gigantea	44	22 (1)	1	5
Nematoda	*Caenorhabditis briggsae* [$]	1	1 (1)	-	-
	Pristionchus pacificus [$]	4	3 (3)	1	2
Petromyzontida	*Petromyzon marinus*	2	2	-	-
Sauropsida	*Anolis carolinensis*	775	42	1	319
Urochordata	*Oikopleura dioica* *	4	2	1	3

Tableau 5 : Résultats de la détection du nombre de copies et du clustering des DIRS1-like rétrotransposons au sein des eucaryotes. (Piednoël et al., 2011)

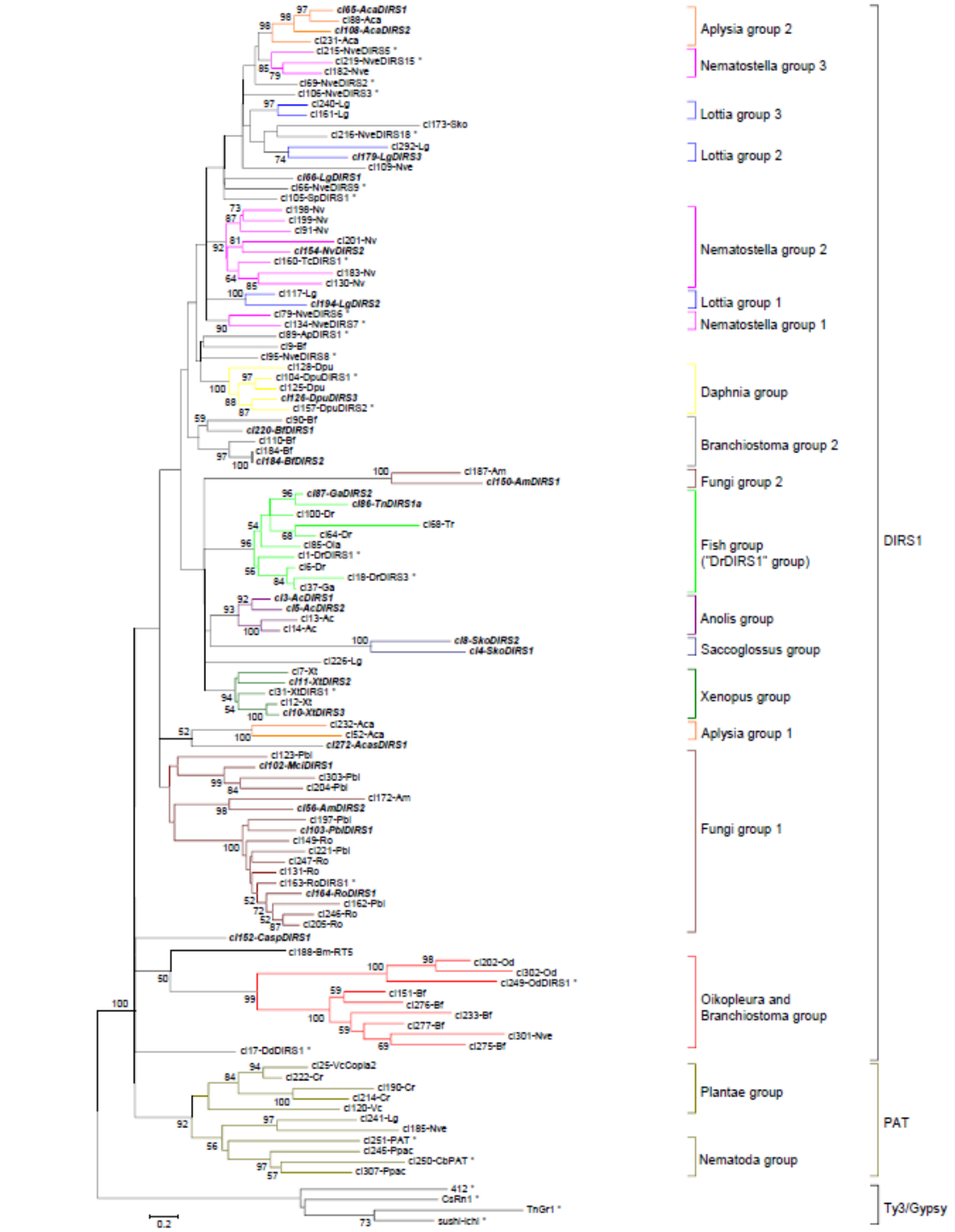


Figure 15 : Arbre phylogénétique enraciné basé sur les séquences d'acides aminés de la *pol* de DIRS1 analogues identifiés. Seules les valeurs de nœud bootstrap plus de 50% sont représentés.
(Piednoël et al., 2011)

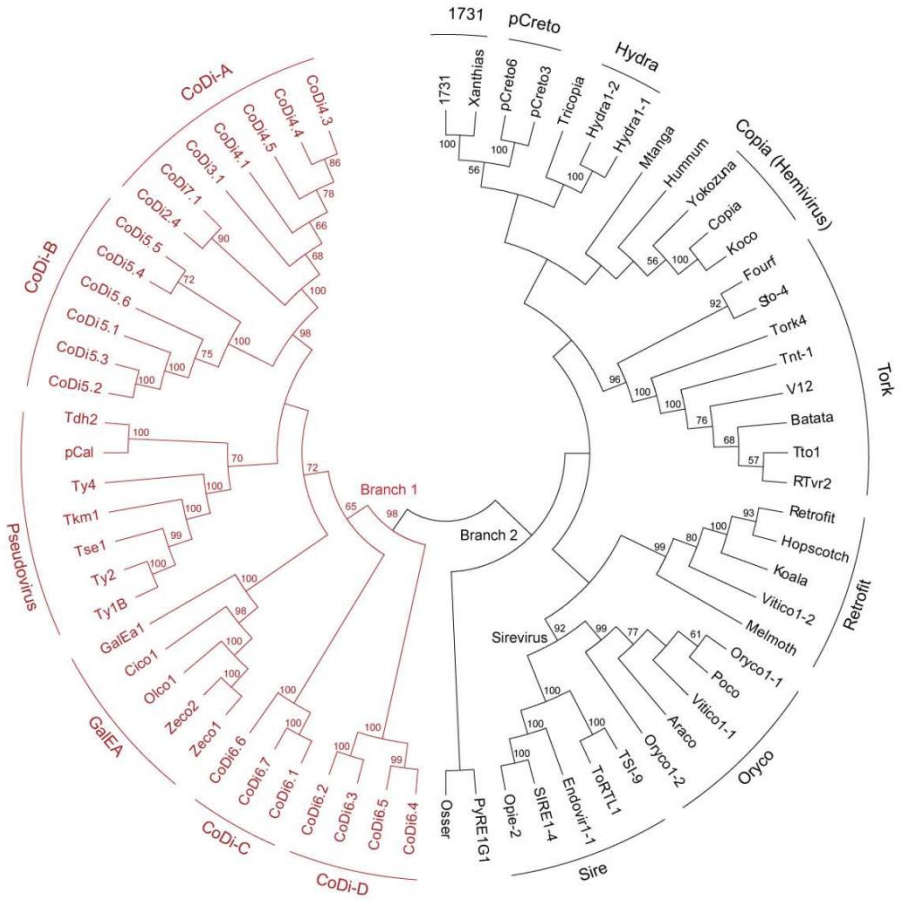


Figure 16 : Phylogénie des éléments Copia basée sur la *pol*.
Cet arbre révèle deux branches principales, la branche 1 et la branche 2 (en rouge et noir, respectivement) (Llorens et al., 2009).

La phylogénie peut également permettre au sein d'une superfamille de définir de grands clades d'éléments et de voir les relations entre les différents clades d'éléments. C'est ce qu'ont fait Llorens et al., (2009) (Figure 16) avec les éléments de la superfamille Copia par exemple. Grâce à la phylogénie, ils ont défini deux branches d'éléments. La première branche d'éléments est composée des éléments CoDi décrits au sein des diatomées, des éléments Ty qui font partis de la famille des Pseudovirus et des éléments de la famille des GalEa. Dans la seconde branche on retrouve toutes les autres familles de Copia tel que le Hydra, les Copia et les Tork par exemple.

L'analyse de la distribution des éléments au sein d'un taxon, l'estimation du nombre de copies au sein des génomes, ainsi que l'étude des relations phylogénétiques des éléments permettent d'étudier leur dynamique. Durant ma thèse, j'ai étudié la distribution des éléments du clade GalEa, leur nombre de copies, ainsi que les relations phylogénétiques des ETs nouvellement décrits pour comprendre l'origine de ces ET.

TRAVAUX DE THESE :

RESULTATS ET DISCUSSION

RESULTATS DE THESE

Chapitre III. Distribution des rétrotransposons à LTR au sein des crustacés

Introduction

Pour étudier la dynamique à large échelle d'un clade d'élément, nous avons choisi en premier lieu un taxon d'espèces, en dehors des espèces modèles habituelles, avec différentes caractéristiques tel qu'un grand nombre d'espèces, qui nous permet d'avoir par exemple un grand nombre de données de séquençage. Nous recherchons également une grande diversité de style de vie et d'habitats (côtiers ou benthiques, marins, d'eau douce, terrestre, etc) car nous savons que les ETs réagissent différemment aux stress des environnements. Et enfin nous recherchons des espèces avec une grande variation de taille de génome car la présence, au sein d'un taxon, d'espèces avec de fortes variations de taille de génomes peut s'expliquer par une grande part des ETs au sein des génomes de grande taille.

Lors du choix de nos modèles biologiques, nous nous sommes intéressés à des espèces liées aux sites hydrothermaux profonds car ce sont des milieux hypervariables, ce qui peut jouer sur la réponse aux stress des éléments avec une forte amplification au sein de ces génomes. Nous souhaitions déterminer s'il existe aussi des espèces phylogénétiquement proches venant d'autres environnements comme par exemple les milieux côtiers. Nous aurions pu, d'après ces critères, nous intéresser aux bivalves et notamment aux moules car le genre *Bathymodiolus* est représenté par plusieurs espèces en milieu hydrothermal. De plus, on retrouve des moules dans différents milieux: côtiers ou eau douce par exemple. En fait, notre choix s'est porté sur les crevettes, plus particulièrement sur la crevette hydrothermale *Rimicaris exoculata* (photo) et plus largement les crustacés, qui possèdent également les caractéristiques recherchées. De plus, l'étude des ETs au sein des crustacés nous permettrait de faire un parallèle avec un autre groupe d'arthropodes bien étudié, celui des hexapodes. C'est pour cela que nous nous sommes intéressés au sous embranchement des crustacés pour notre étude, bien que l'on ne possède que

très peu de données de séquençage de génomes de ces espèces.

Rimicaris exoculata

Le groupe des crustacés est composé d'environ 33 000 espèces. Ce groupe présente une variation de tailles de génomes importante. En effet, une crevette *Chorocaris chacei* a un génome de 15 pg, alors que le crabe *Cyanagraea praedator* présente un génome de 3 pg. Les crustacés sont assez peu étudiés en biologie et notamment en ce qui concerne les ETs. Cependant, il existe quelques rares études réalisées sur ce taxon. Certaines de ces études portent sur les transposons *Mariner* au sein de crabes côtiers (Bui et al., 2007, 2008; Casse et al., 2006). L'espèce *Peneus Monodon* est également assez étudiée, car elle présente un intérêt économique (de la Vega et al., 2007), avec le début du séquençage de son génome (Huang et al., 2011). Cette espèce a également été étudiée pour le transposon Argonaute 4 (Leebonoi et al., 2015). Une seule espèce de crustacés a son génome entièrement séquencé, *Daphnia pulex* en 2010 (*Daphnia* Genomics Consortium (DGC) http://daphnia.cgb.indiana.edu). Deux études concernant les ETs au sein du génome de cette espèce ont été réalisées : l'une porte sur la dynamique des transposons de cette espèce et sur leur rôle dans la recombinaison lors de l'accumulation de mutations au sein du génome (Schaack et al., 2010); l'autre décrit les rétrotransposons à LTR au sein du génome (Rho et al., 2010). A côté de ces différentes études monospécifiques, il n'existe aucune étude réalisée sur les rétrotransposons à LTR à large échelle.

Au sein de notre équipe, le choix s'est porté sur l'étude des ETs chez les crustacés et plus particulièrement des rétrotransposons. L'étude des éléments DIRS au sein des Décapodes (Piednoël et Bonnivard, 2009) m'amena au cours de mon stage de Master 2 à étudier plus largement ces éléments dans le sous-embranchement des crustacés. Une première publication concernant la description des premiers éléments Copia chez des galathées (Terrat et al., 2008) a permis de découvrir une famille particulière: les GalEa. Des éléments Copia ont été découverts chez une espèce modèle de crevette, *Rimicaris exoculata*, appelé CoRex1, 2 et 3. Après une étude plus poussée de ces éléments, nous nous sommes aperçus qu'ils faisaient également partie de la famille des GalEa. Nous avons décidé d'étudier la diversité des éléments Copia au sein des crustacés grâce à l'étude de 25 espèces représentant les différents ordres des crustacés. De plus, nous avons souhaité étudier les éléments Gypsy de la même manière pour compléter cette analyse et pouvoir faire un comparatif avec l'étude réalisée au sein du génome de la daphnie par exemple.

Article

OPEN ACCESS Freely available online

PLOS ONE

LTR-Retrotransposons in *R. exoculata* and Other Crustaceans: The Outstanding Success of GalEa-Like Copia Elements

Mathieu Piednoël[1,2☯], Tifenn Donnart[1☯], Caroline Esnault[1¤], Paula Graça[1], Dominique Higuet[1], Eric Bonnivard[1*]

1 UMR 7138 Systématique Adaptation Evolution, Equipe Génétique et Evolution, Université Pierre et Marie Curie, Paris, France, **2** Systematic Botany and Mycology, University of Munich (LMU), Munich, Germany

Abstract

Transposable elements are major constituents of eukaryote genomes and have a great impact on genome structure and stability. They can contribute to the genetic diversity and evolution of organisms. Knowledge of their distribution among several genomes is an essential condition to study their dynamics and to better understand their role in species evolution. LTR-retrotransposons have been reported in many diverse eukaryote species, describing a ubiquitous distribution. Given their abundance, diversity and their extended ranges in C-values, environment and life styles, crustaceans are a great taxon to investigate the genomic component of adaptation and its possible relationships with TEs. However, crustaceans have been greatly underrepresented in transposable element studies. Using both degenerate PCR and *in silico* approaches, we have identified 35 Copia and 46 Gypsy families in 15 and 18 crustacean species, respectively. In particular, we characterized several full-length elements from the shrimp *Rimicaris exoculata* that is listed as a model organism from hydrothermal vents. Phylogenic analyses show that Copia and Gypsy retrotransposons likely present two opposite dynamics within crustaceans. The Gypsy elements appear relatively frequent and diverse whereas Copia are much more homogeneous, as 29 of them belong to the single GalEa clade, and species- or lineage-dependent. Our results also support the hypothesis of the Copia retrotransposon scarcity in metazoans compared to Gypsy elements. In such a context, the GalEa-like elements present an outstanding wide distribution among eukaryotes, from fishes to red algae, and can be even highly predominant within a large taxon, such as Malacostraca. Their distribution among crustaceans suggests a dynamics that follows a "domino days spreading" branching process in which successive amplifications may interact positively.

Citation: Piednoël M, Donnart T, Esnault C, Graça P, Higuet D, et al. (2013) LTR-Retrotransposons in *R. exoculata* and Other Crustaceans: The Outstanding Success of GalEa-Like Copia Elements. PLoS ONE 8(3): e57675. doi:10.1371/journal.pone.0057675

Editor: Khalil Kashkush, Ben-Gurion University, Israel

Received December 6, 2012; **Accepted** January 24, 2013; **Published** March 4, 2013

Funding: The authors have no support or funding to report.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: eric.bonnivard@upmc.fr

¤ Current address: Section on Eukaryotic Transposable Elements, Laboratory of Gene Regulation and Development, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, Maryland, United States of America

☯ These authors contributed equally to this work.

Introduction

Transposable elements (TEs) have a large impact on genome structure and stability, and are therefore considered as one of the major sources of genetic variability in eukaryotes [1–4]. Environmental variations can promote genome plasticity through transcriptional activation and TE mobilization, often in response to specific stimuli such as biotic stress (e.g., pathogens) and abiotic environmental changes [5–9]. Retrotransposons, a TE class specific to eukaryotes, transpose *via* a RNA intermediate. Five orders of retrotransposons can be defined based on their structural features and their phylogenetic relationships [10]: Long Terminal Repeat retrotransposons (LTR-retrotransposons), tyrosine recombinase encoding retrotransposons (e.g. DIRS1-like elements), Penelope elements, LINEs (Long INterspersed Elements) and SINEs (Short INterspersed Elements). Copia (or Ty1/Copia), Gypsy (or Ty3/Gypsy) and BEL/Pao elements constitute the three superfamilies of LTR-retrotransposons. These elements are related to retroviruses [11] and usually encode two Open Reading Frames (ORFs). The first ORF, the *gag* region, encodes proteins that form the virus-like particles. The second ORF, the *pol* region, is a polyprotein comprising the different domains involved in the retrotransposition mechanism. These domains include an aspartic protease (PR), a reverse transcriptase (RT), a RNase H (RH) and a DDE-type integrase (INT), whose order varies among LTR-retrotransposon superfamilies [12].

Transposable elements have been found in all eukaryotic species investigated so far [10]. However, the TE superfamilies show variable distributions among eukaryotes. For example, LINEs, SINEs retrotransposons and the Tc1/Mariner transposons, have been detected almost ubiquitously [10,13,14]. The Penelope retrotransposons are widely distributed among animal species, but seem to be rare among plants, protists and fungi [15]. The DIRS1-like elements are less frequent but their distribution appears broader than it was previously thought, especially in unikont species, although they remain undetectable in mammals

[16]. Now reported in 61 species, they are widely distributed in some particular phyla such as Decapoda [17]. Finally, LTR-retrotransposons are found in a wide continuous range of species [18,19,20], but a recent analysis of 62 sequenced metazoan genomes underlined their uneven relative abundances among these species [21]. Gypsy elements are the most abundant, the BEL/Pao elements appear intermediate and the Copia retrotransposons constitute a distant third group of low-copy elements.

The decapods (shrimps, lobsters, crabs, etc), and more globally the crustaceans, are a great model to investigate the genomic component of adaptation and its possible relationship with TEs. First, crustaceans form a very large group of arthropods that exhibit great diversity in terms of species, lifestyles (including some parasitic organisms such as *Sacculina carcini*) and are found in various environments (*e.g.* from fresh to highly salty water or from deep-sea vents to terrestrial species). Second, they exhibit great variations in genome size; decapods range from 1.05 Gb in the crab *Carcinus mamas* to 40 Gb in the shrimp *Solenocrangon ferox* [22], with several species (*e.g.* shrimps) that show particularly large genomes and are thus likely to harbor high TE contents [23]. Most of the previous studies on TEs focused on model organisms, such as studies on horizontal transfer across Mammals [24], LINEs and SINEs in human genome [25] or dynamics and impact of TE invasion on the Drosophila genomes [26]. This species sampling bias could potentially affect our knowledge in TE dynamics and evolution. This is particularly striking for marine species such as crustaceans. Given their abundance and diversity, Crustacea and Decapoda have been greatly underrepresented in studies on retrotransposons where few elements have been described to date. LINEs are the most reported retrotransposons in crustaceans with several elements described in the isopod *Porcellio scaber* [27], the ostracod *Darwinula stevensoni* [28], the branchiopod *Daphnia pulex* [29] and several decapods, principally the prawns *Litopenaeus stylirostris*, *Litopenaeus vannamei* and *Penaeus monodon* [30,31]. DIRS1-like elements also constitute a well-studied retrotransposon group within crustaceans. They appear widely distributed among decapods with elements described in 15 diverse species [17]. Interestingly, the study of these elements revealed that they constitute a new DIRS1-like clade, called AllDIRS1 and distant from the elements identified in the *D. pulex* genome [17,32]. This suggests that different TE dynamics occurred among the crustacean orders. By contrast, only a little is known about Penelope elements and LTR-retrotransposon distributions in crustaceans. Penelope elements have been reported only in the prawns *P. monodon* [33] and *Marsupenaeus japonicus* [34]. LTR-retrotransposons are limited essentially to those described in the sequenced genome of *D. pulex* [32] and in galatheid squat lobsters [35,36]. Copia elements, discovered in galatheids using degenerate PCR, define the new GalEa clade, which is widely dispersed among animal species. Indeed, the GalEa-like elements have also been described in phylogenetically distant species, the teleosts *Danio rerio* (Zeco1) and *Oryzias latipes* (Olco1), and the urochordate *Ciona intestinalis* (Cico1) [35].

In this study, we particularly focus on *Rimicaris exoculata*. This deep-sea vent organism may present particular TE characteristics due to its peculiar adaptive abilities and its relatively large genome (10.16 Gb; [17,23]). Deep-sea vents are chemosynthetic environments particularly unstable, where intense physico-chemical shifts are occurring over very short spatial and temporal scales [37–39]. Such unstable environment may be difficult to live in, therefore hydrothermal ecosystems are often considered harsh and stressful. They show however a much higher density of individuals compared with surrounding abyssal plains. *R. exoculata* represents an emblematic species of the Mid-Atlantic Ridge, where populations can reach up to 2500 individuals per square meter [40], and is exceptional among crustaceans for its association with bacteria [41]. It usually lives between 15°C and 30°C, but can endure sudden changes of thermal conditions due to fluid convections and survive the exposure to very high temperature vent emissions [42,43].

While studying DIRS1-like retrotransposons in decapods, we recently characterized RexAlvi1 and RexAlvi2, two elements from *R. exoculata* [17]. Herein we characterized Copia and Gypsy retrotransposons in this species using PCR strategies, and we determined the diversity of these elements among crustaceans using both PCR and *in silico* approaches. We studied 26 species that allow us a broad coverage of the crustacean diversity. We focused in particular on 20 decapods (including 7 other hydrothermal species) that represent the major Decapoda infraorders.

Materials and Methods

Biological Materials

One specimen of *R. exoculata* and one specimen of each shrimps *Alvinocaris markensis*, *Mirocaris fortunata* and *Chorocaris chacei* come from the Mid-Atlantic Ridge vent fields Rainbow and were sampled with the suction sampler of the ROV (Remotely Operated Vehicle) 'Victor 6000' operating from the R/V "Pourquoi pas?" (cruise MoMARETO [44], August 2006, IFREMER). The second specimen of *R. exoculata* was sampled on the same field using the French "Nautile" deep-submergence vehicle operating from the R/V "Pourquoi pas?" (cruise MoMARDREAM-naut [45], July 2007, IFREMER). One specimen of each other hydrothermal decapods were collected using the French "Nautile" deep-submergence vehicle operating from the N.O. "L'Atalante": shrimps *Alvinocaris fusca* and *Nematocarcinus barukovskyi* on the North East Pacific Rise (cruise MESCAL, June 2010, IFREMER); crab *Bythograea thermydron* and galatheid squat lobsters *Munidopsis recta* on the South East Pacific Rise (cruise BIOSPEEDO [46], March-May 2004, IFREMER). The coastal decapods (the caridean shrimps *Palaemon serratus*, *Crangon crangon* and the brachyuran crabs *Maja squinado*, *Necora puber*) and the parasitic barnacle *S. carcini* were collected in French Brittany (Roscoff, 2009). Two specimens of galatheid squat lobsters from seamounts (*Agononida laurentae* and *Eumunida annulosa*) were collected south of New-Caledonia on Norfolk seamounts during the prospecting campaigns Norfolk1 (2001, IRD Nouméa) and Norfolk 2 (2003, MUSORSTOM). The crayfish (*Orconectes limosus*) was collected near Paris (Val d'Oise) and the farmed prawns originated from Thailand (*L. vannamei*, *P. monodon*) were purchased frozen in a grocery store. Hydrothermal specimens were collected during official oceanographic research cruises; other organisms are not endangered or protected species and were not collected in privately-owned or protected areas; so, no specific permits were required for the described field studies.

For all samples, living specimens were fixed immediately after collection in liquid nitrogen for vent species, or in 70% ethanol for the other species. They were then stored at −80°C or 4°C, respectively. DNA from one individual per species was isolated from abdominal muscle tissue using the CTAB method. Dry DNA pellets were resuspended in water.

Detection of LTR-retrotransposons Using Degenerate Primers

To isolate LTR-retrotransposon *pol* fragments, we performed PCRs using several degenerate primer pairs designed within the conserved motifs of the RT/RH domains. Three primers (GD1,

GD2 and GD3) were designed to amplify motifs of Gypsy retrotransposons: 'RMPFGL' (5'-MGNMTGCCNTTYGGNYT-3'), 'LTTDAS' (5'-WSNGCRTCNGTNGSNA-3') and 'ADALSR' (5'-CKNGANASNSCRTCNGC-3'). For Copia retrotransposons, we used the primer pair (CD1/CD2) that previously allowed the detection of elements in the galatheid squat lobsters [35]. CD1 corresponds to the 'KARLVA' motif (5'-ARRGCNMGNYTNGTNGC-3', [35]) and CD2 to the 'YVDD' motif (5'-ANNANRTCRTCNACRTA-3', [47]).

PCR amplifications were performed for 35 cycles (94°C for 45 s, 50.2°C for 1 min and 72°C for 1 min) using 50 ng of DNA, 2.5 U of Taq DNA polymerase (Promega) and 50 pmol of each degenerate primer in a final volume of 25 µL. PCR amplification products were separated on 1% agarose gels. Bands with the expected molecular weight were excised, purified with the Nucleospin Extract kit (Macherey_Nagel) and cloned in pGEM-T vector according to the manufacturer recommendations (Promega, Madison, WI, USA). One to three clones were sequenced (http://www.beckmangenomics.com) and the nucleotide sequences were submitted to the GenBank database (see Table S1 for accession numbers).

Characterization of the Retrotransposons in *R. exoculata*

Sequences obtained with degenerate primers allowed the identification of several new LTR-retrotransposon families in *R. exoculata*. As described in Piednoël and Bonnivard [17], a group of sequences is considered as a family if its highest intra-group divergence is lower than its inter-groups divergence, without overlap of the two distributions. Two PCR walking approaches, 'PCR walking' [48] and 'TE Walking' [17], were then performed to extend large sequences from one representative initial fragment (see Table S1 for sequence reconstruction and primers). PCR amplifications were performed as presented above and for each walking step one to three clones were sequenced. Each new sequence was manually validated as an extension of the initial fragment using a minimum overlap of 50 bp between the two sequences, and a minimum DNA identity of 95%. Chimeric consensus elements were finally determined by joining the different PCR fragments using the Cap contig assembly program included in BioEdit [49].

We developed an efficient strategy that allows characterizing all parts of a full-length LTR-retrotransposon with the fewest possible PCR steps (Figure S1). (1) Detection of fragment of the RT domain using degenerate primers that can be used as an anchor sequence for PCR walking. This anchor sequence is compared with closely related retrotransposons to extrapolate the putative Primer Binding Site (PBS) sequence of the element. (2) Then the 5' edge of the element is obtained using a peculiar 'TE walking' step, we call 'PBS walking', which associates two specific primers designed within the anchor fragment and on the PBS sequence, respectively. When necessary, an additional 'PCR walking' step may be done to extend the 5' edge of the anchor fragment prior to the 'PBS walking'. (3) The 5' LTR sequence is determined by 'PCR walking'. (4) Assuming that both LTRs are almost identical, the missing 3' part of the element is amplified using a pair of specific primers designed in the presumed 3'LTR and in the anchor fragment, respectively.

Transcriptomic Survey

To identify transcriptionally active copies of the elements in *R. exoculata*, total RNAs were isolated from about 20 mg of abdominal muscle tissue (RNeasy mini kit, Quiagen). Prior to cDNA synthesis (Omniscript RT kit with poly(T) primer, Qiagen), RNA isolation products were treated with DNase I (10 U at 37°C during 1h30, inactivation 10 mn at 65°C). To test for DNA contamination within the RNA sample, we performed PCR amplifications using primers specific to the RT domain of each newly described element (primer sequences available upon request, see Table S1 for details). It results in an absence of PCR-amplified fragments, which attests the efficiency of the DNase treatment and the absence of the DNA contamination in the RNA sample. PCR amplifications were performed for 30 cycles (94°C for 45 sec, 54°C for 1 min, and 72°C for 1 min, followed by a final extension step at 72°C for 10 min) using about 50 ng of cDNA, 2.5 U of Taq DNA polymerase (Promega) and 10 pmol of each primer in a final volume of 25 µl.

Data Mining

To identify Copia and Gypsy elements in various crustacean species, we screened several genomic or transcriptomic databases. Gypsy and Copia sequences from the sequenced genome of *D. pulex* were obtained either from National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov) or RepBase (http://www.girinst.org/server/RepBase/index.php). Transcriptomic sequences from Antarctic krill *Euphausia superba* [50] and those of *Euphausia crystallorophias* were kindly provided by JY Toullec (Station biologique de Roscoff); those from the amphipod *Parhyale hawaiensis* were obtained from DOE Joint Genome Institute (http://genome.jgi-psf.org/parha/parha.info.html) and those from the porcelain crab *Petrolisthes cinctipes* from Tagmount [51] (http://sequoia.ucmerced.edu/PetrolESTes/index.php). We also investigated nucleotide collection (nr/nt), expressed sequence tags (est) and whole-genome shotgun (wgs) databases from the NCBI, the Marine Genomics Project database (http://www.marinegenomics.org) and the Penaeus Genome Database (http://sysbio.iis.sinica.edu.tw/page/). Similarity searches were performed using the TBLASTX program [52]. To avoid any bias that would favor the detection of GalEa clade elements [35], two different Copia elements were used as queries: the *Drosophila melanogaster* transposable element Copia (X02599.1) and the chimeric sequence of CoRex2 (herein described). Only the *pol* sequence of GyRex2 (herein described) was used as query to detect Gypsy elements.

To investigate the distribution of GalEa-like elements in all eukaryotes, we performed TBLASTX searches on all NCBI databases using GalEa1 (DQ913005.1) and Zeco1 (DQ91300) *pol* sequence as queries. When possible, chimeric sequences of the newly identified GalEa-like elements were designed. In few cases, the sequences from one species do not overlap themselves, we were thus unable to check whether they belong to the same element or not. Subsequently, we tested the GalEa clade affiliation of the newly identified elements using two different approaches: sequences covering the RT/RH domains were included into phylogenic analyses whereas the remaining sequences were classified using similarity searches using BLAST on the Gypsy Database 2.0 [19]. In the latter case, an element was assigned to the GalEa clade under the two conditions: (i) the five best hits must correspond to the five GalEa1-like elements referenced in the database; and (ii) the difference between the best E-values obtained with GalEa-like and other reference elements must be greater than 1e-10.

Sequence Analysis

Multiple alignments of DNA and protein sequences were constructed using MAFFT [53] and manually curated using BioEdit. Pairwise distances were estimated using the option pairwise deletion of gaps in MEGA5.0 [54] and the p-distance model. Amino acid consensus sequences of elements were

constructed by identifying the most common amino acid for each position. Ambiguously aligned sites within amino acid multiple alignments were removed using BMGE [55]. Phylogenic analyses were conducted using the Neighbor Joining method [56] and the best-fit model JTT+G [57] in MEGA5.0. For all phylogenic analyses, individual clade support was evaluated by non-parametric bootstrapping [58] using 100 bootstrap replicates.

Accession Numbers

The sequences obtained in this study have been submitted to the GenBank database (GenBank: HF548722–HF548824).

The accession numbers of the Copia elements used in phylogenetic analyses are:

Drosophila melanogaster 1731, X07656.1; Xanthias, FJ238509.1; *Arabidopsis thaliana* Araco, AC079131.4; Endovir1-1, AY016208.1; *Drosophila simulans* Copia, D10880.1; *Phaeodactylum tricornutum* CoDi4.4, EU432484.1; CoDi5.1, EU432486.1; CoDi6.4, EU432495.1; CoDi6.6, EU432497.1; CoDi7.1, EU432499.1; *Thalassiosira pseudonana* CoDi5.5, EU432490.1; CoDi6.1, EU432492.1; CoDi6.2, EU432493.1; *Zea mays* Hopscotch, AC084320.10; Opie-2, AC104473.2; Sto-4, AF082133.1; *Nicotiana tabacum* Tnt-1, X13777.1; Tto1, D83003.1; *Volvox carteri* Osser, X69552.1; *Oryza longistaminata* Retrofit, AH005614.1; *Saccharomyces exiguus* Tse1, AJ439547.1; *Saccharomyces cerevisiae* Ty4, M94164.1; *Vitis vinifera* Vitico1-1, AM465428.1; *Bombyx mori* Yokozuna, AB014676.1.

The accession numbers of the Gypsy elements are:

D. melanogaster 17.6, X01472.1; 297, X03431.1; Gypsy, M12927.1; Idefix, AJ009736.1; Springer, AF364549.1; *Trifurcates gratilla* SURI., M75723.1; *Beta vulgaris* Beetle1, AJ539424.1; *Schistosoma mansoni* Boudicca, AY662653.1; *Colletotrichum gloeosporioides* Cgret, AF264032.1 and AF264028.1; *Z. mays* Cinful-1, AF049110.1; CRM, AY129008.1; *Lycopersicon esculentum* Galadriel, AF119040.1; *A. thaliana* Gimli, AL049655.2; *Magnaporthe grisea* Grasshopper, M77661.1; MGLR3, AF314096.1; *Hydra magnipapillata* Hydra2-1, NW_002123104.1; *Pinus radiata* Ifg7, AJ004945.1; *B. mori* Kabuki, AB032718.1; Mag, X17219.1; *Musa acuminata* Monkey, AF143332.1, AF399948.1 and AF399938.1; *Drosophila buzzatii* Osvaldo, AJ133521.1; *Pisum sativum* Peabody, AF083074.1; *Alternaria alternata* Real, AB025309.1; *Oryza sativa* Retrosat-2, AF111709.1; RIRE2, AB030283.1; *Fusarium oxysporum* Skippy, L34658.1; *Strongylocentrotus purpuratus* SPM, NW_001353090.1; *Takifugu rubripes* Sushi-ichi, AF030881.2; *Autographa californica nucleo polyhedrosis* Ted, M32662.1; *Schizosaccharomyces pombe* Tf1, M38526.1; Tf2, L10324.1; *Drosophila virilis* Ulysses, X56645.1; *Ceratitis capitata* Yoyo, U60529.1; *Oryzias latipes* LReO-3, BA000027.2; *Sparus aurata* Saugg1, HQ021461.1. Some DIRS1-like elements were also used as phylogenetic outgroup: *Tetraodon nigroviridis* TnDIRS1, AF442732.1; *Tribolium castaneum* TcDIRS1, AY531876.1; *Strongylocentrotus purpuratus* SpDIRS1, biocadmin.otago.ac.nz/fmi/xsl/retrobase/home.xsl.

Ethics Statement

No specific permits were required for the described field studies. The sampled locations are not privately-owned or protected in any way, and the field studies did not involve endangered or protected species.

Results

Characterization of Copia and Gypsy Elements in *R. exoculata*

To isolate Copia and Gypsy retrotransposons in the hydrothermal shrimp *R. exoculata*, we performed PCR amplifications using degenerate primers. The CD1 and CD2 primers, designed within the conserved "KARLVA" and "YLDD" motifs of the RT (Figure 1), allowed us to amplify and sequence six Copia fragments of ~400 bp. The analysis of these fragments revealed 3 families we called CoRex1-3. The GD1 and GD2 primers, designed within the "RMPFGL" and "LTTDAS" conserved motifs of the RT and RH, led to the identification of 4 Gypsy fragments that cluster into 3 families we called GyRex1-3.

A fast and efficient strategy characterizing all parts of a chimeric full-length retrotransposon in 4 to 5 walking steps (Figure S1) was used on the CoRex1-3 and GyRex1-3 fragments. It associates three complementary walking approaches: the 'PCR walking' and 'TE Walking', as previously described for the characterization of the GalEa and Alvi elements [17,35], and a new method we developed and called 'PBS walking'. This method allows the coverage of the region from the Primer Binding Site (PBS) to the RT in only one walking step (see Material/Methods).

CoRex1 is represented by a 4949 bp chimeric consensus sequence (Figure 1-A), which includes two 217 bp LTRs, and is surrounded by the dinucleotides 5'-TG…CA-3' commonly observed in retrotransposons. The internal region carries a PBS sequence (TGGTAGCAGAGC; position 219), identical to the GalEa1 element PBS and complementary to the 3' end region of *D. melanogaster* tRNAMet gene, and a putative PolyPurine Tract (PPT) signal ($A_5GA_3GAG_3ACGAG$; position 4715). CoRex1 comprises two ORFs (Open Reading Frame). The first ORF encodes a *gag* region (288 amino acids) that holds the zinc-finger motif ($CX_2CX_4HX_4C$) found in all retroviral gag genes. The second ORF exhibits the domains of *pol* region in the order characteristic to Copia: (1) the protease (PR) domain with the typical 'DSGA' motif substituted by a 'DTGC' motif; (2) the integrase (INT) domain with its zinc-finger motif ($HX_4HX_{30}CX_2C$) and DD35E signature; (3) the reverse transcriptase and RNaseH (RT/RH) domains containing all the subdomains of RT sequences [11,12] and the highly conserved TRPDI motif of the RH. CoRex2 is represented by a 4875 bp chimeric consensus sequence (Figure 1-A) harboring shorter LTRs (133 bp) than CoRex1. However, CoRex1 and CoRex2 share the same LTR termini (5'-TGTTA; TATCA-3'). CoRex2 also shares the same PBS as CoRex1 and harbors a putative PPT at the position 4616 ($A_2GAGA_6G_2AG_4GAGA$). We identified a 3220 bp *pol* region (our chimeric sequence including a stop codon at the position 1537 and two frameshifts at the positions 1202 and 3934) that exhibits all the Copia domains and signatures. Upstream of its *pol* region, CoRex2 comprises an altered 522 bp sequence that harbors however the *gag* zinc-finger motif and shows similarity with the *gag* region. Finally, we were not able to characterize CoRex3 in full-length. CoRex3 is represented by a 4128 bp chimeric sequence from the PBS (identical to the CoRex1-2 PBS) to the 3' end of the RT domain (Figure 1-A). All characteristic domains can be found although the *gag* appears highly mutated.

The GyRex1 element is represented by a 4945 bp sequence comprising all domains from the *gag* region to the INT (Figure 1-B). The first 366 amino acid ORF could correspond to the *gag* region, according to similarity searches and the presence of a zinc-finger motif (position 940). The *pol* region (>3330 bp) shows all the signatures from PR to INT domains (but harbors one frameshift). GyRex2 is represented by a 5585 bp chimeric consensus sequence (Figure 1-B), including two 358 bp LTRs surrounded by the dinucleotides 5'-CT…AA-3'. It harbors a PBS sequence (TGGTGACCCTGAAGTA; position 467) complementary to the 3' end region of a *D. melanogaster* tRNATrp gene and similar to the PBS of the Boudicca element from *Schistosoma mansoni* (AAT98609; E-value = $4e^{-157}$ between GyRex2 and Boudicca).

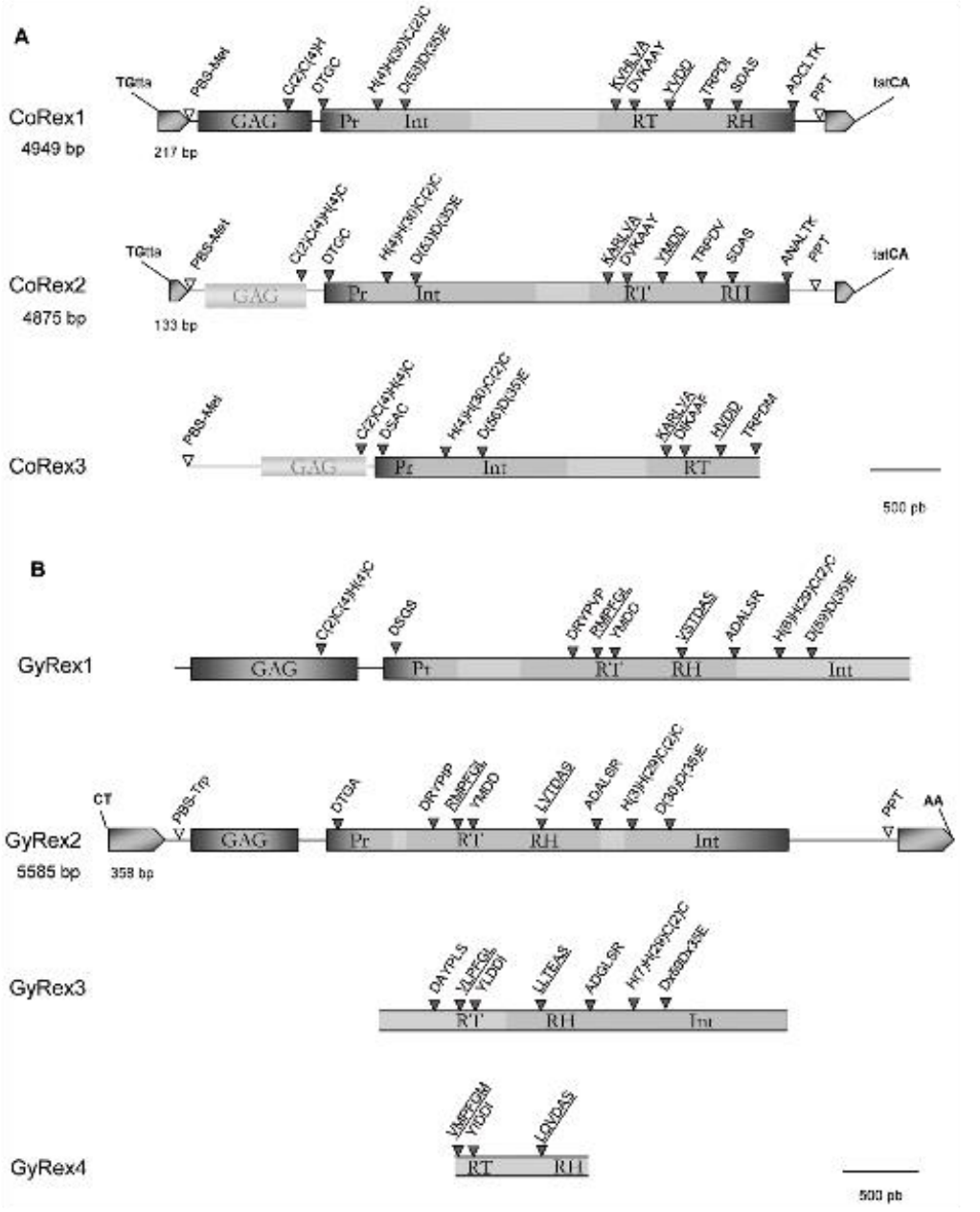


Figure 1. CoRex (A) and GyRex (B) retrotransposons annotation. When an element is described in full-length, its size (in bp), the size of its LTRs and its bordering nucleotides are given. The *gag* and *pol* regions are represented using grey blocks and their conserved domains are indicated by black triangles. Light grey blocks show putative altered *gag* regions. Positions of the Primer Binding Site (PBS) and the PolyPurine Tract (PPT) are indicated by white triangles.
doi:10.1371/journal.pone.0057675.g001

This allowed us to perform the 'PBS walking'. A putative PPT signal ($A_2GA_5T_2AG_3AG$) is observed at the position 5131. GyRex2 harbors two ORFs: (i) a first 235 codon ORF corresponding possibly to the *gag* region even if no zinc-finger motif can be identified, (ii) a second ORF exhibiting the signatures and domains in the order characteristic of Gypsy *pol* region. GyRex3 is only represented by a fragment of the *pol* region (2698 bp) that includes the RT, RH and INT domains (Figure 1-B).

The CoRex1-3 and GyRex1-3 characterization led also to the artifactual amplification of 3 new non-targeted Gypsy elements (GyRex4-6). GyRex4 was identified in its RT/RH domains (Figure 1-B) and appears highly divergent from GyRex1-3 (<33% identity on the 898 bp). GyRex5 is characterized by a 671 bp INT sequence that encompasses a zinc-finger and the DD35E signature. Interestingly, both GyRex4 and GyRex5 show high similarity to an element from the gilt-head bream *Sparus aurata* we called Saugg1 (HQ021461.1), which possesses the same structure than Gmr1-like retrotransposons. Gmr1-like elements are unconventional Gypsy retrotransposons in which the INT domain lies upstream, rather than downstream, of the RT domain [59]. Since the GyRex4-5 sequences do not overlap themselves, they could thus possibly belong to the same element. GyRex6 is represented by a 1160 bp sequence from the PBS position to the beginning of the *pol* region ('DTGA' motif of PR domain at the position 1145), and includes a potential 221 codon *gag* ORF. GyRex6 differs from GyRex1 and GyRex2, but here again we cannot exclude that it does not correspond to a portion of GyRex3-5 because of the lack of overlapping sequences.

Sequences corresponding to three other transposable elements were also identified: two new LINE retrotransposons (LiRex1-2) and one transposon (T-Rex1). LiRex1 (354 bp) appears highly corrupted, although the RT4 motif of the reverse transcriptase [11] is still detectable. LiRex2 (563 bp) is more conserved with the recognizable RT5, RT6 and RT7 motifs. Finally, the T-Rex1 sequence (675 bp) shows high similarity with a transposon from the sea urchin *Strongylocentrotus purpuratus* (XP001188275.1, E-value = $6e^{-54}$).

The *R. exoculata* specimens were collected on hydrothermal vents where they could have been subjected to stresses due to the hypervariability of the environment. They were also exposed to many stresses related to fishing conditions (decompression, temperature variations…) that could also favor the activation of TEs. We performed RT-PCRs on the *R. exoculata* transcripts using primers specific to each element. Transcriptional activity was revealed for CoRex1 and CoRex2. Three CoRex1 (>97% identity) and five CoRex2 (>87% identity) transcript sequences were identified (Table S1), highlighting a preponderance of CoRex2 on the other Copia families within *R. exoculata*. No transcript could be detected for GyRex1-4 and Corex3, which however do not attest to their inactivity in other specimens or conditions.

To determine the CoRex1-3 and GyRex1-6 distributions among species related to *R. exoculata*, we PCR-screened 4 other Alvinocarididae species (*A. lusca*, *A. markensis*, *C. chacei* and *M. fortunata*) as well as two closely related non-hydrothermal shrimps (*C. crangon* and *P. serratus*, [60]) using few combinations of specific primers for each element (Table S1). Elements related to CoRex1-3 and GyRex2 are detected in all hydrothermal shrimps, except CoRex1 that could not be identified in *M. fortunata*. This led to the identification of several new elements: CoAlma1 (*A. markensis*) and CoAllu1 (*A. lusca*) from the CoRex1 family (>97% identity); CoMiro2 (*M. fortunata*), CoAlma2 and CoAllu2 from the CoRex2 family (>87% identity); CoAlma3 and CoMiro3 from the CoRex3 family (>90% identity); and GyMiro2 and GyAlma2 from the GyRex2 family (>79% identity). Finally, Gychoro2, an element that belongs to the same family than GyRex4 (93% identity), was detected in *C. chacei*, whereas GyRex1, GyRex3, GyRex5 and GyRex6 could not be detected in any other species.

Copia and Gypsy Retrotransposons in Crustaceans

To estimate the diversity of Copia and Gypsy elements within crustaceans, we PCR-screened 25 decapods and crustacean species using degenerate primers. We additionally looked for retrotransposons in the crustacean genomic and transcriptomic databases using similarity searches. These two complementary approaches led to the identification of 35 Copia and 46 Gypsy elements distributed among 15 and 18 species, respectively (Figure 2). Sixteen and twenty-nine of these Copia and Gypsy elements were included in phylogenetic analyses based on the RT/RH domain and the remaining sequences were classified using a BLAST-based approach (see Materials and Methods and Table S2).

Gypsy retrotransposons from crustaceans are divided in several clades (Figure 3). One third of the elements group in the CsRN1 clade, including elements from the copepod salmon lice *Lepeophtheirus salmonis* (GyLesa1 and GyLesa5), the cirriped barnacle *S. sarsini* (GySac2) and diverse decapods such as *R. exoculata* (GyRex2), crabs (e.g. GyBy1 from *B. thermydron*), squat lobsters (GyMur1 from *M. recta*). This clade also includes the GyPaha1-3 elements from the amphipod *P. hawaiensis* (Table S2). The Mag clade encompasses seven elements from the branchiopod *D. pulex* (GyDpu15 and GyDpu25), the copepod (GyLesa2 and 3), the cirriped (GySac1), and the krill *E. crystallorophias* (GyEcrys1). To date no Mag clade element has been identified in decapods. Four elements appear to be related to the Gmr1 clade: GyRex4 and Gychoro2 (hydrothermal shrimps), GyMaja1 (spiny spider crab *M. squinado*) and GyLiva4 (prawn *L. vannamei*), which are the first Gmr1-like elements described in protostomes. Several new clades may be also identified using the crustacean elements. For example, GyRex1 seems closely related to GyOrli1 (crayfish *O. limosus*), and the GyLiva6 and GyPeme2 elements from the prawns *L. vannamei* and *P. monodon* are grouped in a very well supported clade. The remaining elements appear to be more or less dispersed within the phylogeny and do not belong to any previously identified clade. Finally, the Gypsy tree mostly differs from the crustacean phylogeny. Clades include elements from distant species and elements from one species belong to distant clades. For example, in *R. exoculata*, GyRex2 is a CsRn1-like element and GyRex4 a Gmr1-like, while GyRex1 and 3 do not belong to any previously defined clade. Three elements from *D. pulex* group into the Mag clade while the two others remain isolated in the phylogeny. The four GyLiva (*L. vannamei*) are divided among four different clades, and the GyLesa (*L. salmonis*) and GySac (*S. sarcini*) elements are split among the CsRN1 and the Mag clades.

In contrast to the Gypsy retrotransposons, the 35 Copia elements from crustaceans appear much less diversified, as they all fall into three clades (Figure 4). Seven of these sequences were previously described as GalEa-like elements [35], including the well-annotated GalEa1 elements (galatheid squat lobsters). Twenty-one new elements, including the CoRex1-3 retrotransposons, belong to this highly supported GalEa clade (Figure 4 and Table S2). It is interesting to note that in terms of diversity various species harbor several GalEa-like families (e.g. at least 4 detected in the *E. superba* transcriptome, 3 in *P. hawaiensis* and 3 in *E. annulosa* genome). The 6 remaining elements belong to three different clades: (i) The three elements from *D. pulex*, which correspond to the two subgroups defined by Rho *et al.* [32], grouped together in

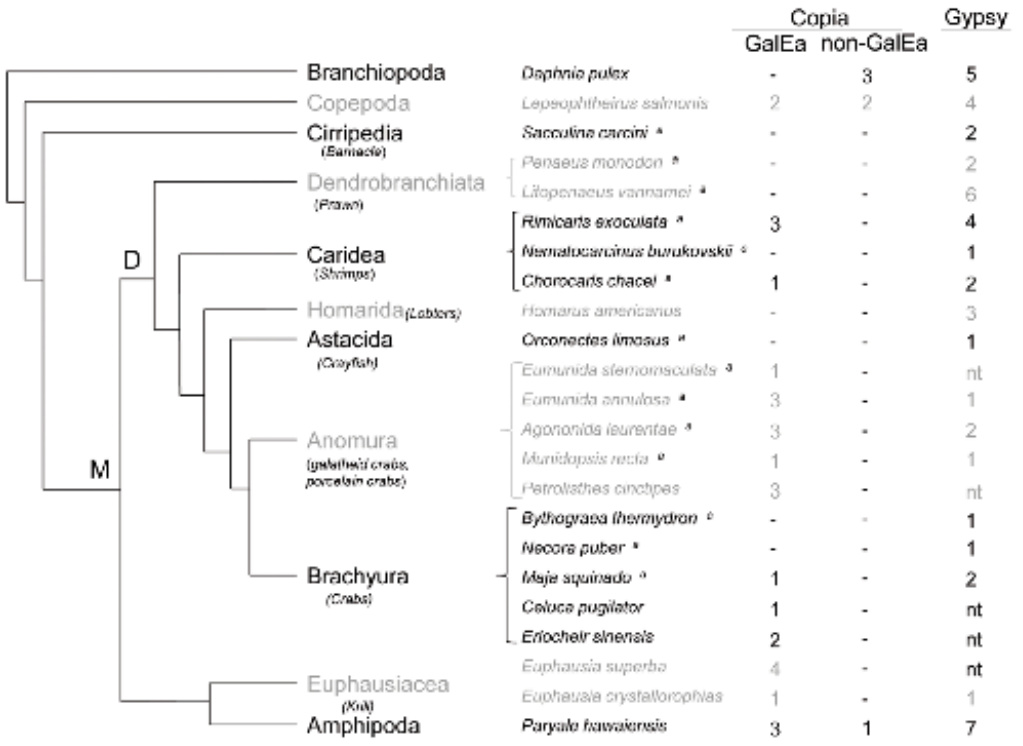


Figure 2. Number of Copia and Gypsy elements studied in crustaceans. Genetic relationships between crustacean classes and orders are represented by a tree topology reconstructed from previous studies (Regier *et al.* 2010, Giribet and Edgecombe, 2011; Ahyong and O'Meally, 2004). **M**: Malacostraca. **D**: Decapoda. For Copia retrotransposons, GalEa and non-GalEa elements are distinguished. Only a few representatives of the Copia elements described in *D. pulex* were studied. nt: not tested; -: no element detected; [a] species screened using degenerate PCRs.
doi:10.1371/journal.pone.0057675.g002

a single clade we called CoDpu; (ii) CoLesa1 (ADNI02013164.1) and Colesa4 (ADNI02043341.1) from the copepod *L. salmonis* grouped in a new clade we called CoLesa1-like that is related to the Sireviruses; and (iii) similarity searches on the CoPaha4 element from the amphipod *P. hawaiensis* revealed that this element is likely related to the Hydra clade (Hydra1-2, E-value = $9e^{-50}$). Interestingly, an additional screen of another Daphnia species, *Daphnia pulicaria* (http://wfleabase.org/blast/), could only reveal Copia elements that belong to the CoDpu clade (data not shown).

Discussion

Crustaceans: a Suitable Study System for Transposable Element Dynamics

Given their abundance, high level of phylogenetic diversification, huge diversity of environment and life styles, and extended range in C-values with particularly large genomes (460-fold variation from 0.14 to 64.62 pg [22]), crustaceans appear a worthy focus for comparative study of metazoan genomes at an intermediate scale (*i.e.* within a subphylum or a class). Crustaceans also appear as one suitable system for a comparative genome evolution study with Hexapoda, one of the most studied group in biology. Indeed, crustaceans are, for example, the second most studied group of "invertebrates", after hexapods, for genome size reports (318 species reported in the Genome Size Database, Gregory 2008). However, crustaceans remain greatly underrepresented in genomics. Only few large-scale genomic sequencing analyses, restricted to branchiopods, have been performed [32]. Nevertheless, the emergence of next-generation sequencing technologies now allows comparative genomic studies for non-model species and/or large genomes [61–64], and led to the recent acquisition of genomic and more especially transcriptomic data for several crustacean species.

Among crustaceans, we focused on *R. exoculata*, listed as a model organism of an extreme deep-sea environment (CAREX, 2010), which dominates the vagile megafauna at many hydrothermal vent sites along the Mid-Atlantic Ridge. *R. exoculata* has been studied in many aspects, such as biogeography/population genetics [65,66], bacterial symbiosis association [41,67] and response to stress [68,69]. *R. exoculata* could also represent an interesting model species for transposable element dynamics because of its extremely variable environment. Our present study, combined with the previous analysis of DIRS1-like retrotransposons in decapods [17], allows us to describe a great diversity of transposable elements in this species. At least 13 TE families are identified, including 2 tyrosine recombinase encoding elements

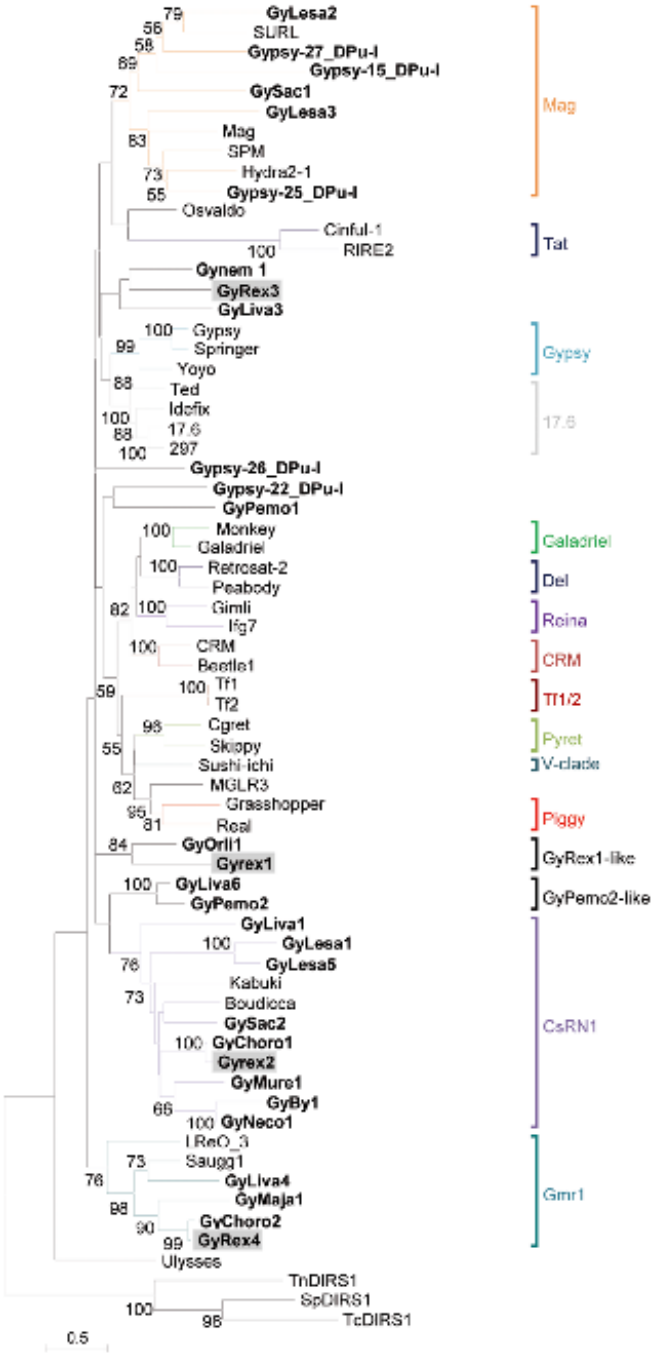

GyLesa2
SURL
Gypsy-27_DPu-I
Gypsy-15_DPu-I
GySac1
GyLesa3
Mag
SPM
Hydra2-1
Gypsy-25_DPu-I
Osvaldo
Cinful-1
RIRE2
Gynem 1
GyRex3
GyLiva3
Gypsy
Springer
Yoyo
Ted
Idefix
17.6
297
Gypsy-26_DPu-I
Gypsy-22_DPu-I
GyPemo1
Monkey
Caladriel
Retrosat-2
Peabody
Gimli
Ifg7
CRM
Beetle1
Tf1
Tf2
Cgret
Skippy
Sushi-ichi
MGLR3
Grasshopper
Rex1
GyOrli1
Gyrex1
GyLiva6
GyPemo2
GyLiva1
GyLesa1
GyLesa5
Kabuki
Boudicca
GySac2
GyChoro1
Gyrex2
GyMure1
GyBy1
GyNeco1
LReO_3
Saugg1
GyLiva4
GyMaja1
GyChoro2
GyRex4
Ulysses
TnDIRS1
SpDIRS1
TcDIRS1
0.5
Mag
Tat
Gypsy
17.6
Caladriel
Del
Reina
CRM
Tf1/2
Pyret
V-clade
Piggy
GyRex1-like
GyPemo2-like
CsRN1
Gmr1

Figure 3. Phylogenetic relationships among Gypsy retrotransposons inferred from Neighbor-Joining analysis of RT/RH amino acid sequences. The crustacean elements are indicated in bold and the four *R. exoculata* elements (GyRex) are highlighted in grey. Statistical support (>50%) comes from non parametric bootstrapping using 100 replicates. DIRS1-like sequences were used as outgroup.
doi:10.1371/journal.pone.0057675.g003

(Alvi1-2), 3 Copia (CoRex1-3) and 6 Gypsy (GyRex1-6) LTR-retrotransposons, as well as 2 LINEs (LiRex1-2) and one transposon (T-Rex1). We noticed that element detection using the degenerate primers approach is usually fairly easy in this species, which confirms the tendency observed during the detection of DIRS1-like elements in hydrothermal shrimps. This seems to be also the case for galatheid squat lobsters (e.g. *E. annulosa*), where a large diversity of retrotransposons is described (DIRS1-like, [17]; GalEa-like, [35]; Gypsy and Pao/Bel, [36]). We hypothesized [17] that such results can be partly related to the copy number in such species having a large genome size [23].

Copia Retrotransposons Seem Relatively Rare in Crustaceans

Thirty-four Copia retrotransposons are now identified in crustaceans. However, we often observed a lack of detection or very low PCR signals in the species we screened for Copia elements using degenerate primers. Although the degenerate primers were designed within very well conserved motifs (Table S3) and are known to be efficient [35,47,70], PCR-screenings led to the identification of 9 Copia retrotransposons in only 7 of the 14 species tested, including 6 Caridea and Anomoura spp. Besides, an additional PCR-screening of 10 other diverse crustaceans could not lead to the detection of any Copia elements. Set apart the choice of primers, the lack of PCR signal could simply be due to the rarity of the elements or their absence from the species studied. Indeed, even if CoDpu elements seem relatively abundant in *D. pulex* [32], the absence or rarity of Copia elements could be a genomic feature frequent in crustacean species. For example, none of these retrotransposons have been reported in repetitive element families of *P. monodon* [71]. Likewise, we could not identify any Copia elements in the well-sequenced transcriptome of *L. vannamei* (141030 contigs available in the *Penaeus* Genome Database: http://sysbio.iis.sinica.edu.tw/page/).

This feature is however not restricted to crustacean species since LTR-retrotransposons are known to be less abundant in animals [10]. Compared to their close relatives, the crustaceans do not differ from the other species. Indeed, de la Chaux and Wagner [21] recently reported that Copia elements have a small relative abundance in hexapods, Copia elements being usually much rarer than the Gypsy or Pao/Bel retrotransposons. They even appear to be absent in one species, *Ixodes scapularis*. In general, it has been shown that Copia elements constitute only a small proportion of LTR-retrotransposons identified in numerous metazoan genomes [21], as well as in fungi [72]. For example, only few were detected in the comparative analysis of TEs content from salamanders [73] and none are described in the draft genome of the pearl oyster [74].

Copia and Gypsy Retrotransposons: Two Opposite Dynamics in Crustaceans

In addition to the fact that Copia elements are much scarcer than Gypsy in metazoan genomes, Copia elements appear clearly less diverse. While studying the evolutionary history of LTR-retrotransposons in eukaryotes, Llorens [20] observed that Gypsy elements have been more successful than their Copia counterparts during evolution. The authors hypothesized that the higher phenotypic plasticity of Gypsy retrotransposons allowed them to diversify much more than Copia elements at distinct geological eras. Our phylogenic analyses of crustacean LTR-retrotransposons also fit this observation. We observed two diametrically opposed patterns for crustacean Copia and Gypsy elements (Figure 3 and 4). Even within a single species such as *R. exoculata*, its GyRex and CoRex elements follow this pattern. The Gypsy elements appear very diverse, widely dispersed among the phylogeny and many clades of Gypsy are represented or are newly described. This large diversity of Gypsy retrotransposons is probably inherited from an ancestral polymorphism in crustacean lineage, where several active element copies within species have been maintained. For example, many crustacean elements belong to the Mag clade, which is believed to be one of the oldest Gypsy clades [20]. The newly described clades (Gyrex1-like, Gynemo2-like; Figure 4) could also result from a diversification of Gypsy elements during the evolution of crustaceans. Alternatively, a higher rate of horizontal transfers could also lead to such diversity, but to date no argument supports this hypothesis. In contrast, the diversity of Copia retrotransposons in crustaceans appears much more restricted and related to the host species. The GalEa clade appears highly predominant comprising 29 elements detected in Decapoda, and more generally in Malacostraca (Figure 2). Only two elements from the copepod *L. salmonis* group into the new CoLesa1 clade, and one element from the amphipod *P. hawaiensis* appears to belong to the Hydra clade. Finally, all the Daphnia elements form the CoDpu clade.

The dynamics of transposable elements is a complex concept, which combines numerous aspects such transposition control mechanisms by the elements themselves and/or the host genome, the element activation by environmental changes (at the genome or ecological levels), the mutation rate, the host migration, the possible domestication events, etc. Moreover, many of these parameters are subject to random events (drift). To get a mental picture of GalEa dynamics, and presumably those of some other elements, we can draw an analogy with a "domino days spreading" branching process in which successive amplifications may interact positively. During the famous worldwide event of toppling domino stones, we can follow the propagation of domino falls along various branches and through several major figures that encompass large, but variable numbers of dominoes. Elements could be represented by the dominoes and the number of copies by the number of falling stones, helping to visualize the TE diffusion within taxa and species during evolution (except that domino structures are pre-designed). Like domino bricks following a restricted number of lines before toppling large structures, few active TEs copies must be inherited prior to a transposition "burst". Many factors could lead to such expansion within a species. For example, it is well illustrated that TE transposition can be activated by stresses [7,9,75] or the colonization of a new environment [76]. It has also been hypothesized that variations in the TE repertoire could promote or be associated with the emergence of new lineages, species, populations or subpopulations [77–79]. Later on, the large domino structures allow the progression to the next structure *via* several paths. Similarly, an initial amplification increases the proportion of young active elements, which allow subsequent derived amplifications in some random lineages, possibly through the transposition of few master copies. Furthermore, the limited number of toppling dominoes between figures may facilitate the random breaking off of their

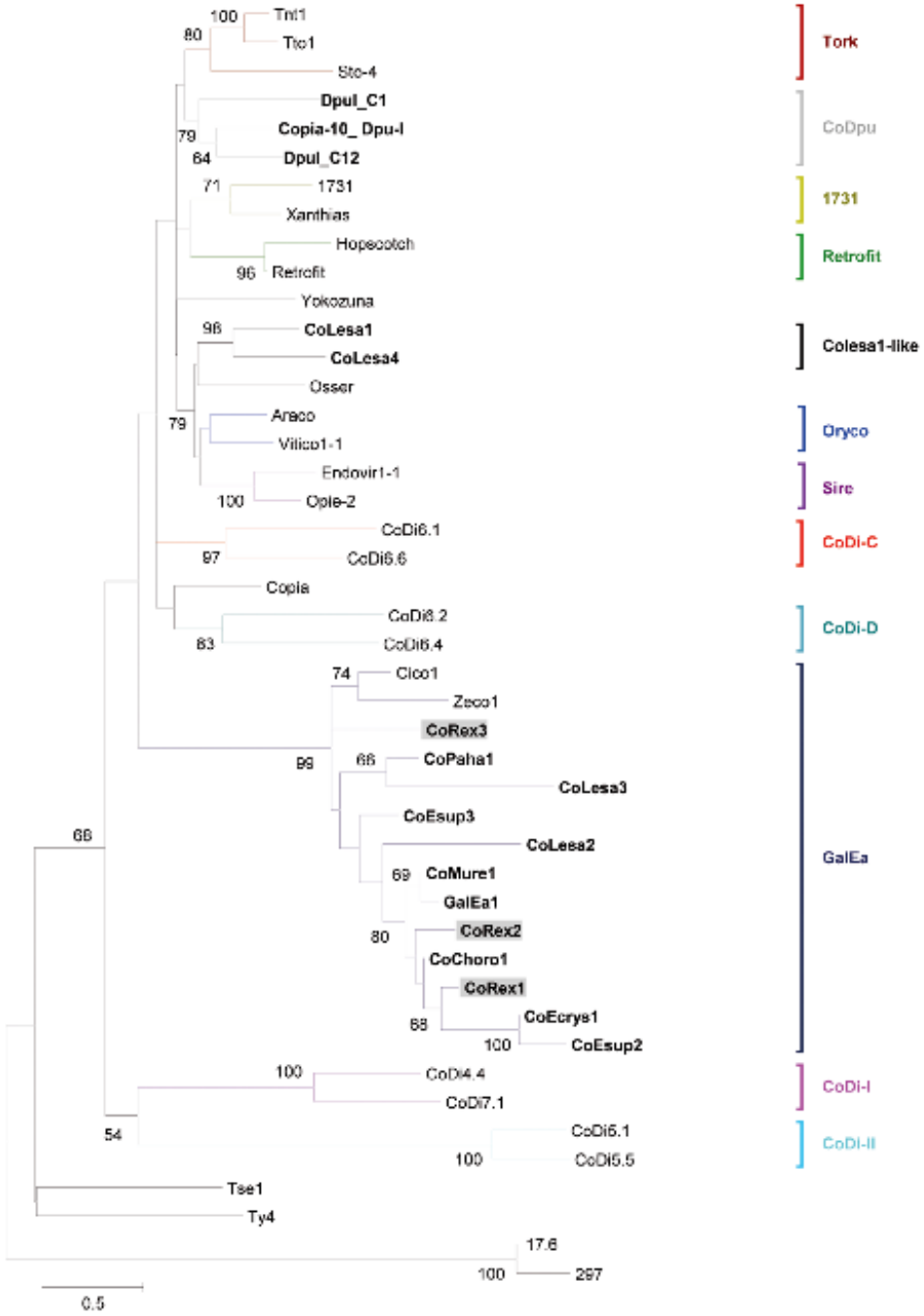

Tnt1
Ttc1
Sto-4
Dpul_C1
Copia-10_ Dpu-I
Dpul_C12
1731
Xanthias
Hopscotch
Retrofit
Yokozuna
CoLesa1
CoLesa4
Osser
Areco
Vilico1-1
Endovir1-1
Ople-2
CoDi6.1
CoDi6.6
Copia
CoDi8.2
CoDi6.4
Clco1
Zsco1
CoRex3
CoPaha1
CoLesa3
CoEsup3
CoLesa2
CoMure1
GalEa1
CoRex2
CoChoro1
CoRex1
CoEcrys1
CoEsup2
CoDi4.4
CoDi7.1
CoDi5.1
CoDi5.5
Tse1
Ty4
17.6
297
0.5
Tork
CoDpu
1731
Retrofit
Colesa1-like
Oryco
Sire
CoDi-C
CoDi-D
GalEa
CoDi-I
CoDi-II

Figure 4. Phylogenetic relationships among Copia retrotransposons inferred from Neighbor-Joining analysis of RT/RH amino acid sequences. The crustacean elements are indicated in bold and the three *R. exoculata* elements (CoRex) are highlighted in grey. Statistical support (>50%) comes from non parametric bootstrapping using 100 replicates. Gypsy sequences were used as outgroup.
doi:10.1371/journal.pone.0057675.g004

progression along some paths. Similarly, evolutionary forces may drive the extinction of some elements within a lineage when elements are maintained too long at a low copy number. In a funny parallel, the high diversity of dominoes features may also reflect the element diversity and the evolution. Changes in the material or color of dominoes, which are much more numerous in the figures, may reflect TEs mutations and the recent use of "slow stones" may represent variable speed of evolution. Likewise, to ensure the success of major figures, builders design rescue paths in case of failure of the main circuit, which can easily be compared to the TE dynamics through horizontal transfers.

In the case of the crustacean phylogeny, such a model could have led to the current GalEa distribution and could explain the three transitions observed in the Copia content: (i) the expansion of GalEa-like elements in a common ancestor; (ii) the predominance of GalEa-like elements in decapods and euphausiaceans; and (iii) the loss of Copia elements in some species. The expansion of GalEa-like elements prior to multicrustacean radiation is supported by their presence in most Malacostraca and in the only copepod tested. It could be hypothesized that GalEa-like elements have been horizontally transferred to the multicrustacean ancestor (*i.e.* Copepoda, Cirripedia and Malacostraca according to Regier [80]) and then invaded its genome. However, since they are present in various metazoans (see section below), GalEa-like elements should have been already present in the multicrustacean common ancestor. The GalEa-like element absence in branchiopods remains to be confirmed by the study of other species outside the daphnia group. The phylogenetic distance between Branchiopoda and Multicrustacea supports this hypothesis. Indeed, the relationships within Pancrustacea remain controversial as several studies describe Branchiopoda as a sister-group to Hexapoda instead of Multicrustacea [81–83]. In such a case, Copia retrotransposons from branchiopods are expected to be as different from GalEa as those observed in hexapods [20].

In addition to the GalEa-like element distribution, the detection of several other Copia elements in amphipods and copepods suggests that the Copia repertoire of crustacean or multicrustacean ancestor comprises elements from several clades. Since the GalEa-like elements appear to be exclusive to decapods and euphausiaceans, by implication the other Copia retrotransposons have been rarely amplified and have been progressively lost. Most likely, a slow mutational decay of other Copia retrotransposons, which are usually in low copy number except in plants [21,35,72], led to this loss in many lineages. Besides, the success in maintaining GalEa-like elements within multicrustaceans appears to be species- or lineage-specific. The fact that only some Copia are able to counteract the evolutionary erosion forces suggests that the dynamics of the different elements may be related to particular ability of each of them to amplify under peculiar conditions in some genomes. For example, the tobacco Tnt1 retrotransposons tightly control their activation by restricting expression to specific conditions, as they possess in their promoter regulatory motifs similar to those involved in activation of plant defense genes [7,75,84]. GalEa-like elements seem to have been lost in few species, such as prawns, while they seem to have undergone some secondary expansions in others infraorders, such as in galatheid squat lobsters or caridean shrimps. This could explain their uneven distribution among Decapoda. Interestingly, similar expansions of DIRS1-like elements have also been observed in these lineages [17].

To reinforce the idea that few specific Copia elements could, from time to time, increase their transpositional activity and so broaden their occurrence in some particular host taxa, it appears necessary to study Copia diversity in other metazoan groups at roughly the same scale of study. For this, it may be interesting to survey the distribution of CoDpu-like elements among Branchiopoda, and/or to study Copia elements diversity in another taxon such as Hexapoda. To date, six clades of Copia retrotransposons have been described in wingled hexapods: 1731, Copia, GalEa, Humnum, Mtanga and Tricopia [20]. Interestingly, as observed in crustaceans, the distribution of TE clades among species appears also highly related to the host phylogeny. For example, whereas the Copia clade is widely distributed in Insecta [85–87], the Tricopia, Mtanga and Humnum clades have been detected in only one species [20,88].

GalEa-like Retrotransposons in Eukaryotes

The success of GalEa-like elements in crustaceans raises the question of their distribution in others organisms. When they defined the GalEa clade, Terrat *et al.* [35] described GalEa1 related elements in 3 fishes and 1 acidian. The GalEa clade is actually more widely distributed among animals. We retrieved GalEa-like retrotransposons through BLAST searches using GalEa1 and Zeco1 *pol* domain as queries, which now allow us to report such elements in more than 50 species (Table S2). Many of these species are of course crustaceans (16 species). There are also numerous fishes (18 species), as GalEa-like elements appears widely distributed in teleost fishes, which are the subject of many sequencing projects. GalEa-like elements are also present in diverse molluscs (7 species), as well as some Chordata, Cnidaria, Ctenophora, Echinoderma and Hemichordata. Two elements (CoPorcru1 and CoPorcru2) were also detected outside metazoans, in the red algae *Porphyridium cruentum*. This fits the previous identification of some similar GalEa-like elements in another red algae, *Porphyra yezoensis* (PyRE10G, AB286055) and suggests that GalEa-like elements are probably ancient in eukaryotes, at the exception of the hypothesis of multiple horizontal transfers. To determine the relatedness between these different GalEa-like retrotransposons, we performed a phylogenetic analysis based on the RT-RH domain of 42 elements that represent 33 species (Figure S2). Within the well-supported GalEa clade (bootstrap value 92%), the two red algae elements (CoPorcru1-like) form a distinct group from all other elements. Three other groups can also be defined. Almost all elements from crustaceans group in a same subclade (CoRex1-like), except CoRex3 and CoLesa2. Likewise, all elements from fishes belong to a monophyletic group (bootstrap value 97%) and form, with CoCre1 (*Crepidula fornicata*) and CoSaccoglo1 (*S. kowalevskii*), a subclade we called Zeco1-like (bootstrap value 89%). The last subclade, CoPali1-like (bootstrap value 99%), contains one element from the sea urchin *Paracentrotus lividus* and one from the cuttlefish *Sepia officinalis*. The remaining elements, especially those from molluscs, appear more or less dispersed within the phylogeny. GalEa-like elements have a widespread distribution, being highly represented in at least 3 groups of organisms: Malacostraca, Teleostei and probably part of Mollusca. For a better understanding of the distribution of GalEa-like retrotransposons, we wonder whether their predominance is

a peculiar feature of Malacostraca, or whether similar feature can be observed in other species clades.

Supporting Information

Figure S1 Characterization strategy of full-length LTR-retrotransposons. A copia retrotransposon is used as example. For each of the five steps, the known part of the element is represented by a full line whereas the walking part is indicated by colored dotted arrow: red, PCR or TE Walking; green, PBS Walking; purple: PCR using specific primers. The conserved domains used to design the degenerate primers and the PBS sequences are represented by blue and green triangles, respectively.
(TIF)

Figure S2 Phylogenetic relationships among GalEa-like retrotransposons inferred from Neighbor-Joining analysis of RT/RH amino acid sequences. Statistical support (>50%) comes from non parametric bootstrapping using 100 replicates. Two to three representative elements of the other Copia clades are also included to the phylogeny. Gypsy sequences were used as outgroup.
(TIF)

Table S1 Report of the sequences obtained from PCR approaches. For each element, the host species, name, length and accession number are given, as well as the PCR methodology and primers used.
(XLSX)

Table S2 List of GalEa-like retrotransposons identified. For each element, the corresponding host species and the accession number(s) are indicated. The GalEa nature of the elements was determined following different classification methods: Figure B and SupData E correspond to the phylogenetic analyses; BlastP to the BLAST-based classification method, for which the best GalEa and non-GalEa hits are given with the corresponding E-values.
(XLSX)

Table S3 Comparison of CD1 and CD2 primers with Copia sequences. Dissimilarities at nucleic or amino-acid levels are indicated in red.
(XLSX)

Acknowledgments

We are grateful to Stephane Hourdez and Nicolas Rabet for generously providing samples. We thank Jean Yves Toullec for freely providing transcriptomic sequences from Antarctic krill and Nelly Léger for rimicaris RNA sample. We kindly acknowledge Angela Atwood-Moore for English revisions of the manuscript. The authors wish also to thank chief scientists, captains and crews of the oceanographic cruises (Norfolk1 2001, Norfolk2 2003, BIOSPEEDO 2005, MoMARETO 2006, MoMARDREAM-Naut 2007 and MESCAL 2010) and the crew of the submersibles (Nautile and ROV Victor6000). We would like to thank two anonymous referees for useful comments on this manuscript.

Author Contributions

Carried out the in silico element detection: DH EB. Carried out molecular analyses: MP TD CE PG. Performed the phylogenetic analyses: MP TD. Conceived and designed the experiments: EB. Analyzed the data: TD CE. Wrote the paper: MP EB.

References

1. Finnegan DJ (2012) Retrotransposons. Curr Biol 22: R432–437. doi:10.1016/j.cub.2012.04.025.
2. Bémont C, Vieira C (2006) Genetics: junk DNA as an evolutionary force. Nature 443: 521–524. doi:10.1038/443521a.
3. Kazazian HH Jr (2004) Mobile elements: drivers of genome evolution. Science 303: 1626–1632. doi:10.1126/science.1089670.
4. Finkeroff NV (1999) Transposable Elements As a Molecular Evolutionary Force. Annals of the New York Academy of Sciences 870: 251–264. doi:10.1111/j.1749-6632.1999.tb08886.x.
5. Bennetzen JL (2000) Transposable element contributions to plant gene and genome evolution. Plant Mol Biol 42: 251–269.
6. Capy P, Gasperi G, Biémont C, Bazin C (2000) Stress and transposable elements: co-evolution or useful parasites? Heredity (Edinb) 85 (Pt 2): 101–106.
7. Melayah D, Bonnivard E, Chalhoub B, Audeon C, Grandbastien MA (2001) The mobility of the tobacco Tnt1 retrotransposon correlates with its transcriptional activation by fungal factors. Plant J 28: 159–168.
8. Lisch D (2009) Epigenetic regulation of transposable elements in plants. Annu Rev Plant Biol 60: 43–66. doi:10.1146/annurev.arplant.59.032607.092744.
9. Fablet M, Vieira C (2011) Evolvability, epigenetics and transposable elements. BioMol Concepts 2: 333–341. doi:10.1515/BMC.2011.035.
10. Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, et al. (2007) A unified classification system for eukaryotic transposable elements. Nat Rev Genet 8: 973–982. doi:10.1038/nrg2165.
11. Xiong Y, Eickbush TH (1990) Origin and evolution of retroelements based upon their reverse transcriptase sequences. EMBO J 9: 3353–3362.
12. Capy P, Langin T, Higuet D, Maurer P, Bazin C (1997) Do the integrases of LTR-retrotransposons and class II elements transposases have a common ancestor? Genetica 100: 63–72.
13. Ohshima K, Okada N (2005) SINEs and LINEs: symbionts of eukaryotic genomes with a common tail. Cytogenet Genome Res 110: 475–490. doi:10.1159/000084981.
14. Piskurek O, Jackson DJ (2011) Tracking the ancestry of a deeply conserved ruminant SINE domain. Mol Biol Evol 28: 2727–2730. doi:10.1093/molbev/msr115.
15. Arkhipova IR (2006) Distribution and phylogeny of Penelope-like elements in eukaryotes. Syst Biol 55: 875–885.
16. Piednoël M, Gonçalves IR, Higuet D, Bonnivard E (2011) Eukaryote DIRS1-like retrotransposons: an overview. BMC Genomics 12: 621. doi:10.1186/1471-2164-12-621.
17. Piednoël M, Bonnivard E (2009) DIRS1-like retrotransposons are widely distributed among Decapoda and are particularly present in hydrothermal vent organisms. BMC Evol Biol 9: 86. doi:10.1186/1471-2148-9-86.
18. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, et al. (2005) Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res 110: 462–467. doi:10.1159/000084979. Accessed June 2012.
19. Llorens C, Futami R, Covelli L, Domínguez-Escribá L, Viu JM, et al. (2011) The Gypsy Database (GyDB) of mobile genetic elements: release 2.0. Nucleic Acids Res 39: D70–74. doi:10.1093/nar/gkq1061. Accessed June 2012.
20. Llorens C, Muñoz-Pomer A, Bernad L, Botella H, Moya A (2009) Network dynamics of eukaryotic LTR retroelements beyond phylogenetic trees. Biol Direct 4: 41. doi:10.1186/1745-6150-4-41.
21. de la Chaux N, Wagner A (2011) BEL/Pao retrotransposons in metazoan genomes. BMC Evol Biol 11: 156. doi:10.1186/1471-2148-11-156.
22. Rees DJ, Belelle C, Glémet H, Dufresne F (2008) Large genomes among raridean shrimp. Genome 51: 159–163. doi:10.1139/g07-108.
23. Bonnivard E, Catrice O, Ravaux J, Brown SC, Higuet D (2009) Survey of genome size in 28 hydrothermal vent species covering 10 families. Genome 52: 524–536. doi:10.1139/g09-027.
24. Gilbert C, Schaack S, Fruchonne C (2010) [Mobile elements jump between parasites and vertebrate hosts]. Med Sci (Paris) 26: 1025–1027. doi:10.1051/medsci/20102612025.
25. Roy-Engel AM (2012) LINEs, SINEs and other retroelements do birds of a feather flock together? Front Biosci 17: 1345–1361.
26. Brookfield JFY (2011) Host–parasite relationships in the genome. BMC Biol 9: 67. doi:10.1186/1741-7007-9-67.
27. Burke WD, Malik HS, Jones JP, Eickbush TH (1999) The domain structure and retrotransposition mechanism of R2 elements are conserved throughout arthropods. Mol Biol Evol 16: 502–511.
28. Schön I, Arkhipova IR (2006) Two families of non-LTR retrotransposons, Syrinx and Daphne, from the Darwinulid ostracod, Darwinula stevensoni. Gene 371: 296–307. doi:10.1016/j.gene.2005.12.007.
29. Rho M, Tang H (2009) MGEScan-non-LTR: computational identification and classification of autonomous non-LTR retrotransposons in eukaryotic genomes. Nucleic Acids Res 37: e143. doi:10.1093/nar/gkp752.
30. de la Vega E, Degnan BM, Hall MR, Wilson KJ (2007) Differential expression of immune-related genes and transposable elements in black tiger shrimp (Penaeus monodon) exposed to a range of environmental stressors. Fish Shellfish Immunol 23: 1072–1088. doi:10.1016/j.fsi.2007.05.001.

31. Hurst SE, Tamuls WG, Robertson LM, Garcia DK (2008) Evidence of multiple retrotransposons in two litopenaeid species. Anim Genet 39: 363–373. doi:10.1111/j.1365-2052.2008.01739.x.
32. Rho M, Schaack S, Gao X, Kim S, Lynch M, et al. (2010) LTR retroelements in the genome of Daphnia pulex. BMC Genomics 11: 425. doi:10.1186/1471-2164-11-425.
33. Dalle Nogare DE, Clark MS, Elgar G, Frame IG, Pendar RTM (2002) Xena, a full-length basal retroelement from tetraodontid fish. Mol Biol Evol 19: 247–255.
34. Koyama T, Kondo H, Aoki T, Hirono I (2012) Identification of Two Penelope-Like Elements with Different Structures and Chromosome Localization in Kuruma Shrimp Genome. Marine biotechnology (New York, NY). doi:10.1007/s10126-012-9474-z.
35. Terrat Y, Bernivard E, Higuet D (2008) GalEa retrotransposons from galaxhrid squat lobsters (Decapoda, Anomura) define a new clade of Ty1/copia-like elements restricted to aquatic species. Mol Genet Genomics 279: 63–73. doi:10.1007/s00438-007-0295-0.
36. Terrat Y (2009) Caractérisation de rétrotransposons chez des crustacés décapodes anomoures, les Galathées [S.l.]: [s.n.].
37. Sarradin P-M, Caprais J-C, Rao R, Kerouel R, Aminot A (1999) Chemical environment of the hydrothermal mussel communities in the Lucky Strike and Menez Gwen vent fields, Mid Atlantic Ridge. Cahiers de biologie marine 40: 93–104.
38. Van Dover CL, German CR, Speer KG, Parson LM, Vrijenhoek RC (2002) Evolution and biogeography of deep-sea vent and seep invertebrates. Science 295: 1253–1257. doi:10.1126/science.1067361.
39. Dusak RP, Johnson HP (2002) Hydrothermal systems. Stirring the oceanic incubator. Science 296: 1406–1407. doi:10.1126/science.1072075.
40. Desbruyères D, Segonzac M (1997) Handbook of Deep-sea Hydrothermal Vent Fauna. Editions Quae. 284 p.
41. Pernard J, Cambon-Bonavita M-A, Zbinden M, Lepoint G, Jeanin A, et al. (2012) Inorganic carbon fixation by chemosynthetic ectosymbionts and nutritional transfers to the hydrothermal vent host-shrimp Rimicaris exoculata. ISME J. doi:10.1038/ismej.2012.87.
42. Segonzac M, De Saint-Laurent M, Casanova B (1993) L'énigme du compartiment tegumentaire des crevettes Alvinocarididae des sites hydrothermaux de la dorsale medio-atlantique. Cahiers de biologie marine 34: 535–571.
43. Ravaux J, Gaill F, Le Bris N, Sarradin P-M, Jollivet D, et al. (2003) Heat-shock response and temperature resistance in the deep-sea vent shrimp Rimicaris exoculata. J Exp Biol 206: 2345–2354.
44. Sarrazin J, Sarradin P-M, Allais A-G, Momareto Cruise Participants X (2006) MoMARETO: a cruise dedicated to the spatio-temporal dynamics and the adaptations of hydrothermal vent fauna on the Mid-Atlantic Ridge. InterRidge News 15: 24–33.
45. Gaill F, Ballu V (2007) Cruise MoMARDREAM-Naut and other MoMAR experiments at Rainbow and lucky Strike in summer 2007. InterRidge News 16.
46. Jollivet D, Lallier FH, Barnay A-S (2004) The BIOSPLEIDO cruise: a new survey of hydrothermal vents along the South East Pacific Rise from 7°24'S to 21°53'S. InterRidge News 13: 20–26.
47. Havell AJ, Dunbar L, Anderson R, Pearce SR, Hartley R, et al. (1992) Ty1-copia group retrotransposons are ubiquitous and heterogeneous in higher plants. Nucleic Acids Res 20: 3639–3644.
48. Devic M, Albert S, Delseny M, Roume TJ (1997) Efficient PCR walking on plant genomic DNA. Plant physiology and biochemistry 35: 331–339.
49. Hall T (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symposium Series 41: 95–98.
50. Clark MS, Thorne MAS, Toullec J-Y, Meng Y, Guan LL, et al. (2011) Antarctic krill 454 pyrosequencing reveals chaperone and stress transcriptome. PLoS ONE 6: e15919. doi:10.1371/journal.pone.0015919.
51. Tagmount A, Wang M, Lindquist E, Tanaka Y, Teranishi KS, et al. (2010) The porcelain crab transcriptome and PCAD, the porcelain crab microarray and sequence database. PLoS ONE 5: e9327. doi:10.1371/journal.pone.0009327.
52. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403–410. doi:10.1016/S0022-2836(05)80360-2.
53. Katoh K, Asimenos G, Toh H (2009) Multiple alignment of DNA sequences with MAFFT. Methods Mol Biol 537: 39–64. doi:10.1007/978-1-59745-251-9_3.
54. Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol Biol Evol 24: 1596–1599. doi:10.1093/molbev/msm092.
55. Cracoole A, Gribaldo S (2010) BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. BMC Evol Biol 10: 210. doi:10.1186/1471-2148-10-210.
56. Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4: 406–425.
57. Jones DT, Taylor WR, Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci 8: 275–282.
58. Felsenstein J (1985) Confidence Limits on Phylogenies: An Approach Using the Bootstrap. Evolution 39: 783–791. doi:10.2307/2408678.
59. Goodwin TJD, Poulter RTM (2002) A group of deuterostome Ty3/gypsy-like retrotransposons with Ty1/copia-like pol-domain orders. Mol Genet Genomics 267: 481–491. doi:10.1007/s00438-002-0679-0.
60. Shank TM, Black MB, Halanych KM, Lutz RA, Vrijenhoek RC (1999) Miocene radiation of deep-sea hydrothermal vent shrimp (Caridea: Bresiliidae): evidence from mitochondrial cytochrome oxidase subunit I. Mol Phylogenet Evol 13: 244–254. doi:10.1006/mpev.1999.0642.
61. Macas J, Neumann P, Navrátilová A (2007) Repetitive DNA in the pea (Pisum sativum L.) genome: comprehensive characterization using 454 sequencing and comparison to soybean and Medicago truncatula. BMC Genomics 8: 427. doi:10.1186/1471-2164-8-427.
62. Kelly LJ, Leitch IJ (2011) Exploring giant plant genomes with next-generation sequencing technology. Chromosome Res 19: 939–953. doi:10.1007/s10577-011-9246-z.
63. Piednoel M, Aberer AJ, Schnerweis GM, Macas J, Novak P, et al. (2012) Next-Generation Sequencing Reveals the Impact of Repetitive DNA Across Phylogenetically Closely Related Genomes of Orobanchaceae. Molecular biology and evolution. doi:10.1093/molbev/mss168.
64. Pagán HJT, Macas J, Novák P, McCulloch ES, Stevens RD, et al. (2012) Survey sequencing reveals elevated DNA transposon activity, novel elements, and variation in repetitive landscapes among vesper bats. Genome Biol Evol 4: 575–585. doi:10.1093/gbe/evs038.
65. Teixeira S, Cambon-Bonavita M-A, Serrão EA, Desbruyères D, Arnaud-Haond S (2011) Recent population expansion and connectivity in the hydrothermal shrimp Rimicaris exoculata along the Mid-Atlantic Ridge. Journal of Biogeography 38: 564–574. doi:10.1111/j.1365-2699.2010.02408.x.
66. Teixeira S, Serrão EA, Arnaud-Haond S (2012) Panmixia in a fragmented and unstable environment: the hydrothermal shrimp Rimicaris exoculata disperses extensively along the Mid-Atlantic Ridge. PLoS ONE 7: e38521. doi:10.1371/journal.pone.0038521.
67. Zbinden M, Shillito B, Le Bris N, de Villardi de Montlaur C, Roussel E, et al. (2008) New insights on the metabolic diversity among the epibiotic microbial community of the hydrothermal shrimp Rimicaris exoculata. Journal of Experimental Marine Biology and Ecology 359: 131–140. doi:10.1016/j.jembe.2008.03.009.
68. Ravaux J, Corzin D (2009) Hydrothermal vent shrimps display low expression of the heat-inducible hsp70 gene in nature. MARINE ECOLOGY-PROGRESS SERIES 396: 153–156.
69. Corzin D, Shillito B, Chertemps T, Tanguy A, Léger N, et al. (2010) Identification of differentially expressed genes in the hydrothermal vent shrimp Rimicaris exoculata exposed to heat stress. Mar Genomics 3: 71–78. doi:10.1016/j.margen.2010.05.002.
70. Vinyas DF, Cummings MP, Konieczny A, Azuebel FM, Rodermel SR (1992) copia-like retrotransposons are ubiquitous among plants. PNAS 89: 7124–7128.
71. Huang S-W, Liu Y-Y, You E-M, Liu T-T, Shu H-Y, et al. (2011) Fosmid library end sequencing reveals a rarely known genome structure of marine shrimp Penaeus monodon. BMC Genomics 12: 242. doi:10.1186/1471-2164-12-242.
72. Muszewska A, Hoffman-Sommer M, Grynberg M (2011) LTR retrotransposons in fungi. PLoS ONE 6: e29425. doi:10.1371/journal.pone.0029425.
73. Sun C, Shepard DB, Chong RA, López Arriaza J, Hall K, et al. (2012) LTR retrotransposons contribute to genomic gigantism in plethodontid salamanders. Genome Biol Evol 4: 168–183. doi:10.1093/gbe/evr139.
74. Takeuchi T, Kawashima T, Koyanagi R, Gyoja F, Tanaka M, et al. (2012) Draft genome of the pearl oyster Pinctada fucata: a platform for understanding bivalve biology. DNA Res 19: 117–130. doi:10.1093/dnares/dsr045.
75. Grandbastien M-A, Audeon C, Bonnivard E, Casacuberta JM, Chalhoub B, et al. (2005) Stress activation and genomic impact of Tnt1 retrotransposons in Solanaceae. Cytogenet Genome Res 110: 229–241. doi:10.1159/000084957.
76. Vieira C, Nardon C, Arpin C, Lepetit D, Biémont C (2002) Evolution of genome size in Drosophila. is the invader's genome being invaded by transposable elements? Mol Biol Evol 19: 1154–1161.
77. Oliver KR, Greene WK (2009) Transposable elements: powerful facilitators of evolution. Bioessays 31: 703–714. doi:10.1002/bies.200800219.
78. Oliver KR, Greene WK (2011) Mobile DNA and the TE-Thrust hypothesis: supporting evidence from the primates. Mob DNA 2: 8. doi:10.1186/1759-8753-2-8.
79. Jurka J, Bao W, Kojima KK (2011) Families of transposable elements, population structure and the origin of species. Biol Direct 6: 44. doi:10.1186/1745-6150-6-44.
80. Regier JC, Shultz JW, Zwick A, Hussey A, Ball B, et al. (2010) Arthropod relationships revealed by phylogenomic analysis of nuclear protein-coding sequences. Nature 463: 1079–1083. doi:10.1038/nature08742.
81. Montagné N, Desdevises Y, Soyez D, Toullec J-Y (2010) Molecular evolution of the crustacean hyperglycemic hormone family in ecdysozoans. BMC Evol Biol 10: 62. doi:10.1186/1471-2148-10-62.
82. Dircksen H, Neupert S, Predel R, Verleyen P, Huybrechts J, et al. (2011) Genomics, transcriptomics, and peptidomics of Daphnia pulex neuropeptides and protein hormones. J Proteome Res 10: 4678–4691. doi:10.1021/pr200284e.
83. Giribet G, Edgecombe GD (2012) Reevaluating the arthropod tree of life. Annu Rev Entomol 57: 167–186. doi:10.1146/annurev-ento-120710-100659.
84. Casacuberta JM, Grandbastien MA (1993) Characterisation of LTR sequences involved in the protoplast specific expression of the tobacco Tnt1 retrotransposon. Nucleic Acids Res 21: 2087–2093.

85. Yoshioka K, Kanda H, Takamatsu N, Togashi S, Kondo S, et al. (1992) Efficient amplification of Drosophila simulans copia directed by high-level reverse transcriptase activity associated with copia virus-like particles. Gene 120: 191–196.
86. Ohbayashi F, Shimada T, Sugasaki T, Kawai S, Mita K, et al. (1998) Molecular structure of the copia-like retrotransposable element Yokozuna on the W chromosome of the silkworm, Bombyx mori. Genes Genet Syst 73: 345–352.
87. Peterson-Burch BD, Voytas DF (2002) Genes of the Pseudoviridae (Ty1/copia retrotransposons). Mol Biol Evol 19: 1832–1845.
88. Ruiz CJB, Ranson H, Wang X, Besansky NJ (2002) Structure and evolution of mtanga, a retrotransposon actively expressed on the Y chromosome of the African malaria vector Anopheles gambiae. Mol Biol Evol 19: 149–162.

Des crustacés … aux champignons

Nous avons mis en évidence des dynamiques différentes entre les éléments Gypsy et Copia au sein des crustacés. Les Gypsy sont divers et bien présents au sein des espèces étudiées. Nous en trouvons facilement quand nous les recherchons que ce soit par approche "humide" ou *in silico*. Les Copia sont plus rares au sein des génomes et très difficiles à trouver par l'approche "humide". De plus, ils sont moins diversifiés et assez rares au sein des crustacés : un seul clade d'éléments Copia est représentée au sein des malacostraca: les GalEa. La diversité des éléments et leur nombre de copies au sein des génomes, peuvent jouer sur la facilité ou non de les retrouver au sein des génomes par une méthode par PCR. Un faible nombre de copies des éléments peut rendre la détection plus difficile tandis qu'une grande diversité des éléments peut permettre une détection par amorces dégénérées plus facile. Il existe donc un biais de détection des éléments par approche par PCR.

Les éléments GalEa sont présents chez les crustacés et également chez beaucoup d'autres métazoaires. Il y a eu une très bonne réussite de ces éléments chez les métazoaires, car même si la distribution est morcelée, leur présence est maintenue dans de nombreux taxons, tels que les mollusques et les actinoptérygiens (poissons). Les éléments GalEa sont aussi le groupe majoritaire des Copia au sein des crustacés. Nous pouvons suite à cet article nous poser différentes questions : Les GalEa sont-ils le clade de Copia majoritaire dans d'autres taxons, comme les téléostéens ou les mollusques, ou leur dominance est une caractéristique propre aux crustacés? En effet, on retrouve des GalEa chez des mollusques et nous pourrions étudier la part de ces éléments parmi les Copia au sein des génomes. Nous nous sommes, de plus, rendu compte qu'un autre élément Copia non GalEa était présent chez le crustacé *Paryale hawaiensis*. Cet élément fait partie du clade des Hydra1-2, d'abord décrite chez une espèce de cnidaire et que l'on retrouve aussi au sein d'une espèce de téléostéen *Danio rerio* (Llorens et al., 2009). Nous nous demandons donc si au sein des métazoaires, un tel autre clade de Copia pourrait présenter une réussite identique à celle des GalEa? Les éléments GalEa présentent un modèle de dynamique particulier : le « Domino's Day Spreading » qui associe un faible nombre moyen de copies au sein des espèces mais des «bursts» importants qui peuvent expliquer leur maintien au sein des génomes. Nous nous posons donc la question de savoir si un autre clade de Copia présente le même modèle de dynamique que celui des GalEa ?

Pour répondre à ces questions, nous pensons d'abord étudier la présence des Copia chez

d’autres groupes de métazoaires. D’abord, la distribution dans un groupe de même importance et qui présente une même diversité en terme de nombre d’espèces, de mode de vie, de taille de génome, et l’avantage d’avoir quelques génomes séquencés dans quelques ordre différents, par exemple les mollusques. Et aussi étudier d’autres clades d’éléments Copia tels que les Hydra1-2 au sein des métazoaires.

On pourrait aussi vérifier où les GalEa existent en dehors des métazoaires et s’ils peuvent être majoritaires parmi les Copia. Nous savons déjà que les GalEa sont présents chez les deux espèces de rhodophytes (algues rouges) *Porphyridium cruentum* (Piednoël et al., 2013) et *Porphyra yezoensis* (Peddigari et al., 2008). Nous pourrions étudier la répartition des éléments GalEa au sein du subphylum des rhodophytes, mais les données génomiques disponibles sur les algues rouges restent jusqu’à présent très limitées. Seul 7 génomes (dans 4 ordres d’espèces) sont disponibles, ce qui ne nous permet pas encore de faire d’analyse comparative pour la recherche et l’étude des éléments Copia, notamment des GalEa. Nous ne pouvons donc pas savoir si les GalEa sont les éléments Copia majoritairement représentés chez les algues rouges.

Par la suite, nous nous sommes rendu compte que la distribution des éléments GalEa était incomplète. En effet, lors de la recherche des éléments GalEa en dehors des crustacés, par l’approche *in silico*, nous nous étions penchés uniquement sur les bases de données nucléotidiques. Malheureusement, les données concernant les éléments de champignons sont déposés dans des bases de données protéiques. Ainsi, nous n’avions pas pu trouver d’éléments GalEa au sein des champignons lors de la première recherche. Lors de notre étude des GalEa de rhodophytes, les interrogations dans les bases de données protéiques ont révélé un élément GalEa chez *Metarhizium anisopliae,* un ascomycète. Grâce à ce nouvel élément, nous avons pu réinterroger les bases de données et retrouver des éléments GalEa au sein de nombreuses espèces de champignons, et plus précisément chez les pezizomycotina qui représentent le sous embranchement d'ascomycète majoritaire (environ 90 % des ascomycètes connus) avec plusieurs dizaines de milliers d'espèces décrites. Nous avons donc décidé d’étudier les éléments GalEa dans ce phylum groupe frère des métazoaires. Nous voulions d’abord savoir si le clade GalEa était également le clade majoritaire des Copia chez ces autres espèces et également si leur dynamique semblait être compatible avec le modèle du « Domino’s Days Spreading ».

Les eumycètes, ou champignons, sont un bon taxon pour l'étude des ETs. En effet, ils

présentent plus de 100 000 espèces connues, et une grande variation de taille de génome, entre 10^7 bp et 10^8 bp (plus grande variation que chez les crustacés). Les champignons ont une grande diversité de style de vie et d'habitats. Ce sont principalement des organismes terrestres parasites, symbiotiques, ou saprophytes, mais toujours hétérotrophes. Ils ont envahi la plupart des milieux : eaux, sols, intestins et excréments des herbivores, et des champignons hydrothermaux sont également décrits (Le Calvez et al., 2009). Ils sont fondamentalement absorbotrophes et ne se nourrissent donc que de nutriments présents dans leur environnement, soit fournis par un hôte symbiotique, soit obtenus par digestion extracellulaire par des enzymes lytiques. Il existe également des associations de champignons (principalement des ascomycètes) qui vivent en association avec un organisme photosynthétique pour constituer un lichen.

Les champignons sont des espèces très étudiées car beaucoup sont des pathogènes de plantes et d'espèces cultivées. Leur étude a donc un intérêt agronomique et beaucoup d'espèces ont leurs génomes séquencés, ou en cours de séquençage. Il existe notamment le projet 1000 génomes, en partenariat avec JGI, qui regroupe un maximum de données. Plusieurs études à large échelle ont porté sur la recherche de rétrotransposons au sein des champignons. Muszewska et al., (2013) se sont intéressés à la distribution des rétrotransposons à Tyrosine Recombinase DIRS et Ngaro, et en ont décrit dans de nombreuses espèces, mais de manière morcelée. Muszewska et al., (2011) ont également étudié les rétrotransposons à LTR chez les champignons (Figure 17). Sur les 59 espèces étudiées, 53 présentent des éléments Gypsy et 52 des éléments Copia. On remarque une forte variation de nombre d'éléments entre les espèces. En effet, certains génomes présentent beaucoup d'éléments à LTR (*Postia placenta*), tandis que d'autres paraissent presque vides (genre *Aspergillus*). Cela est dû aux proportions des deux superfamilles qui diffèrent fortement. On retrouve principalement des éléments Gypsy, tandis que les éléments Copia ont l'air plus rare. Dans la plupart des génomes, les éléments Gypsy sont en plus grand nombre, avec jusqu'à 2689 copies (*P. placenta*) contre seulement 274 copies pour les Copia (*Pyrenophora tritici-repentis*). Seules 11 espèces ont une proportion de Copia plus forte que celle des Gypsy. Même si les conclusions restent fiables, les résultats de cette étude doivent tenir compte du biais dans l'échantillonnage. Certaines espèces sont sur-représentées (11 souches de *Coccidioides posadasii*) ainsi que le genre *Aspergillus* (8 espèces). La base des critères de distinction des éléments Gypsy et Copia de cette étude ne permettaient pas de distinguer les différents clades de rétrotransposons.

Nous avons décidé de faire une étude plus complète des éléments GalEa, afin de connaitre leur distribution au sein des champignons, et également une étude plus générale des éléments Copia pour définir si les GalEa représentent le clade majoritaire. Nous disposons d'un grand nombre de génomes complètements séquencés au sein des champignons, contrairement aux crustacés, et avons donc privilégié une approche *in silico*. De ce fait, nous avons pu savoir quelle part du génome les GalEa pouvaient représenter, en étudiant en tout plus de 80 espèces de pezizomycotina possédant des GalEa dont une quarantaine ont un génome complètement séquencé.

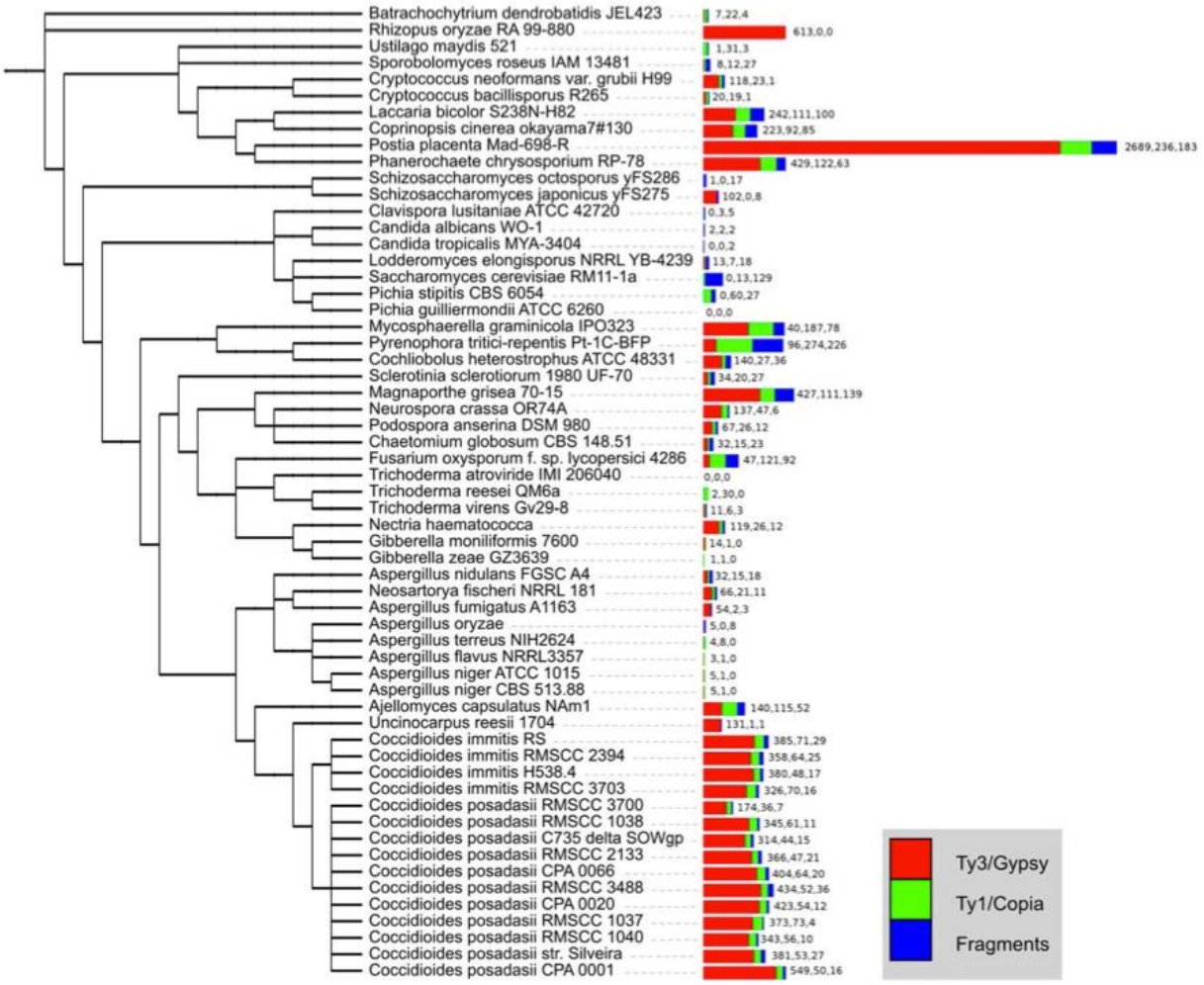


Figure 17: Nombre de rétrotransposons à LTR par génome au sein des champignons. La longueur de la barre est proportionnelle au nombre de rétrotransposons détectés, les nombres de copies potentiellement encore active de chaque superfamille d'éléments sont indiqués dans l'ordre à côté de chaque barre. Le nombre de fragments (éléments non définis) est également noté.

Chapitre IV. Etude de la distribution des GalEa au sein des champignons (eumycètes).

Article

GalEa retrotransposons in Fungi

Tifenn DONNART[1], Florian MAUMUS[2], Mathieu PIEDNOEL[1,3], Dominique HIGUET[1], Éric BONNIVARD[1,*]

[1] Institut de Biologie Paris-Seine, UMR7138 "Evolution Paris-Seine" UPMC-CNRS, Equipe : Eucaryotic genome evolution, Bât A, 4ème étage, porte 414, Case 5, 7 quai Saint Bernard, 75252 Paris Cedex 05

[2] Unité de Recherches en Génomique-Info (UR INRA 1164), INRA, Centre de recherche de Versailles, bat.18, RD10, Route de Saint Cyr, 78026 Versailles Cedex, France

[3] Max Planck Institute for Plant Breeding Research**,** Köln, Germany

[*]Corresponding author

Email adresses :
TD : tifenn.donnart@gmail.com
FM : fmaumus@gmail.com
MP : piednoel@closun.snv.jussieu.fr
DH : dominique.higuet@upmc.fr
EB : eric.bonnivard@upmc.fr

Introduction

Transposable elements (TEs) have been identified in all eukaryotic species investigated so far where they can constitute a large fraction of the genome (Wicker et al., 2007). Because of their important effects on genome structure and stability, they are considered as one of the major sources of genetic variability in eukaryotes (Biémont and Vieira, 2006; Fedoroff, 1999; Finnegan, 2012; Kazazian, 2004). TEs show a very large diversity in terms of structural features, sequences and replication mechanisms (Eickbush and Jamburuthugoda, 2008; Wicker et al., 2007), which greatly impact their dynamics in the genomes. For example, while LTR (Long Terminal Repeats) elements make up the largest proportion of plant genomes, they are less predominant in animals. The host distribution and their abundance among genomes thus greatly depend on both the element type and the taxon studied. Contrary to the other TEs, retrotransposons replicate via a "copy and paste" mechanism, which relies on an RNA intermediate and are subdivided in five major orders: LTR retrotransposons, LINEs (Long INterspersed Elements), SINEs (Short INterspersed Elements), Penelope and YR (tyrosine recombinase encoding) elements. The LINEs, SINES and LTR retrotransposons have been detected almost ubiquitously. In contrast, the Penelope are widely distributed among animal species, but seem to be rare among plants, protists and fungi (Arkhipova, 2006) and the YR retrotransposons (e.g. DIRS1-like elements) are less frequent with a patchy distribution in unikont species (Piednoël and Bonnivard, 2009; Piednoël et al., 2011). At a finer scale, one can distinguish the distribution of the different superfamilies that make up these orders. In particular, it has been shown that the three LTR retrotransposon superfamilies (BEL/Pao, Copia and Gypsy) have uneven relative abundances among eukaryotes (Llorens et al., 2009; Piednoël et al., 2011). The Gypsy and Copia elements are widely distributed among the genomes of plants, fungi and animals. Gypsy elements are clearly most abundant when Copia elements are absent in a wide range of species. Therefore, the BEL/Pao elements appear often more abundant in metazoans than Copia retrotransposons (de la Chaux and Wagner, 2011).

In fungi, first TEs have been described in yeast (Cameron et al., 1979). The presence of Gypsy and Copia retrotransposons was later reported in filamentous fungi, with Gypsy being the most abundant (Daboussi and Capy, 2003; Dobinson and Hamer, 1993; Muszewska et al., 2011). As these elements require a multi-compound machinery to be mobile, they easily become non-autonomous, with the presence of numerous truncated elements or of traces in

genomes. So, although some retrotransposons appear still functional (e.g. the Gypsy element MAGGY from *Magnaporthe grisea;* Nakayashiki et al., 1999), most of the detected copies harbor many stop codons or frameshift in coding regions. This effect is reinforce by divers silencing mechanisms developed by fungi to limit the expansion of TEs, especially the RIP (Repeat-Induced Point mutation), a homology-based process by which repetitive DNA accumulates GC to AT transition mutations (Clutterbuck, 2011; Cuomo et al., 2007; Idnurm and Howlett, 2003, Graïa et al., 2001). When genomic studies of LTR retrotransposons usually concern one or few species (Amselem et al., 2011; Amyotte et al., 2012, Santana et al., 2014), a large-scale analysis on 45 diverse assembled genomes was recently reported (Muszewska et al., 2011). It reveals that LTR retrotransposon contents are very variable (from 0 to 2689 copies for Gypsy and from 0 to 274 copies for Copia), even between closely related species. Their expansions in fungi usually involve both an increase in copy number of individual elements and an increase in the number of different elements. Copia retrotransposons are often poorly represented as 6 species appear devoid of Copia elements and 19 harbor less than 10 copies.

In the present study, we will only focus on the Copia retrotransposons. These elements usually encode two open reading frames: the *gag*, which encodes proteins that form the virus-like particles, and the *pol*, which encodes various enzymatic activities like an aspartic protease (PR), a reverse transcriptase (RT), a RNase H (RH) and a DDE-type integrase (INT) that are involved in the transposition mechanism. As every LTR retrotransposons, Copia elements are flanked by two direct LTRs (usually between 100 and 500 bp in length), which encompass the promoter and regulatory regions. These two LTRs are supposed to be identical when the element inserts into the genome. Phylogenetic analyses of Copia retrotransposons have revealed several families of elements regrouped into two major branches (Llorens et al., 2009, 2011). While the branch 2 is highly diverse and widely distributed among eukaryotes (5 clades among arthropods, 4 clades among land plants, as well as clades related to fungi, red or green algae, and other metazoans), the Branch 1 harbors a low diversity and seems restricted to few hosts. This last branch encompasses the original *Pseudovirus* elements (*Ty* retrotransposons normally found in fungi) together with four clades of CoDi-like elements from diatoms and the GalEa clade. The first GalEa element, GalEa1, has been initially described in galatheids (**Gal**atheid ***E****uminida* ***a****nnulosa*, Terrat et al., 2008), and few closely related elements have been identified in teleosts (Zeco1 from *Danio rerio* and Olco1 from *Oryzias latipes*) and a urochordate (Cico1 from *Ciona intestinalis*). Numerous GalEa

sequences have been also identified from some environmental samples of micro-organisms collected during Sargasso Sea surveys (Maumus et al., 2009). However, it remained impossible to determine which organisms they originated from. Subsequent studies confirmed the presence of GalEa elements in Rhodophyta genomes, with PyRE10G element identified in the red algae *Porphyra yezoensis* (Peddigari et al., 2008) and two elements (CoPorcru1 and CoPorcru2) in *Porphyridium cruentum* (Piednoël et al., 2013). We also recently revealed that these elements have been actually more successful among metazoan species as previously thought with some elements also identified in Mollusca, Chordata, Cnidaria, Ctenophora, Echinoderma and Hemichordata (Piednoël et al., 2013). The presence of GalEa elements in two really distant eukaryote clades suggest that these elements are ancient, but their hypothesized absence in Chlorophyta, in Streptophyta (plantae) or fungi as well as the low number of Rhodophyta harboring such elements may raise the question of possible horizontal transfers.

In the present study, we find evidence of the presence of GalEa elements in several Ascomycota. We decided to take advantage of the exploding amount of genomic data (1000 fungal genomes project; http://1000.fungalgenomes.org/home/) to carry out the first large-scale GalEa retrotransposons research in fungi. We revealed their wide distribution in Pezizomycotina, a particular subphylum of Ascomycota. At last, our previous analyses have shown that GalEa elements were highly predominant in Malacostraca in comparison with the other clades of Copia (Piednoël et al., 2013). We thus analyzed the relative abundance of the different clades of Copia within fungi to test whether this particular pattern is widely shared among eukaryotes or specific to Malacostraca.

Methods

Distribution of GalEa elements in fungi

To determine which species of fungi harbor GalEa elements, we performed tBLASTn analyses on all genomes available on the fungal genomics resource from MycoCosm (http://genome.jgi-psf.org/programs/fungi/index.jsf, release August 2014) as well as on all genomic or transcriptomic databases available on the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov). Elements from three different phyla have been used as queries: one element previously identified in Metazoa (GalEa1, DQ913005.1) and two

element from fungi (CoMetani, n$) and from Rhodophyta (CoGrala, n$). To discriminate potential identified sequences that belong to other clades of Copia, we also used the Copia element from *Drosophila melanogaster* (X02599.1) as a query. When possible, chimeric sequences of the newly identified GalEa elements from the transcriptomic or unassembled genomic data have been designed and labeled as reference elements (Supdata1).

To check the GalEa clade affiliation of these newly identified elements, we used two complementary approaches: sequences covering the RT/RH domains were included into phylogenic analyses whereas the remaining sequences were classified using similarity searches using BLAST on the Gypsy Database 2.0. In the latter case, an element was assigned to the GalEa clade under the two conditions: (i) the five best hits must correspond to the five GalEa1-like elements referenced in the database; and (ii) the difference between the best E-values obtained with GalEa-like and other reference elements must be greater than 1e-10.

Relative genomic abundance of GalEa elements in fungi genomes

Forty two assembled genomes that possess GalEa elements were downloaded from the DOE Joint Genome Institute and the Broad Institute of MIT and Harvard databases, to estimate both the number of copy of GalEa elements and the proportion of GalEa elements among Copia retrotransposons. For that purpose, we first isolate all potential retrotransposon sequences bounded by two conserved LTRs using LTR Harvest (Ellinghaus et al., 2008) using the following parameters: LTR length comprised between 100 and 1000 bp, distance between two LTRs ranging from 3000 and 11000 bp and similarity between two LTRs greater than 80 %. To discriminate the GalEa elements from the other Copia sequences, we performed a blastx on a custom database composed of RT/RH domains from 122 Pao/BEL, 116 Gypsy, 97 GalEa (including elements from metazoan and Rhodophyta (Piednoël et al., 2013), as well as GalEa from fungi previously identified) and 67 other Copia retrotransposons (for example: Tork-like, CoDi-like, Oryco-like and 1731-like elements).

The newly identified Copia copies may be corrupted by insertion of various sequences such as microsatellites, MITEs, transposons or other retrotransposons. Muszewska et al., (2011) have described several complex transposons of fungi, which architecture encompass a mix of retrotransposons and transposons domains bordered by conserved LTR. Such sequences may lead to bias in the estimation of abundance of Copia elements among the genomes. We thus clustered, species by species, the newly identified copies into families using BLASTclust

(http://toolkit.tuebingen.mpg.de/blastclust) and then aligned the sequences from the same family (MAFFT, http://mafft.cbrc.jp/alignment/server/) to remove all unshared sequence larger than 20bp. The orphan sequences were individually analyzed using BLAST to confirm that the overall correspond to a Copia element and to eliminate potential other transposable element insertions.

Finally, the genomes characterized as harboring at least one GalEa element have been screened using RepeatMasker (options -nolow -no_is -pa 8 –frag 380000 -div 20) to recover all Copia related sequences, including those not detected with LTR Harvest. For that purpose we used a custom database (one for each sepcies) comprising all curated Copia sequences. If no GalEa element is detected with LTR Harvest, then the database only includes the sequences previously obtained by tBLASTn analyses.

Phylogenetic analyses

We used BLASTx (e-15) and tools from SMart (Zytnicki and Quesneville, 2011) to retrieve RT/RH domains of at least 200 amino acid from the GalEa sequences identified in fungi. 43 copies representative of the diversity in GalEa identified in fungi (data not shown) were included in some phylogenetic analyzes performed according to Piednoel, Donnart et al. (2013). As ripped sequences (AT-content >70%) are strongly corrupted, their protein sequences were manually constructed using the DNA to protein tool at http://bio.lundberg.gu.se/edu/translat.html. Reference element sequences that were used in phylogenetic analyses correspond to GalEa retrotransposons previously deposited (Piednoël et al., 2013; Terrat et al., 2008) and to Copia sequences that could be accessed in GenBank.

In deep characterization of GalEa elements of fungi

To highlight the structural variations of GalEa elements in fungi, we selected 7 elements obtained from LTRHarvest that are distant within the CoGaFu clade. When several copies of an element were available, multiple alignments of DNA were constructed using MAFFT (Katoh et al., 2009) and manually curated using BioEdit. The boundaries of the LTRs were manually analyzed, as well as element length and amino acid sequences of conserved domains. The presence of ORFs was determined with the ORF Finder tool (http://www.ncbi.nlm.nih.gov/gorf/gorf.html) and putative PPT with LTR_finder (http://tlife.fudan.edu.cn/ltr_finder/). GalEa elements from fungi were compared with those of

metazoans previously annotated in Terrat et al. (2008) and Piednoël et al. (2013). Analyses of the nucleic acid folding and hybridization predictions on the CHS were performed on the Mfold web server (Zuker, 2003; http://mfold.rit.albany.edu/?q=mfold/DNA-Folding-Form).

Results

Distribution of GalEa elements in fungi

According to tBLASTn analyses, the presence of GalEa elements in Fungi seems restricted to one group of Ascomycota: the subphylum of Pezizomycotina (Fig 1). More precisely, GalEa elements were found in four classes of Leotiomyceta: 10 species of Eurotiomycetes, 14 of Leotiomycetes, 34 of Sordariomycetes and 17 of Dothideomycetes; for a total of 75 species. Considering only the species with assembled genomes, 51 harbor GalEa elements, corresponding to 28% of the 182 genomes tested for these four classes (see Supdata 2 for details). Only 14 genomes are available for the 4 remaining classes of Pezizomycotina; too few to give any reliable conclusion. On the contrary numerous and diverse Saccharromycotina and Basidiomycota species are sequenced (31 and 132, respectively), but all appear devoid of GalEa retrotransposons.

Of course the distribution at small scale is greatly influenced by the choice of the species nominated in genome projects, mainly centered on fungi of interest, like pathogenic species. However the large number of genomes presented reveals that GalEa retrotransposons have a patchy distribution within each class. In Leotiomycetes and Sordariomycetes, we found these elements in almost all orders tested (Fig 1 and Supdata 2), while in Eurotiomycetes and Dothideomycetes they are not detected in two well sampled orders (Onygenales and Botryosphaeriales, respectively). Generally, within an order only some species have GalEa elements; but they appear well represented in few orders such as Helotiales, Glomerellales or Xylariales; being detected in various families and genus. At last, distribution may also be variable within an order. For example considering Eurotiales, GalEa are observed in only 3 out of 19 *Aspergillus* species and 2 out of 13 *Penicillium* species, whereas all the 3 available *Talaromyces* genomes harbor GalEa retrotransposons.

Abundance of GalEa elements in fungi genomes

Copia retrotransposons (both GalEa and other Copia elements) were screened in 43

completely sequenced genomes wherein GalEa elements were detected. In 9 of them we have only detected one or two very short deleted sequences (Supdata 2), therefore these genomes were removed from the analysis. In the 34 remaining species the GalEa amount was estimated in different ways (Table 1).

Class name	Order name	Species	Genome size (Mb)	Sequences obtained with LTR Harvest			Sequences obtained with RepeatMasker					
							Number of sequences >3 Kb			Portion of the genome (%)		
				GalEa	other Copia	GalEa/Copia	GalEa	other Copia	GalEa/Copia	GalEa*	other Copia	GalEa/Copia
Dothideomycetes	Capnodiales	*Aulographum hederae*	31.98	1	1	0,50	2	1	0,67	0,22	0,16	0,58
	Pleosporales	*Amniculicola lignicola*	49.57	77	94	0,45	122	150	0,45	2,91	2,60	0,53
	Pleosporales	*Lentithecium fluviatile*	54.69	18	22	0,45	37	94	0,28	0,69	1,43	0,33
	Pleosporales	*Pyrenophora teres*	33.58	0	0		1	ne		0,15	ne	
	Pleosporales	*Pyrenophora tritici-repentis*	37.84	77	128	0,38	92	278	0,25	1,99	6,35	0,24
	Pleosporales	*Sporormia fimetaria*	25.90	0	3	0,00	2	0		0,10	0,11	0,47
	incertae sedis	*Cenococcum geophilum*	177.57	1262	386	0,77	4317	1336	0,76	16,76	5,78	0,74
Eurotiomycetes	Eurotiales	*Neosartorya fischeri*	32.55	1	11	0,08	1	13	0,07	0,03	0,35	0,08
	Eurotiales	*Talaromyces aculeatus*	37.27	12	6	0,67	19	7	0,73	0,46	0,20	0,70
	Eurotiales	*Talaromyces mameffei*	28.64	19	32	0,37	16	36	0,31	0,44	0,98	0,31
	Eurotiales	*Talaromyces stipitatus*	35.69	40	44	0,48	51	52	0,50	1,27	1,51	0,46
Leotiomycetes	Erysiphales	*Erysiphe pisi*	49.38	13	40	0,25	23	60	0,28	0,48	1,09	0,31
	Helotiales	*Botryotinia fuckeliana*	42.66	3	2	0,60	6	1	0,86	0,10	0,02	0,80
	Helotiales	*Cadophora sp*	70.46	34	54	0,39	123	284	0,30	1,45	3,29	0,31
	Helotiales	*Chalara longipes*	52.43	4	8	0,33	11	21	0,34	0,13	0,35	0,28
	Helotiales	*Meliniomyces bicolor*	82.38	144	29	0,83	603	106	0,85	5,62	1,06	0,84
	Helotiales	*Meliniomyces variabilis*	55.86	3	21	0,13	5	56	0,08	0,15	0,80	0,16
	Helotiales	*Sclerotinia sclerotiorum*	38.33	1	18	0,05	1	20	0,05	0,10	0,81	0,11
	incertae sedis	*Oidiodendron maius*	46.43	11	53	0,17	17	118	0,13	0,25	1,38	0,15
Sordariomycetes	Glomerellales	*Colletotrichum higginsianum*	49.08	0	0		1	ne		0,16	ne	
	Glomerellales	*Colletotrichum graminicola*	51.60	19	35	0,35	151	53	0,74	2,52	0,95	0,73
	Glomerellales	*Colletotrichum acutata*	50.04	0	1	0,00	0	1	0,00	0,06	0,01	0,82
	Glomerellales	*Verticillium albo-atrum*	32.83	0	3	0,00	1	3	0,25	0,01	0,06	0,13
	Glomerellales	*Verticillium dahliae*	33.83	1	24	0,04	1	24	0,04	0,20	0,50	0,28
	Hypocreales	*Cordyceps militaris*	32.27	13	3	0,81	46	9	0,84	3,71	1,05	0,78
	Hypocreales	*Metarhizium robertsii*	39.14	0	3	0,00	2	3	0,40	0,03	0,14	0,17
	Magnaporthales	*Gaeumannomyces graminis*	43.62	0	7	0,00	0	50		0,06	1,18	0,05
	Magnaporthales	*Magnaporthe poae*	39.50	0	2	0,00	0	1		0,05	0,07	0,41
	Ophiostomatales	*Ophiostoma piceae*	32.84	0	0		1	ne		0,03	ne	
	Sordariales	*Chaetomium globosum*	34.89	5	13	0,28	7	15	0,32	0,13	0,79	0,14
	Xylariales	*Anthostoma avocetta*	56.23	0	12	0,00	7	96		0,12	1,41	0,08
	Xylariales	*Daldinia eschscholzii*	37.55	0	0		2	ne		0,12	ne	
	Xylariales	*Hypoxylon sp. CO27-5*	46.59	1	0	1,00	4	ne		0,09	ne	
	Xylariales	*Hypoxylon sp. EC38*	47.30	0	1	0,00	1	1	0,50	0,03	0,03	0,47

Table 1. Copia retrotransposon detection in Fungi genomes. For each species, the number of sequences detected using LTR Harvest and the number of sequences larger than 3kb obtained with the RepeatMasker program are given. The percentage of genome covered by GalEa or other Copia elements is estimated from all sequences detected with RepeatMasker.

ne: not estimated.

We first considered the potentially well-conserved retrotransposons that still possess 2 LTRs and so are recognizable by the LTRharvest software. We thus identified 1669 sequences of GalEa elements and 925 sequences of other Copia elements. Considering the repartition of the GalEa elements detected, the copy number per genome is usually relatively low, with two-third of the species harboring fewer than 5 copies. GalEa elements could not be detected in 12 of them, although the genomes have been chosen because they harbor GalEa sequences. This is probably because these species contain only few copies, which LTRs are altered. Indeed, previous BLAST analyses always reveal a small number of deleted sequences in these species (data not shown). Seven species show between 10 and 20 copies and only 6 species harbor more copies (34, 40, 77, 77 and 144), with a particularly huge number of copies for *Cenococcum geophilum* (1262). This suggests that the recent element activity is relatively low, resulting either from the inactivation of most genomic copies or from a strong regulation of the copy number. The loss of elements in some Pezizomycotina classes or orders (Figure 1) could be facilitated by this low copy number. However, the relatively low copy number observed in genomes has to be regarded with precautions since only well-conserved copies are considered based on the preservation of their LTRs. The copy number per genome appears highly variable, even within some of the orders or genus exanimated. In Helotiales, the high number of copies detected in *Meliniomyces bicolor* (144 copies) contrasts with the few copies identified in other species. Likewise, the two species *Pyrenophora teres* and *Pyrenophora tritici-repentis* have highly variable number of copies (0 and 77, respectively); and this is also true to a lesser extent for the three species of the genus *Colletotrichum.*

The number of GalEa copies was compared to that of other Copia retrotransposons. Among the species that harbor less than five GalEa copies, 15 also harbor few other Copia elements, when the 6 remaining species show between 11 and 24 copies. Considering the species that harbor between 11 and 40 GalEa copies, 2 harbor the same quantity of other Copia, 5 harbor significantly more other Copia and *Cordyceps militaris* and *Talaromyces aculeatus* possesses a majority of GalEa elements. At last, for the four species that harbor a high number of GalEa, *Pyrenophora tritici-repentis* has two times more other Copia elements, *Amniculicola lignicola* possess the same amount of both, whereas GalEa are plainly the predominant Copia

elements in *C. geophilum* and *M. bicolor*. Overall, GalEa and other Copia show an equivalent amount in 18 species, other Copia appear predominant in 12 species and GalEa elements are more numerous in only 4 species.

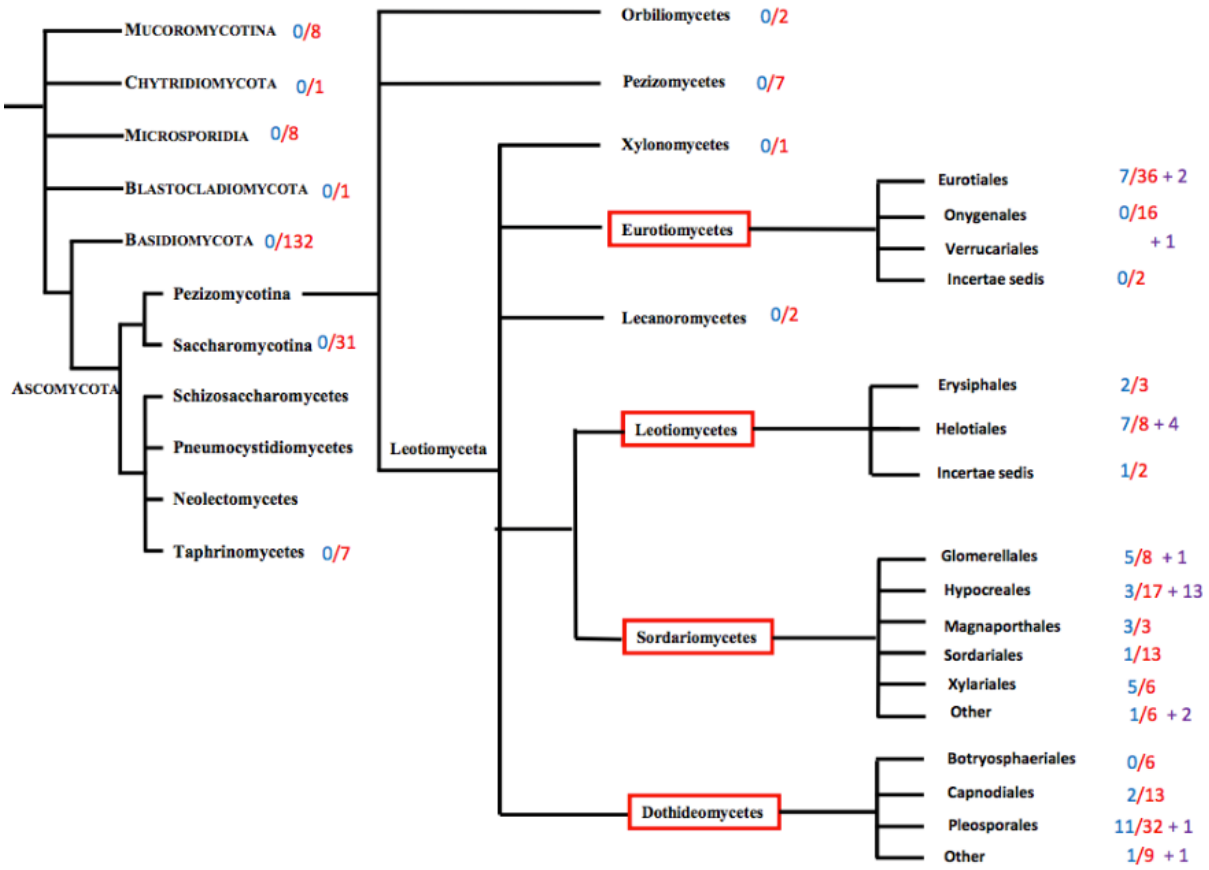


Figure 1. Distribution of GalEa elements among the Fungi groups tested. Species phylogeny is redrawn from MycoCosm. The four Pezizomycotina classes in which GalEa elements were detected are framed in red. For each group, we indicate the number of assembled genomes harboring GalEa sequences (in blue) compared to the total number of genomes analyzed (in red), as well as the number of other Fungi species that also have GalEa retrotransposons according to BLAST analyses (in purple). See Supdata 1 for species details.

At a second step, to get another estimation of the copy number which overcomes LTR sequences conservation; we looked at the Copia sequences larger than 3kb (the smallest size obtained with LTRharvest) detected with RepeatMasker. We identified such sequences in 9 on the 12 species in which no GalEa element could previously be detected using LTRharvest. The three remaining species only harbor some short copies (16 copies about 1500 bp in *Gaeumannomyces graminis*, 14 copies between 1200 and 2500 bp in *Colletotrichum acutata*

and 13 copies between 800 and 2500 bp in *Magnaporthe poae*). In most species the amount of GalEa sequences larger than 3kb is, as expected, a little higher than the number of copies estimated using LTRharvest; but the two estimations remain substantially the same. In very few cases, this amount appears surprisingly smaller, because of stretches of Ns in sequences. However the results greatly differ for five species for which the copy number is increased by a factor 3 to 8 (*C. geophilum*, *Cadophora sp*, *M. bicolor*, *Colletotrichum graminicola* and *C. militaris*). Considering the proportion of GalEa among Copia copies, even if the values slightly change, the ratios remain quite identical for most species whatever the estimation parameter used. However, the ratio significantly differs in *C. graminicola* for which GalEa retrotransposons appear with RepeatMasker strongly predominant. The same result is observed to a lesser extent in *Botryotinia fuckeliana*.

These results were confirmed when we considered the proportion of the genome occupied by Copia retrotransposons estimated with RepeatMasker. Most of the species harbor a small percentage of Copia in their genome (less than 1.5 % for both GalEa and other Copia elements). This proportion appears relatively high (between 3.5 to 22.5% of the genome) in seven species. This can result from a high quantity of both types of elements (*Amniculicola lignicola*) or from a high proportion of other Copia retrotransposons (*Pyrenophora tritici-repentis* and *Cadophora sp*). On the 6 species in which the GalEa elements are a majority, 4 show a high amount of these elements in their genome. In *C. graminicola*, *C. militaris* and *M. bicolor*, in which GalEa elements make up 2.5 %, 3.7 % and 5.6 % of the genome, respectively, the genomic amount of other Copia is less than 1 %. At last, *C. geophilum* harbor not only a large proportion of other Copia retrotransposons (5.5%), but mainly a huge proportion of GalEa elements (more than 16 %).

Phylogenetic analysis of GalEa retrotransposons from fungi

To infer the relationships among the various GalEa elements from fungi and estimate their diversity, we performed phylogenetic analyses based on alignment of amino acid RT/RH domain sequences. The phylogenetic tree (Figure 2) comprises 94 GalEa sequences, including 42 elements from 27 diverse species of Fungi, 14 elements from 4 rhodophyta species and 38 elements from metazoans (representative of 33 species of crustaceans, chordates, cnidarians, mollusks, fishes…). We also included 25 Copia reference retrotransposons representative of each group previously described and 2 Gypsy elements used as outgroups. Adding elements

from fungi do not affect the overall structure of Copia phylogenetic tree and the monophyly of GalEa-like retrotransposons remains supported. Elements from metazoans cluster in a single clade. In contrast, elements from of the rhodophyta species split into several clades for which the phylogenetic relationships remain poorly resolved. At last, all the elements from fungi cluster together into the highly supported monophyletic group that we called CoGaFu. It is interesting to note that RIPped sequences from fungi group in the CoGaFu clade and do not have a great effect on the bootstrap values (Supdata 3). Indeed, when both sequences are available, the original and the RIPped copies of an element group together in the phylogeny (e.g. CoVerda 1 & 2 from *Verticillium dahliae* or CoOima1 & 2 from *Oidiodendron malus*). The phylogenetic relationships within the CoGaFu clade remains poorly resolved. It is thus not possible to test whether the elements phylogeny mirrors the host phylogeny. However, it looks that species from a same genus may share very close sequences (e.g. CoPyte1 and CoPytri3 from *Pyrenophora teres* and *P. tritici-repentis*; or CoTama1 and CoTasti1 & 3 from *Talaromyces marneffei* and *T. stipitatus*).

In-depth characterization of GalEa elements of Fungi

To describe the characteristics of retrotransposons of the CoGaFu clade, we detailed the structure of 7 elements that represent different fungi orders. Major features of the 5' and 3' parts of the sequences are described (data not shown). Moreover, several characteristics, such as their length, the structure of their LTRs and amino acid sequences of conserved domains are compared to those of 6 metazoan GalEa elements in Supdata 4.

The length of the 7 GalEa retrotransposons ranges from 5640 bp (CoBlugra1 from *B. graminis*) to 6485 bp (CoTasti2 from *T. stipitatus*), with an average length of 5985 bp. GalEa elements from fungi thus appear larger than those from previously described in metazoans (the largest to date, CoRex1 is 4750 bp long). As observed for Zeco1 from *D. rerio*, fungi GalEa elements induce a 5bp Target Site Duplication (data not shown). As part of GalEa retrotransposons, the elements from fungi share numerous features with the elements from metazoans, such as LTRs bordered by 5'-TGT and 3'-CA, with an average size of 200 bp (from 163 to 254 for CoGaFu); a large single ORF (sometimes corrupted by stop codon or frameshift according to the copy considered); the HHCC and DD(35)E signatures of the integrase amino acid sequence; the ADxxTK motif at the end of the RH domain; and a great variability in the putative PolyPurine Tract signal.

The major features that distinguish Galea elements from fungi concern the Primer Binding Site (PBS). Previous studies charaterized the structure of GalEa's PBS as a conserved TGGTAGCAGAGC sequence, complementary to the 3' end region of *D. melanogaster* $tRNA^{Met}$ gene, located just after the end of the 5' LTR (Terrat et al. 2008). Such a sequence is absent in all fungi elements. One can however identify at this position a strictly conserved 9bp sequence CTGATCAGT (Figure 3A). Including the last A nucleotide of the 5' LTR to this motif forms a palindromic sequences. Interestingly, the end of the palindromic sequence is always complementary to a 6bp motif (ACAGAT) located within the next 50 bp (Figure 3A), which allow to form a hairpin structure (Figure 3B). We propose to call this particular feature the Conserved Hairpin Site (CHS). Moreover, it seems that for 5 of 8 elements analyzed their CHS may be the extremity of a more complex secondary structure with a larger hairpin that show a bulge or an internal loop located at the LTR-CHS junction (data not shown).

At last, the GalEa elements from fungi also slighty differ from the other GalEa retrotransposons on various conserved motifs. The zinc-finger motif (C(2)C(4)C(4)H) in the *gag* region for example is characteristic of the fungi elements (C(2)C(4)H(4)C in metazoans). In the RT domain, fungi elements harbor two KSRLVI and QTDD motifs (instead of the KARLVA and YVDD usual motifs, respectively), that frame a highly conserved DITQAY motif sequence. More contrastingly, the usual TRPDI motif at the beginning of the RH domain is replaced by a CQPEA conserved motif.

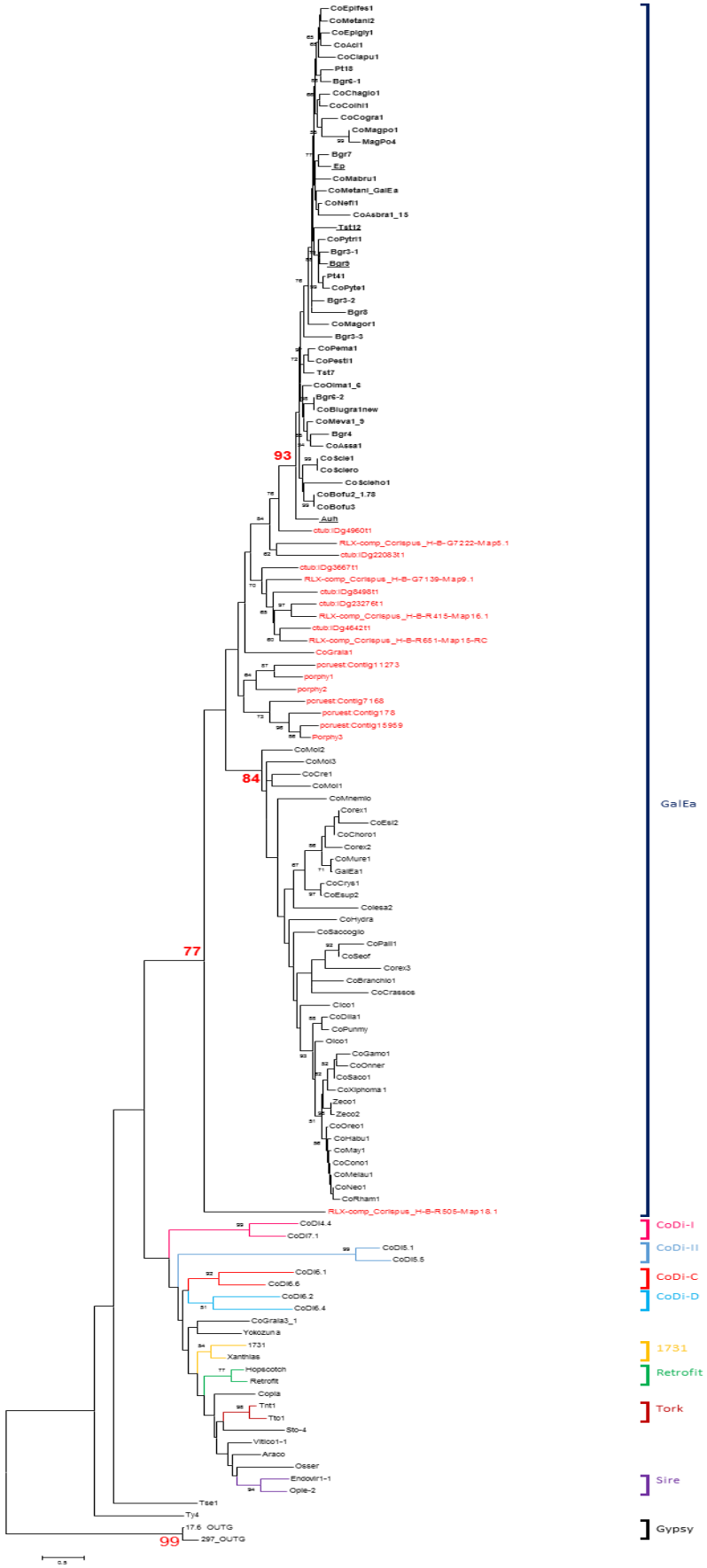
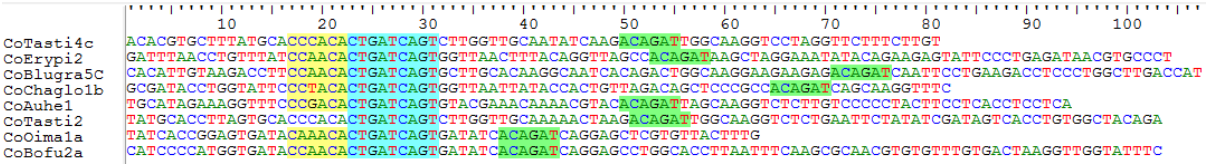

CoEpifes1
CoMetani2
CoEpigly1
CoAcil1
CoClapu1
Pt18
Bgr6-1
CoChaglo1
CoColihi1
CoCogra1
CoMagpo1
MagPo4
Bgr7
Ep
CoMabru1
CoMetani_GalEa
CoNefi1
CoAsbra1_15
Tst12
CoPytri1
Bgr3-1
Bgr9
Pt41
CoPyte1
Bgr3-2
Bgr8
CoMagor1
Bgr3-3
CoPema1
CoPesti1
Tst7
CoDima1_6
Bgr6-2
CoBlugra1new
CoMeva1_9
Bgr4
CoAssa1
93
CoScle1
CoSclero
CoScleho1
CoBofu2_1.78
CoBofu3
Auh
ctub:IDg4960t1
RLX-comp_Corispus_H-B-G7222-Map5.1
ctub:IDg22083t1
ctub:IDg3667t1
RLX-comp_Corispus_H-B-G7139-Map9.1
ctub:IDg8498t1
ctub:IDg23276t1
RLX-comp_Corispus_H-B-R415-Map16.1
ctub:IDg464211
RLX-comp_Corispus_H-B-R651-Map15-RC
CoGraila1
pcruest:Contig11273
porphy1
porphy2
pcruest:Contig7168
pcruest:Contig178
pcruest:Contig15959
Porphy3
CoMol2
CoMol3
84
CoCre1
CoMol1
CoMnemio
Corex1
CoEsi2
CoChoro1
Corex2
CoMure1
GalEa1
CoCrys1
CoEsup2
Colesa2
CoHydra
CoSaccoglo
77
CoPalli1
CoSeor
Corex3
CoBranchio1
CoCrassos
Cloco1
CoDila1
CoPunmy
Oloco1
CoGamo1
CoOnner
CoSaco1
CoXiphoma1
Zeco1
Zeco2
CoOreo1
CoHabu1
CoMay1
CoCono1
CoMelau1
CoNeo1
CoRham1
RLX-comp_Corispus_H-B-R505-Map18.1
CoDi4.4
CoDi7.1
CoDi5.1
CoDi5.5
CoDi6.1
CoDi6.6
CoDi6.2
CoDi6.4
CoGraila3_1
Yokozuna
1731
Xanthias
Hopscotch
Retrofit
Copia
Tnt1
Tto1
Sto-4
Vitloc1-1
Araco
Oeser
Endovir1-1
Ogle-2
Tse1
Ty4
17.6_OUTG
99
297_OUTG
GalEa
CoDi-I
CoDi-II
CoDi-C
CoDi-D
1731
Retrofit
Tork
Sire
Gypsy

Figure 2. Phylogenetic relationships among Copia retrotransposons inferred from Neighbor-Joining analysis of RT/RH amino acid sequences. GalEa elements are colored according to their host: Fungi in blue, rhodophyta in Red and metazoan in green. Statistical support (>50%) comes from non-parametric bootstrapping using 100 replicates. Gypsy retrotransposon sequences were used as outgroup.

A



B

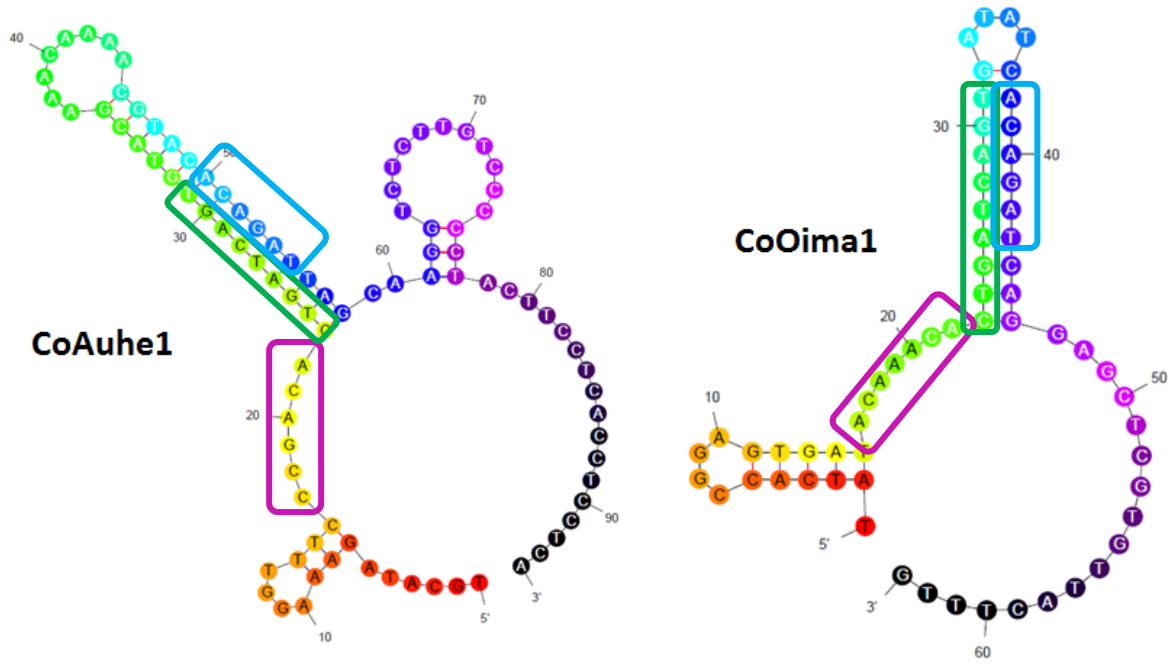


Figure 3. The Conserved Hairpin Site of GalEa retrotransposons of Fungi. (A) Sequences of the CHS of seven full-length elements from *Talaromyces stipitatus* (CoTasti3), *Chaetomium globosum* (CoChaglo1), Botryotinia fuckeliana (CoBofu2), Oidiodendron malus (CoOima1), *Aulographum hederae* (CoAuhe1), *Erysiphe pisi* (CoErypi2) and *Blumeria graminis* (CoBlugra1). (B) Hairpin secondary structure of the CHS. Highlighted element features encompass the end of the LTRs (in pink), the conserved palindromic motif of the CHS (in blue) and the complementary conserved sequence (in green).

Discussion

GalEa retrotransposons were first described in decapods but are widely distributed among metazoans. Study of Copia elements in crustaceans (Piednoël et al. 2013) reveals two features: (1) they seem relatively rare, especially compared to Gypsy elements, which supports the hypothesis of the Copia retrotransposon scarcity in metazoans already established in other taxa (de la Chaux and Wagner, 2011). (2) The GalEa elements are highly predominant within Malacostraca and show a species- or lineage-dependent distribution that may be related to their dynamic in a ‘‘domino days spreading’’ branching process. A major aim of our study was to compare these results with those obtained in fungi, considering two characteristic of these organisms. Fungi belong to the Unikonta and are thus closely related to metazoans, forming with them (and few minor groups) the Opisthokonta (Torruella et al., 2012). Fungi genomes are slight (www.zbi.ee/fungal-genomesize/). For example, the 1C values among 1254 Ascomycota vary from 0.007 to 3.12 pg (mean 0.04 pg), with only three genomes larger than 1 pg. Such genome sizes are clearly smaller than those observed in crustaceans whose 1C values are always larger than 1 pg (Animal Genome Size Database, www.genomesize.com).

Focusing on GalEa elements, we did not look in depth at the distribution of Copia retrotransposons in fungi. However our results seem to agree that there amount is highly variable among genomes, and they often appear scarce. This is consistent with previous a study showing that gypsy elements greatly prevail among LTR retrotransposons in fungi (Muszewska et al., 2011). Moreover, we must take into consideration that, in absence of sequenced genomes, the study on crustacean was made using a PCR approach; with all the classical limits linked to such a detection method. So, it does not seem exaggerated to argue that global distribution patterns of Copia elements within metazoans or fungi appear similar. For half of the 43 tested genomes, selected on the presence of GalEa elements, Copia retrotransposons comprise less 0.5 % of the genome and only 9 genomes are covered by more than 2%. Two of them have a large amount of Copia sequences (about 7% in *M. bicolor* and 8% in *P. tritici-repentis*). An unexpected huge extent of Copia retrotransposons is observed in *C. geophilum* where they represent more than 22% of the genome. This result is coherent with the great number of large Copia insertions detected with both LTRharvest and Repeat masker (1648 and 5653 copies, respectively) and doubtless related to the fact that this species possess one of the ten biggest genomes (177 Mb) among Ascomycota.

Considering the prevalence of GalEa elements among Copia retrotransposons, our results on Pezizomycotina clearly differ from those obtained in crustaceans. In none class or order GalEa elements appear predominant. Actually, in more than half of the species they account for less than a third of Copia retrotransposons, and often much less. Equivalent proportions of both element types are estimated in five species. The extent of GalEa elements exceeds 1% in only eight species, being the majority in five of them. They also represent 70% of Copia retrotransposons in *T. aculeatus*, but in a small quantity (0.4 % of the genome with less than 20 large copies). These species where GalEa element prevail are dispersed throughout the phylogeny. If the distribution of GalEa seems influenced by phylogeny (CoGaFu elements being very common in some orders such as Helotiales, Magnaporthales and Xylariales), this is not the case in quantitative analyses. Except perhaps for the genus *Talaromyces* in which the three genomes available possess several GalEa elements (between 20 to 50 large copies), which contrasts with other species of the Eurotiales order that possess none or just one copy of CoGaFu. If confirmed, such a feature may help to differentiate *Talaromyces* from *Penicillium* species, two genus closely related and difficult to distinguish. Conversely, species from the genus *Pyrenophora* or *Colletotrichum* display significant differences in their number of GalEa elements. It suggests sudden amplification in number of copies likely resulting from bursts of transposition. It could also be the case for species having a huge number of copies such as *C. geophilum* and *M. bicolor*, especially since they have large clusters (data not shown). So, even if the repartition between GalEa and other Copia retrotransposons in Fungi appears different of the one described in crustaceans, the species- or lineage-dependent distribution of GalEa, their variability in copy number and the small number of GalEa clusters obtained per species (data not shown) still agree with a dynamic related to a ‘‘domino days spreading’’ dynamics model.

GalEa retrotransposons remain close to each other between species; which is confirmed by the monophyly of the elements, which are close enough to also encompass ripped copies in the clade. The common origin of GalEa from the CoGaFu clade is confirmed by the absence of a classical PBS, replaced by their singular CHS. We suspect that the CHS plays the role of the PBS considering its position and the strong conservation of the 9 bp palindromic sequence. However, to our knowledge, there is no model that could explain the role of the CHS. Reverse transcription of most retrovirus and LTR retrotransposons required cellular tRNAs to serve as primers of minus-strand strong-stop DNA synthesis (Gabriel and Boeke, 1993) (Leis et al., 1993). However, few LTR retrotransposons are known to have developed other strategies to

ensure reverse transcription initiation. It is the case of Tf1 from *Schizosaccharomyces pombe*, which uses a self-priming mechanism to initiate synthesis of reverse transcript instead of primer derived from a tRNA (Lin and Levin, 1997). Similarly, the initiation of reverse transcription of Rous sarcoma virus has been shown to require the formation of an additional RNA stem-loop structure (Cobrinik et al., 1991). However, such a mechanism seems not suitable for CoGaFu retrotransposons. It especially requires a perfect complementarity between the 11 bp PBS and the first nucleotides of the Tf1 mRNA. But we have not found any U5-inverted repeat sequence that could be complementary to the conserved palindromic 9 nucleotides of the CHS in CoGaFu elements. Moreover, the PBS of Tf1 is not palindromic; even if it is difficult to determine if the palindromic sequence in CoGaFu directly plays a role or whether it is involved only in order to initiate a hairpin structure. Indeed, it is the hairpin structure that appears fundamental for Tf1 transposition since although the loop is required, the specific sequence of the nucleotides within the loop seems unimportant for function (Lin and Levin, 1998). Another particular structure implied in reverse transcription start that combines palindromic sequence and loop is the dimer initiation site (or DIS), a highly conserved stem-loop sequence found in many retroviruses (Berkhout and van Wamel, 1996 ; Dirac et al., 2001). However, this hairpin structure is observed in addition to the PBS and the palindromic sequence localized at the tip.

The common origin of GalEa retrotransposons from fungi, which define a new clade distinct from that of metazoans, raises the question of the origin of GalEa elements in eukaryotes. The presence of GalEa in numerous species of both fungi and metazoans suggest that GalEa elements are probably ancient in Opisthokonta and then diverge following the separation of Fungi and Metazoa. Thereafter, GalEa elements persist in various groups of metazoans and to a lesser extent in fungi. The loss of GalEa retrotransposons in several groups of fungi is consistent with the small genome size of these organisms, probably related to a small number of copies, and the ‘‘domino days spreading’’ dynamics. Especially since the absence of GalEa elements appear well supported only in Basidiomycota, according to the 132 genomes tested, but remain unclear in other groups. Of course we cannot completely rule out the possibility of a very old horizontal transfer between Pezizomycotina and metazoans. Such an hypothesis is more likely in the case of red algae, of which only four are known to possess GalEa retrotransposons. This small number may be partly due to the slight amount of genomic data available. Moreover, the phylogenetic relationships between elements from Rhodophyta

remains unclear, particularly their monophyly. So, at present, only the phylogenic distance that separate Rhodophyta from Opisthokonta and the absence of GalEa element detection outside these two groups argue for a horizontal transfer between red algae and fungi, or metazoan, or a third still indeterminate group.

References

Amselem, J., Cuomo, C.A., van Kan, J.A.L., Viaud, M., Benito, E.P., Couloux, A., Coutinho, P.M., de Vries, R.P., Dyer, P.S., Fillinger, S., et al. (2011). Genomic analysis of the necrotrophic fungal pathogens Sclerotinia sclerotiorum and Botrytis cinerea. PLoS Genet. *7*, e1002230.

Amyotte, S.G., Tan, X., Pennerman, K., Jimenez-Gasco, M. del M., Klosterman, S.J., Ma, L.-J., Dobinson, K.F., and Veronese, P. (2012). Transposable elements in phytopathogenic Verticillium spp.: insights into genome evolution and inter- and intra-specific diversification. BMC Genomics *13*, 314.

Arkhipova, I.R. (2006). Distribution and phylogeny of Penelope-like elements in eukaryotes. Syst. Biol. *55*, 875–885.

Berkhout, B., and van Wamel, J.L. (1996). Role of the DIS hairpin in replication of human immunodeficiency virus type 1. J. Virol. *70*, 6723–6732.

Biémont, C., and Vieira, C. (2006). Genetics: junk DNA as an evolutionary force. Nature *443*, 521–524.

Cameron, J.R., Loh, E.Y., and Davis, R.W. (1979). Evidence for transposition of dispersed repetitive DNA families in yeast. Cell *16*, 739–751.

De la Chaux, N., and Wagner, A. (2011). BEL/Pao retrotransposons in metazoan genomes. BMC Evol. Biol. *11*, 154.

Clutterbuck, A.J. (2011). Genomic evidence of repeat-induced point mutation (RIP) in filamentous ascomycetes. Fungal Genet. Biol. FG B *48*, 306–326.

Cobrinik, D., Aiyar, A., Ge, Z., Katzman, M., Huang, H., and Leis, J. (1991). Overlapping retrovirus U5 sequence elements are required for efficient integration and initiation of reverse transcription. J. Virol. *65*, 3864–3872.

Cuomo, C.A., Güldener, U., Xu, J.-R., Trail, F., Turgeon, B.G., Di Pietro, A., Walton, J.D., Ma, L.-J., Baker, S.E., Rep, M., et al. (2007). The Fusarium graminearum genome reveals a link between localized polymorphism and pathogen specialization. Science *317*, 1400–1402.

Daboussi, M.-J., and Capy, P. (2003). Transposable elements in filamentous fungi. Annu. Rev. Microbiol. *57*, 275–299.

Dirac, A.M., Huthoff, H., Kjems, J., and Berkhout, B. (2001). The dimer initiation site hairpin mediates dimerization of the human immunodeficiency virus, type 2 RNA genome. J. Biol. Chem. *276*, 32345–32352.

Dobinson, K.F., and Hamer, J.E. (1993). The ebb and flow of a fungal genome. Trends Microbiol. *1*, 348–352.

Eickbush, T.H., and Jamburuthugoda, V.K. (2008). The diversity of retrotransposons and the properties of their reverse transcriptases. Virus Res. *134*, 221–234.

Ellinghaus, D., Kurtz, S., and Willhoeft, U. (2008). LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC Bioinformatics *9*, 18.

Fedoroff, N.V. (1999). Transposable Elements As a Molecular Evolutionary Force. Ann. N. Y. Acad. Sci. *870*, 251–264.

Finnegan, D.J. (2012). Retrotransposons. Curr. Biol. CB *22*, R432–R437.

Gabriel, A., and Boeke, J. (1993). Retrotransposon reverse transcription (Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.).

Graïa, F., Lespinet, O., Rimbault, B., Dequard-Chablat, M., Coppin, E., and Picard, M. (2001). Genome quality control: RIP (repeat-induced point mutation) comes to Podospora. Mol. Microbiol. *40*, 586–595.

Idnurm, A., and Howlett, B.J. (2003). Analysis of loss of pathogenicity mutants reveals that repeat-induced point mutations can occur in the Dothideomycete Leptosphaeria maculans. Fungal Genet. Biol. FG B *39*, 31–37.

Katoh, K., Asimenos, G., and Toh, H. (2009). Multiple alignment of DNA sequences with MAFFT. Methods Mol. Biol. Clifton NJ *537*, 39–64.

Kazazian, H.H., Jr (2004). Mobile elements: drivers of genome evolution. Science *303*, 1626–1632.

Leis, J, Aiyar, A., and Cobrinik, D. (1993). Reverse transcriptase (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.).

Lin, J.H., and Levin, H.L. (1997). Self-primed reverse transcription is a mechanism shared by several LTR-containing retrotransposons. RNA N. Y. N *3*, 952–953.

Lin, J.H., and Levin, H.L. (1998). Reverse transcription of a self-primed retrotransposon requires an RNA structure similar to the U5-IR stem-loop of retroviruses. Mol. Cell. Biol. *18*, 6859–6869.

Llorens, C., Muñoz-Pomer, A., Bernad, L., Botella, H., and Moya, A. (2009). Network dynamics of eukaryotic LTR retroelements beyond phylogenetic trees. Biol. Direct *4*, 41.

Llorens, C., Futami, R., Covelli, L., Domínguez-Escribá, L., Viu, J.M., Tamarit, D., Aguilar-Rodríguez, J., Vicente-Ripolles, M., Fuster, G., Bernet, G.P., et al. (2011). The Gypsy Database (GyDB) of mobile genetic elements: release 2.0. Nucleic Acids Res. *39*, D70–D74.

Maumus, F., Allen, A.E., Mhiri, C., Hu, H., Jabbari, K., Vardi, A., Grandbastien, M.-A., and Bowler, C. (2009). Potential impact of stress activated retrotransposons on genome evolution in a marine diatom. BMC Genomics *10*, 624.

Muszewska, A., Hoffman-Sommer, M., and Grynberg, M. (2011). LTR retrotransposons in fungi. PloS One *6*, e29425.

Nakayashiki, H., Kiyotomi, K., Tosa, Y., and Mayama, S. (1999). Transposition of the retrotransposon MAGGY in heterologous species of filamentous fungi. Genetics *153*, 693–703.

Peddigari, S., Zhang, W., Sakai, M., Takechi, K., Takano, H., and Takio, S. (2008). A copia-like retrotransposon gene encoding gypsy-like integrase in a red alga, Porphyra yezoensis. J. Mol. Evol. *66*, 72–79.

Piednoël, M., and Bonnivard, E. (2009). DIRS1-like retrotransposons are widely distributed among Decapoda and are particularly present in hydrothermal vent organisms. BMC Evol. Biol. *9*, 86.

Piednoël, M., Gonçalves, I.R., Higuet, D., and Bonnivard, E. (2011). Eukaryote DIRS1-like retrotransposons: an overview. BMC Genomics *12*, 621.

Piednoël, M., Donnart, T., Esnault, C., Graça, P., Higuet, D., and Bonnivard, E. (2013). LTR-Retrotransposons in R. exoculata and Other Crustaceans: The Outstanding Success of GalEa-Like Copia Elements. PLoS ONE *8*, e57675.

Santana, M.F., Silva, J.C.F., Mizubuti, E.S.G., Araújo, E.F., Condon, B.J., Turgeon, B.G., and Queiroz, M.V. (2014). Characterization and potential evolutionary impact of transposable elements in the genome of Cochliobolus heterostrophus. BMC Genomics *15*, 536.

Terrat, Y., Bonnivard, E., and Higuet, D. (2008). GalEa retrotransposons from galatheid squat lobsters (Decapoda, Anomura) define a new clade of Ty1/copia-like elements restricted to aquatic species. Mol. Genet. Genomics MGG *279*, 63–73.

Torruella, G., Derelle, R., Paps, J., Lang, B.F., Roger, A.J., Shalchian-Tabrizi, K., and Ruiz-Trillo, I. (2012). Phylogenetic relationships within the Opisthokonta based on phylogenomic analyses of conserved single-copy protein domains. Mol. Biol. Evol. *29*, 531–544.

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J.L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., et al. (2007). A unified classification system for eukaryotic transposable elements. Nat. Rev. Genet. *8*, 973–982.

Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. *31*, 3406–3415.

Zytnicki, M., and Quesneville, H. (2011). S-MART, a software toolbox to aid RNA-Seq data analysis. PloS One *6*, e25988.

Discussion supplémentaire

Nous avons analysé la part des éléments Copia et plus précisément GalEa au sein des espèces de champignons, grâce à 2 approches différentes. Lors de l'étude du nombre de copies de ces éléments au sein des génomes (Table 1), nous pouvons considérer des éléments potentiellement complets, ici des éléments avec des LTR et les domaines internes; retrouvées grâce à LTR Harvest. LTR Harvest ne permet pas de récupérer les copies avec des LTR trop divergentes, si l'une des 2 LTR est perdue, ou encore si une séquence s'est insérée au sein de l'élément augmentant de manière considérable sa taille. Avec RepeatMasker (RM), en fixant une taille minimum d'élément à 3kb, on estime le nombre de copies des éléments avec des domaines conservés, pas forcément de pleine taille, tout en s'affranchissant des traces et des petits fragments. Cela nous permet de trouver des éléments partiels, par exemple ayant perdus leurs LTR, mais avec une *pol* assez conservée. Chacune des 2 approches ne nous permet pas de récupérer le nombre de copies totales au sein d'un génome mais nous donne des résultats complémentaires. Ainsi, en les combinant, nous affinons l'estimation du nombre de copies.

Comme attendu le nombre des copies de plus de 3Kb estimé avec RepeatMasker (RM) est plus grand ou égal qu'avec LTR Harvest, plus stringent sur la recherche. Dans quelques rares cas, les copies détectées par LTR Harvest ne le sont pas par RM. Cela peut s'expliquer par la présence de grandes répétitions de N au sein des séquences identifiées par LTR Harvest. Celles-ci sont alors scindées en 2 par RM et passent hors du seuil de 3kb. A l'inverse, RM est lui capable de comptabiliser les éléments de grandes tailles indépendamment de la présence des 2 LTR.

Les résultats de RM et LTR Harvest sont globalement semblables pour 28 espèces. Pour 6 espèces, on obtient un nombre beaucoup plus élevé de copies avec RM (facteur 3). Ceci est particulièrement vrai pour des génomes où LTR Harvest repère déjà beaucoup de copies (Ceg, Cado), même si ce n'est pas toujours le cas (Amli, Ptt). Cela s'observe également pour des espèces présentant peu de copies avec des LTRs intègres (Antav, Cm). Une telle différence est difficile à interpréter en l'absence d'analyse fine des séquences concernées. On peut cependant émettre l'hypothèse que dans ces génomes les éléments sont assez remaniés au point de ne plus avoir de LTR identifiable même s'il reste globalement reconnaissable par RM car la qualité de l'assemblage et du séquençage du génome est variable d'une espèce à l'autre. Nous avons

recherché à faire un comparatif sur un grand nombre d'espèce et nous ne recherchions pas à avoir une vision exhaustive et fine de tous les éléments de tous les génomes. Un élément peut être présent au sein d'un génome et non détectable par LTR Harvest et son analyse ne sera donc pas faite par RM. Toutefois, nous avons privilégié de travailler sur des éléments corrects, ainsi nos différents filtre par LTR Harvest, blast et clustering nous permet de récupérer des séquences propres dont on peut être sûre. Au final l'utilisation de ces 2 approches parait judicieuse et permet plus de sécurité. Ainsi pour la suite, nous pourrons continuer d'utiliser LTR Harvest pour récupérer des éléments pleine taille potentiellement bien conservés au sein des espèces, à partir desquels on pourra faire une estimation du nombre de copies plus précise grâce à RM, tout en remarquant les cas particuliers montrant une forte différence entre les 2.

D'après ces résultats, au sein des champignons, les GalEa n'ont pas l'air d'être majoritaire parmi les éléments Copia. Une question qui se pose est donc de savoir si le succès des éléments GalEa au sein des crustacés est dû à une propriété du phylum ou des GalEa. Pour répondre à cela, nous pensons donc qu'une étude chez un second groupe de métazoaires est indispensable. Nous avons décrit des éléments de type GalEa chez d'autres espèces de crustacés et de téléostéens ainsi que chez d'autres espèces de métazoaires (mollusques, chordés, cnidaires, cténaires, échinodermes et hémichordés) (Piednoël et al., 2013).

DISCUSSION ET PERSPECTIVES

Chapitre V. Discussion

Dynamique des éléments Gypsy et Copia

Au sein des crustacés, nous avons décrit deux dynamiques très différentes pour deux types de rétrotransposons à LTR, les Gypsy et les Copia, présentant une structure assez proche. Tandis que les éléments Gypsy sont bien présents et variés chez les crustacés, les éléments Copia sont limités à certaines espèces et appartiennent majoritairement au clade des GalEa. Llorens et al. dans leur GypsyDataBase (2009; http://gydb.org/index.php/Main_Page) ont regroupé de nombreuses informations concernant les clades composant les éléments, comme les taxons hôtes, la branche sur laquelle on retrouve l'élément (d'après les arbres phylogénétiques des éléments non enracinés), le clade de l'élément et la présence de domaines supplémentaires comme l'enveloppe ou les chromodomaines (Tableau 6, et Tableau 1 dans l'introduction). Ainsi, dans le phylum des métazoaires, on retrouve 17 familles d'éléments Gypsy dans de nombreux taxons tels que les vertébrés, les arthropodes ou les nématodes. De plus, au sein d'un groupe d'hôte on peut avoir plusieurs clades différents de Gypsy. Par exemple, chez les deutérostomiens, il a été décrit 5 clades différents les Tor 1, 2, 4, Cigr-1 et les Gmr1. A l'inverse, les éléments Copia décrits appartiennent à seulement 7 clades au sein de 3 grands taxons : les arthropodes, les cnidaires et les « arthropodes marins » (regroupant en fait des crustacés, des actinoptérygiens et des urochordés) (cf. Tableau 1). On observe la même diversité chez des champignons : 8 clades d'éléments Gypsy (Tableau 6) sont décrits contre seulement 3 clades de Copia (Tableau 1). Les données concernant la présence d'éléments Copia au sein des champignons a été récemment complété par nos travaux. Pour l'instant nous nous sommes concentré sur les éléments GalEa sans regarder dans combien de clade se répartissent les Copia non GalEa. Dans le phylum des plantes, 6 clades de Gypsy sont présents chez des plantes terrestres et 2 clades chez des algues vertes tandis que l'on retrouve 4 clades de Copia au sein des plantes terrestres, 1 clade chez des algues vertes et 2 clades chez les algues rouges, dont le clade des GalEa récemment décrit.

Host phyla	Branch	Genus and/or cluster	clade	Env	Chromodomain
Viridiplantae	Branch 1 *(Chromoviridae)*	Plants	CRM	no	no
Viridiplantae	Branch 1 *(Chromoviridae)*	Plants	*Del*	no	yes
Viridiplantae	Branch 1 *(Chromoviridae)*	Plants	*Galadriel*	no	yes
Viridiplantae	Branch 1 *(Chromoviridae)*	Plants	*Reina*	no	yes
Chlorophyta	Branch 1 *(Chromoviridae)*	Plants	*REM1*	no	yes
Cryptophyta	Branch 1 *(Chromoviridae)*	Plants	*G-Rhodo*	no	yes
Fungi	Branch 1 *(Chromoviridae)*	Fungi and vertebrates	*Pyggy*	no	yes
Fungi	Branch 1 *(Chromoviridae)*	Fungi and vertebrates	*MGLR3*	no	yes
Fungi	Branch 1 *(Chromoviridae)*	Fungi and vertebrates	*Pyret*	no	yes
Fungi	Branch 1 *(Chromoviridae)*	Fungi and vertebrates	*Maggy*	no	yes
Fungi	Branch 1 *(Chromoviridae)*	Fungi and vertebrates	*MarY1*	no	yes
Fungi	Branch 1 *(Chromoviridae)*	Undefined	*Tse3*	no	yes
Fungi	Branch 1 *(Chromoviridae)*	Fungi and vertebrates	*TF1-2*	no	no
Fungi	Branch 1 *(Chromoviridae)*	Undefined	*Ty3*	no	no
Vertebrata	Branch 1 *(Chromoviridae)*	Fungi and vertebrates	*V-clade*	no	yes
Amoebozoa	Branch 1 *(Chromoviridae)*	Undefined	*Skipper*	no	yes
Viridiplantae	Branch 2	*Athila/Tat*	*Athila*	yes	no
Viridiplantae	Branch 2	*Athila/Tat*	*Tat*	no	no
Arthropoda	Branch 2	*Errantiviridae*	*17.6*	yes	no
Arthropoda	Branch 2	*Errantiviridae*	*Gypsy*	yes	no
Arthropoda	Branch 2	*Metaviridae*	*412/mdg1*	no	no
Arthropoda	Branch 2	*Metaviridae*	*Micropia/mdg3*	no	no
Bilateria	Branch 2	*Mag*	*A-clade*	no	no
Bilateria	Branch 2	*Mag*	*B-clade*	no	no
Bilateria	Branch 2	*Mag*	*C-clade*	no	no
Deuterostomia	Branch 2	*Metaviridae*	*Gmr1*	no	no
Arthropoda	Branch 2	*Metaviridae*	*Osvaldo*	Only *Osvaldo*	no
Nematoda	Branch 2	*Metaviridae*	*Cer2-3*	no	no
Nematoda	Branch 2	*Metaviridae*	*Cer1*	no	no
Protostomia	Branch 2	*Metaviridae*	CsRN1	no	no
Deuterostomia	Branch 2	*Metaviridae*	*Tor1*	no	no
Deuterostomia	Branch 2	*Metaviridae*	*Tor4*	some elements	no
Deuterostomia	Branch 2	*Metaviridae*	*Tor2*	no	no
Deuterostomia	Branch 2	*Metaviridae*	*Cigr-1*	no	no

Tableau 6: Tableau regroupant les clades Gypsy par branches sur l'arbre phylogénétique, par phylum d'hôtes, par clade et présence de domaines *env* ou Chromodomaines. (Llorens, GypsyDatabase) Branch1-2 correspond à la branche sur laquelle se placent les éléments dans l'arbre phylogénétique des Gypsy. Tableau complémentaire du Tableau 1 des Copia en introduction.

Pour l'instant, d'après nos résultats au sein des champignons, les GalEa ne sont pas majoritaires parmi les Copia. Une étude plus complète des éléments Copia au sein des génomes de champignons nous permettra de savoir si on retrouve un clade de Copia prédominant. De plus les GalEa semblent présents chez presque toutes les espèces chez lesquelles nous les avons recherchés au sein des crustacés, tandis que chez les champignons, et plus particulièrement les pezizomycotina, on en retrouve des éléments GalEa que chez 49 espèces sur les 184 séquencées.

Une question que l'on se pose est de savoir si la domination des GalEa au sein des crustacés est lié aux ETs ou bien lié au phylum ? Pour répondre à cela nous avons besoin de réaliser une étude des éléments GalEa au sein d'un second groupe de métazoaire où nous savons qu'il existe des GalEa. Nous avons 2 choix de taxons parmi lesquels on connait un bon nombre d'espèces présentant des GalEa: les actinoptérygiens ou les mollusques. Les crustacés sont au niveau du sous embranchement, tandis que les actinoptérygiens sont au niveau de la classe (niveau plus petit) et les mollusques au niveau de l'embranchement (niveau plus grand). Si nous voulons réaliser une étude à large échelle, il semble plus intéressant de réaliser cette étude chez les mollusques. Nous avons donc souhaité vérifier si la dynamique des éléments Copia et Gypsy retrouvée chez les crustacés était similaire au sein d'un taxon équivalent: les mollusques. Des études de distribution des éléments DIRS, BEL/Pao et PLE ont été réalisées, présentant ces éléments avec une distribution parcellaire au sein des eucaryotes (Arkhipova, 2006; de la Chaux and Wagner, 2011; Piednoël et al., 2011). Nous avons souhaité élargir notre étude comparative au sein des Mollusques à ces rétrotransposons.

Les mollusques sont un embranchement de métazoaires qui présentent plus de 117 000 espèces avec une forte diversité de style de vie (libre ou parasite) et d'habitats (terrestre, marin, eau douce ou des milieux extrêmes comme les sources hydrothermales). Leurs génomes présentent une grande variation de taille: *Lottia gigantea* (une patelle) présente un génome de 0,36 pg (soit 10 fois moins grand que le génome humain) alors que *Diplommatina collarifera collarifera* (un type d'escargot de mer) possède un génome 19 fois plus important de 6,71 pg (genomesize.com). Compte tenu de la diversité des ETs à étudier et de la taille du taxon des mollusques, nous avons décidé d'utiliser une approche *in silico*. Le principe de cette étude est de réaliser une recherche par blast dans toutes les bases de données disponibles, sur le même principe que celui développé pour l'étude des Copia chez les champignons. Nous avons obtenus des résultats chez 36 espèces: 17 bivalves (dont 3 génomes complètement séquencés:

Crassostrea gigas, Pinctada fucata et *Pinctada martensii),* 14 gastéropodes (dont 3 génomes complètement séquencés: *Aplysia californica, Lottia gigantea* et *Biomphalaria glabrata)* et 5 céphalopodes. Dans le cadre de données transcriptomiques et de génomes partiels, une espèce a été retenue si 1 des 5 types d'ET ressortait lors de la recherche. Nous n'avons pas trouvés d'éléments dans les 5 autres classes de mollusques (les Solénogastres, Caudofovéates, Polyplacophores, Monoplacophores et Scaphopodes). Ces classes présentent un faible nombre d'espèces décrites avec très peu de données génomiques ou transcriptomiques disponibles.

Nos résultats confirment que les éléments Gypsy sont bien répartis au sein des espèces de mollusques (Tableau 7). On en retrouve chez au moins 24 des 36 espèces étudiées avec jusqu'ici 70 familles décrites (familles définies sur le critère d'au moins 80% d'identité sur les séquences alignées). Ces éléments Gypsy présents sont divisés dans au moins 14 clades (Figure 18). Un tiers est regroupé dans un clade spécifique (bootstrap de 73) dans lequel on retrouve des éléments de 13 espèces de gastéropodes et bivalves (mais pas d'éléments de céphalopodes). Les autres éléments Gypsy sont dispersés dans la phylogénie. Ils se regroupent soit dans des clades spécifiques avec uniquement des éléments de mollusques, soit avec des éléments de crustacés. Les éléments de Gypsy de mollusques ne suivent pas la phylogénie des espèces. En effet, les clades comprennent des éléments provenant d'espèces différentes et les éléments d'une même espèce peuvent appartenir à des clades éloignés. Par exemple, chez le gastéropode *Bithymia siamensis*, 7 familles d'éléments Gypsy se retrouvent dans 5 clades différents. Effectivement, chez les 24 espèces, nous observons un très forte diversité intra-spécifique. Au sein d'une même espèce nous pouvons retrouver un grand nombre de familles: 6 chez *Villosa lienosa* et *Placobranchus ocellatus,* ou 8 chez *Lymnaea stagnalis* et *Biomphalaria glabrata* (cf. Tableau 7).

Espèces		Copia			Gypsy	PaoBel	PLE	DIRS
		GalEa	Hydra 1-2	autre Copia				
Alasmidonta varicosa	B				GyAlva1 & 2			
Azumapecten farreri	B				CFG1			
Crassostrea angulata	B			CoCrassang1&2				
*Crassostrea gigas**	B	CoCrassos1-3				Cragibel1-7	PeCragi1-5	YCragi1&2
Elliptio complanata	B				GyEllico1			
Meretrix meretrix	B				GyMeme1			YMeme1
Mizuhopecten yessoensis	B				PYG1			YMiye1&2
Mytilus californianus	B					Mycabel1&2	PeMyca1	
Mytilus galloprovincialis	B	CoMyga1-3			GyMyga1&2	Mygabel1&2		YMyga1
*Pinctada fucata**	B					Pinfubel1-5	PePifu1-4	
*Pinctada martensii**	B				GyPima1&2			
Placopecten magellanicus	B	CoPlama1					PePlama1-3	
Ruditapes philippinarum	B				GyRuphi1-4	Ruphibel1	PeRuphi1&2	YRuphi1
Sinonovacula constricta	B				GySico1&2			YSico1
Spisula solidissima	B				GySpiso1&2			
Tegillarca granosa	B					Tegrabel1		
Villosa lienosa	B				GyVilli1-6		PeVili1-4	YVili1&2
Doryteuthis pealeii	C	CoDope1						
Euprymna scolopes	C	CoEusco5	CoEusco2&4	CoEusco3	GyEusco1-3		PeEusco1&2	
Idiosepius paradoxus	C	CoIdio1						
Octopus vulgaris	C		CoOcto1&2		GyOcto1 & 2		PeOcto1	
Sepia officinalis	C	CoSeof1	CoSeof2	CoSeof3	GySeof1-3	Seofbel1	PeSeof1-5	
*Aplysia californica**	G	CoApca2	CoApca1			Apcabel1&2	PeApca1&2	21
Aplysia kurodai	G				GyApku1			
*Biomphalaria glabrata**	G				GyBigla1-8		PeBigla1	
Bithynia siamensis	G				GyBisia1-7	Bisibel1	PeBisi1&2	YBisi1-3
Crepidula fornicata	G	CoCre1			GyCrefo1	Crefobel1	PeCrefo1	YCrefo1-4
Elysia timida	G		CoElyti1		GyElyti1-3			
Haliotis diversicolor	G	CoHadi1						
Ilyanassa obsoleta	G				GyIlob1			YIlob1&2
Littorina saxatilis	G				GyLisa1 & 2			YLisa1-4
*Lottia gigantea**	G	CoLogi1				Logibel1-7	PeLogi1&2	7
Lottia scutum	G						PeLoscu1	
Lymnaea stagnalis	G				GyLysta1-8			
Physa acuta	G				GyPhyac1			
Placobranchus ocellatus	G				GyPlaco1-6		PePlaco1	YPlaco1
		15	7	4				
			26		70	30	37	52

Tableau 7 : Résultats de la distribution de rétrotransposons au sein de mollusques

B : Bivalve ; G : gastéropode, C : céphalopode ; * : génome complètement séquencé. Les éléments Gypsy PYG1 et CFG1 avaient déjà été décrits dans la littérature (Wang et al., 2008), ainsi que les familles de DIRS d'*Aplysia californica* et *Lottia gigantea* (Piednoël et al., 2011). Lorsque nous n'avons pas testé la présence d'un élément ou que nous ne l'avons pas trouvé chez une espèce, nous avons laissé une case blanche.

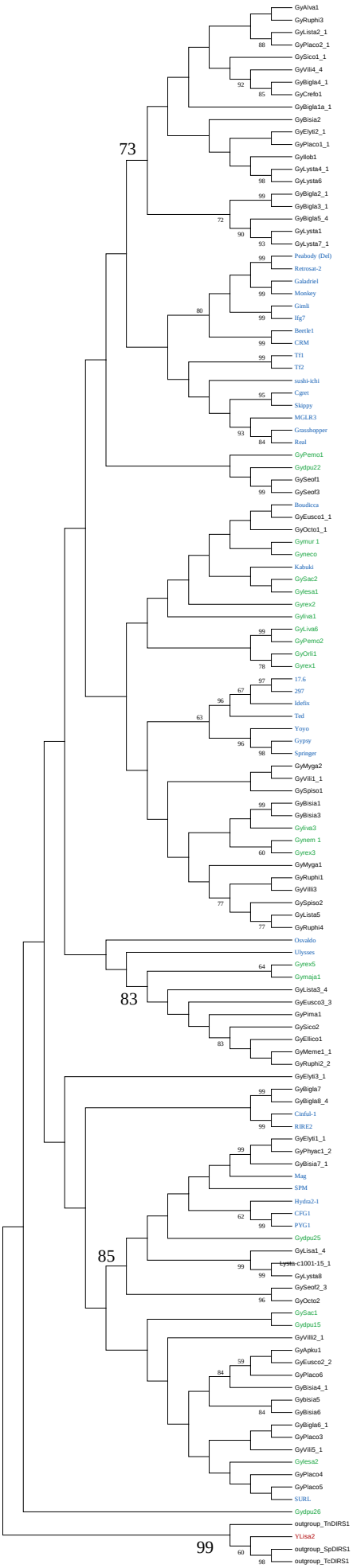

GyAlva1
GyRuphi3
GyLista2_1
GyPlaco2_1
GySico1_1
GyVili4_4
GyBigla4_1
GyCrefo1
GyBigla1a_1
GyBisia2
GyElyti2_1
GyPlaco1_1
Gyllob1
GyLysta4_1
GyLysta6
GyBigla2_1
GyBigla3_1
GyBigla5_4
GyLysta1
GyLysta7_1
Peabody (Del)
Retrosat-2
Galadriel
Monkey
Gimli
Ifg7
Beerle1
CRM
Tf1
Tf2
sushi-ichi
Cgret
Skippy
MGLR3
Grasshopper
Real
GyPemo1
Gydpu22
GySeof1
GySeof3
Boudicca
GyEusco1_1
GyOcto1_1
Gymur 1
Gyneco
Kabuki
GySac2
Gylesa1
Gyrex2
Gyliva1
GyLiva6
GyPemo2
GyOrli1
Gyrex1
17.6
297
Idefix
Ted
Yoyo
Gypsy
Springer
GyMyga2
GyVili1_1
GySpiso1
GyBisia1
GyBisia3
Gyliva3
Gynem 1
Gyrex3
GyMyga1
GyRuphi1
GyVilli3
GySpiso2
GyLista5
GyRuphi4
Osvaldo
Ulysses
Gyrex5
Gymaja1
GyLista3_4
GyEusco3_3
GyPima1
GySico2
GyEllico1
GyMeme1_1
GyRuphi2_2
GyElyti3_1
GyBigla7
GyBigla8_4
Cinful-1
RIRE2
GyElyti1_1
GyPhyac1_2
GyBisia7_1
Mag
SPM
Hydra2-1
CFG1
PYG1
Gydpu25
GyLisa1_4
Lysta c1001-15_1
GyLysta8
GySeof2_3
GyOcto2
GySac1
Gydpu15
GyVilli2_1
GyApku1
GyEusco2_2
GyPlaco6
GyBisia4_1
Gybisia5
GyBisia6
GyBigla6_1
GyPlaco3
GyVili5_1
Gylesa2
GyPlaco4
GyPlaco5
SURL
Gydpu26
outgroup_TnDIRS1
YLisa2
outgroup_SpDIRS1
outgroup_TcDIRS1
73
83
85
99

Figure 18 : Arbre phylogénétique représentant les relations entre les rétrotransposons Gypsy après une analyse Neighbor-Joining sur la RT / RH des éléments en acides aminés. Les éléments de crustacés décrits dans la précédente étude sont indiqués en vert, les éléments de mollusques sont en noir et les éléments de référence de Gypsy décrit dans la bibliographie sont en bleu. Nous avons laissé les bootstraps supérieur à 60 et agrandi les bootstraps intéressants pour notre analyse. Les séquences DIRS1 ont été utilisées comme outgroup (bootstrap à 99), ainsi qu'une séquence de DIRS1 de mollusque (en rouge).

Concernant les éléments Copia, ils ne sont détectés que chez 14 espèces de mollusques et on en dénombre en tout 26 familles (Tableau 7). On retrouve une diversité intra-spécifique avec 1 à 5 familles de Copia au sein des espèces (3 familles chez *Mytilus galloprocincialis*, 5 familles chez *Euprymna scolopes*). Présents chez 11 des 14 espèces, les éléments du clade des GalEa paraissent majoritaires parmi les Copia avec 16 familles décrites. Un second clade de Copia est également bien représenté chez les mollusques, le clade Hydra1-2. Préalablement décrit chez le cnidaire *Hydra magnipapillata*, il est également présent chez d'autres espèces telles que le poisson zèbre (*Danio rerio*) (Llorens et al., 2009). On retrouve 6 éléments de ce clade Hydra 1-2 chez 5 espèces de mollusques dont 3 céphalopodes. Nous avons retrouvé également 4 familles d'éléments n'appartenant ni au clade GalEa ni au clade Hydra1-2 chez 3 espèces différentes.

Quel que soit le groupe (crustacés ou mollusques) ou l'échelle (embranchement ou phylum) que l'on regarde, les éléments Gypsy sont bien présents au sein d'un grand nombre d'espèces avec une grande diversité. Les Copia sont plus difficiles à trouver et ont l'air d'être moins divers. Le clade GalEa représente la très grande majorité des éléments Copia chez les crustacés et plus de la moitié chez les mollusques. Un quart des éléments Copia de mollusques se regroupent dans le second clade Hydra1-2. La similarité des résultats obtenus dans ces 2 taxons d'espèces montrent que les dynamiques des éléments Copia et Gypsy précédemment décrites ne sont donc pas une particularité des crustacés.

Les éléments BEL/Pao, PLE et DIRS sont détectés dans un nombre équivalent mais limité d'espèces. On retrouve 30 familles de BEL/Pao au sein des 11 espèces, avec une diversité intra-spécifique importante, jusqu'à 7 familles chez *L. gigantea*. Nous avons décrit 52 familles de DIRS au sein de 16 espèces de mollusques, avec une diversité intra-spécifique relative: entre 1 à 4 familles. Néanmoins, l'étude des éléments DIRS réalisé par Piednoël et al., (2011) a révélé jusqu'à 7 et 21 familles pour les éléments DIRS au sein des génomes complètements séquencés de *L. gigantea* et de *A. californica*. Enfin, les éléments PLE ont été identifiés au sein

de 14 espèces avec 37 familles. On retrouve également une diversité intra-spécifique relative avec 1 à 5 familles (1 *chez Cepidula fornicata*, 5 chez *S. officinalis).* Les PLE semblent variés au sein des espèces où ils sont présents (5 espèces avec au moins 3 familles de PLE).

Jusqu'à aujourd'hui, lorsqu'un clade d'éléments était présent au sein d'un groupe d'espèces, on décrivait ce clade uniquement dans le taxon dans laquelle il avait été trouvé, chez les plantes, les champignons ou les métazoaires. Par exemple, dans la superfamille des Copia, le clade Tork est décrit chez les plantes, le clade 1731 uniquement chez les arthropodes ou encore les clades CoDi chez les diatomées. De la même façon pour les rétrotransposons Gypsy, le clade CsRN1 se retrouve chez les protostomiens et le clade Reina chez les plantes. Le clade des GalEa a été décrit chez des métazoaires en premier lieu mais également chez des algues rouges et chez des champignons, ce qui en fait le premier clade d'éléments à LTR décrit chez plus d'un phylum et même chez 3 grands phylums d'eucaryotes. Cette répartition peut être propre aux GalEa ou alors lié à la manière de le rechercher. En effet, il existe d'autres études similaires avec d'autres clades d'ETs; par exemple les transposons Tc1/Mariner qui ont été retrouvés ubiquitairement chez les eucaryotes. Il pourrait être intéressant de voir si on retrouve une répartition identique pour des éléments tels que Hydra1-2, plus largement chez les métazoaires ou au-delà, chez les champignons ou les plantes par exemple.

Les éléments GalEa de crustacés présentait trois singularités: présence dans des espèces phylogénétiquement éloignées, une distribution discontinue et une apparente restriction aux espèces aquatiques. Ce dernier point est clairement remis en cause par les études menées au sein des champignons et des mollusques.

Pourquoi une différence de dynamique entre les éléments Gypsy et Copia ?

La dynamique des ETs est un concept complexe, qui combine de nombreux aspects comme les mécanismes de contrôle de la transposition par les éléments eux-mêmes et / ou le génome de l'hôte, l'activation de l'élément par des changements environnementaux (au niveau du génome ou au niveau écologique), etc. Beaucoup de ces paramètres sont soumis à des événements aléatoires. Les éléments Gypsy et Copia présentent une dynamique différente qui pourraient être présentée comme suit. Les Gypsy ont une dynamique qui suit la théorie de la Reine Rouge : l'évolution permanente d'une espèce est nécessaire pour maintenir son aptitude à survivre suite aux évolutions des espèces avec lesquelles elle co-évolue. Cette hypothèse

suppose que l'environnement d'un groupe concurrentiel d'organismes (principalement les autres organismes vivants tels que des prédateurs ou des parasites) se modifierait en permanence, si bien que l'effort d'adaptation serait toujours à recommencer, et l'extinction toujours aussi probable. On peut appliquer cette théorie aux ETs au sein d'un génome. Les éléments auraient besoin d'évoluer sans cesse pour pouvoir se maintenir au sein des génomes dans lesquels ils sont intégrés, se réamplifiés en échappant aux systèmes de régulation. Cela implique plusieurs copies et une évolution continue avec un taux de transposition basale assez élevée ou du moins une possibilité de bouger assez souvent au sein des génomes, et cela expliquerait leur très grande diversité, au sein d'un taxon et au sein même d'une espèce.

Une autre dynamique possible observée pour des éléments Copia (comme des GalEa), et probablement certains autres éléments, est une dynamique où quelques copies d'éléments actifs doivent être héritées avant de connaitre un « burst » de transposition. Cela correspond à une faible diversité et un petit nombre de copies mais avec une possibilité d'explosion de la transposition. Cela expliquerait la distribution très morcelée et les variations du nombre de copies au sein des génomes. Ce type de dynamique est une dynamique de « Domino day spreading ». Pour la décrire, on peut établir une analogie avec un jeu, le domino's days, un événement mondial dont le but est de renverser le plus de dominos possible. Lors de cet évènement télévisé, nous pouvons suivre la propagation de chutes de dominos le long de différentes branches qui passent également par plusieurs grandes figures. Chaque branche et figure présente un nombre variable de dominos. Les éléments Copia pourraient être représentés par ces dominos et le nombre de copies par le nombre de chutes de dominos. Ces chutes nous aident à visualiser la diffusion des éléments au sein des taxons et des espèces au cours de l'évolution. Comme les dominos qui suivent un nombre restreint de lignes avant de renverser des grandes structures.

De nombreux facteurs pourraient conduire à une telle expansion brutale au sein d'une espèce. Par exemple, la transposition d'éléments peut être activée par des stress ou la colonisation d'un nouvel environnement. Au sein du jeu, les grandes structures de dominos permettent de progresser aux structures suivantes *via* plusieurs chemins. De même, une amplification initiale augmente la proportion de « jeunes » éléments actifs, qui permettent des amplifications dérivés ultérieures dans certains lignages aléatoires, éventuellement par le biais de la transposition de quelques copies maîtresses. En outre, le nombre limité de dominos peut faciliter la rupture aléatoire lors de leur progression le long des chemins. De même, les forces évolutives peuvent

conduire à l'extinction de certains éléments au sein d'une lignée lorsque les éléments sont maintenus trop longtemps à un faible nombre de copies et ainsi expliquer la distribution fragmentée.

La différence de dynamique entre les éléments Copia et Gypsy au sein des métazoaires et des champignons peut s'expliquer grâce à différentes hypothèses. Les premières hypothèses sont liées aux domaines complémentaires que les éléments Gypsy de métazoaires et de champignons peuvent avoir; tels qu'un chromodomaine ou une enveloppe, contrairement aux éléments Copia pour qui ces domaines ont été assez peu décrits. L'ajout de domaines pourrait expliquer une différence de dynamique.

Les chromodomaines sont des domaines protéiques de 40-50 acides aminés impliqués dans le remodelage de la chromatine et la régulation de l'expression génique chez les eucaryotes (Cavalli and Paro, 1998; Koonin et al., 1995). Une action possible des chromodomaines est liée au fait qu'ils peuvent cibler certains sites d'intégration, notamment dans l'hétérochromatine (Gao et al., 2008). Ainsi, la fusion d'un chromodomaine à l'intégrase du rétrotransposon Tf1 redirige son intégration. L'accumulation d'éléments dans l'hétérochromatine va influer sur les deux partenaires. Pour l'hôte, l'hétérochromatine peut être majoritairement responsable des variations de taille du génome, comme dans le cas d'*A. thaliana* (Hall et al., 2006). Pour l'élément, il sera moins contre-sélectionné si son intégration à lieu dans l'hétérochromatine (Boeke and Devine, 1998), sans pour autant que cela bloque nécessairement sa capacité à transposer (Ke et al., 1997). Les Gypsy à chromodomaine pourraient donc ainsi posséder un avantage expliquant en partie une répartition plus large et un plus grand nombre de copies.

La présence d'une enveloppe au sein des ETs peut également expliquer en partie la dynamique des éléments Gypsy. En effet, les enveloppes sont des protéines virales, qui sont présentes au sein des virus dit enveloppés (contrairement aux virus nus). Les virus enveloppés ont la particularité de pouvoir sortir de leur cellule hôte sans déclencher la mort de celle-ci. L'enveloppe virale a une grande importance pour l'infection de la cellule par le virus, la stabilité envers les influences externes, ainsi qu'une plus grande capacité de changement de la surface du virus. Les Gypsy sont proche des rétrovirus. On classe donc les rétrovirus chez les ETs. Mais les virus ont également leur propre classification dans laquelle on retrouve les éléments Gypsy (Figure 19). Ainsi, grâce aux protéines d'enveloppe, les éléments Gypsy pourraient avoir un potentiel infectieux plus fort ce qui expliquerait leur distribution au sein de

nombreuses espèces.

La question est de savoir si les éléments Gypsy observés dans notre étude possèdent effectivement ces types de domaines supplémentaires.

Une autre hypothèse concerne la séquence des éléments. En effet, les éléments Gypsy et Copia présentent une séquence de la RT, RH, et INT différente entre eux. Des études phylogénétiques ont montré que les éléments DIRS et BEL/Pao sont plus proches des éléments Gypsy que les éléments Copia (Figure 10, introduction). De plus la séquence de la *pol* des éléments Gypsy est proche de celle des rétrovirus (Figure 19) et les éléments Gypsy présentent une position de leur INT au sein de leur *pol* en 3' de la RT et RH, comme chez les rétrovirus, tandis que les Copia présentent leur INT en 5' de leur RT et RH. Sans comprendre comment cette différence de séquence ou de structure peut jouer sur la dynamique des éléments, ces différences peuvent être un élément de réponse.

Un autre facteur possible est le fort nombre de copies des éléments Gypsy. En effet, le fait d'avoir un grand nombre de copies au sein des génomes facilite la dispersion des éléments, donc leur maintien et peut augmenter la diversité suite à l'évolution des copies indépendamment les unes des autres. On sait que ces éléments présentent, au sein des métazoaires et des champignons, un fort nombre de copies au sein des génomes (De la Chaux et Wagner, 2011 et Muszewska et al., 2011).

Une autre idée est une différence de dynamique du à une réponse différente aux transferts horizontaux. En effet, on sait que les ETs répondent différemment aux stress qui peuvent venir de l'environnement, et ainsi faciliter les transferts horizontaux.

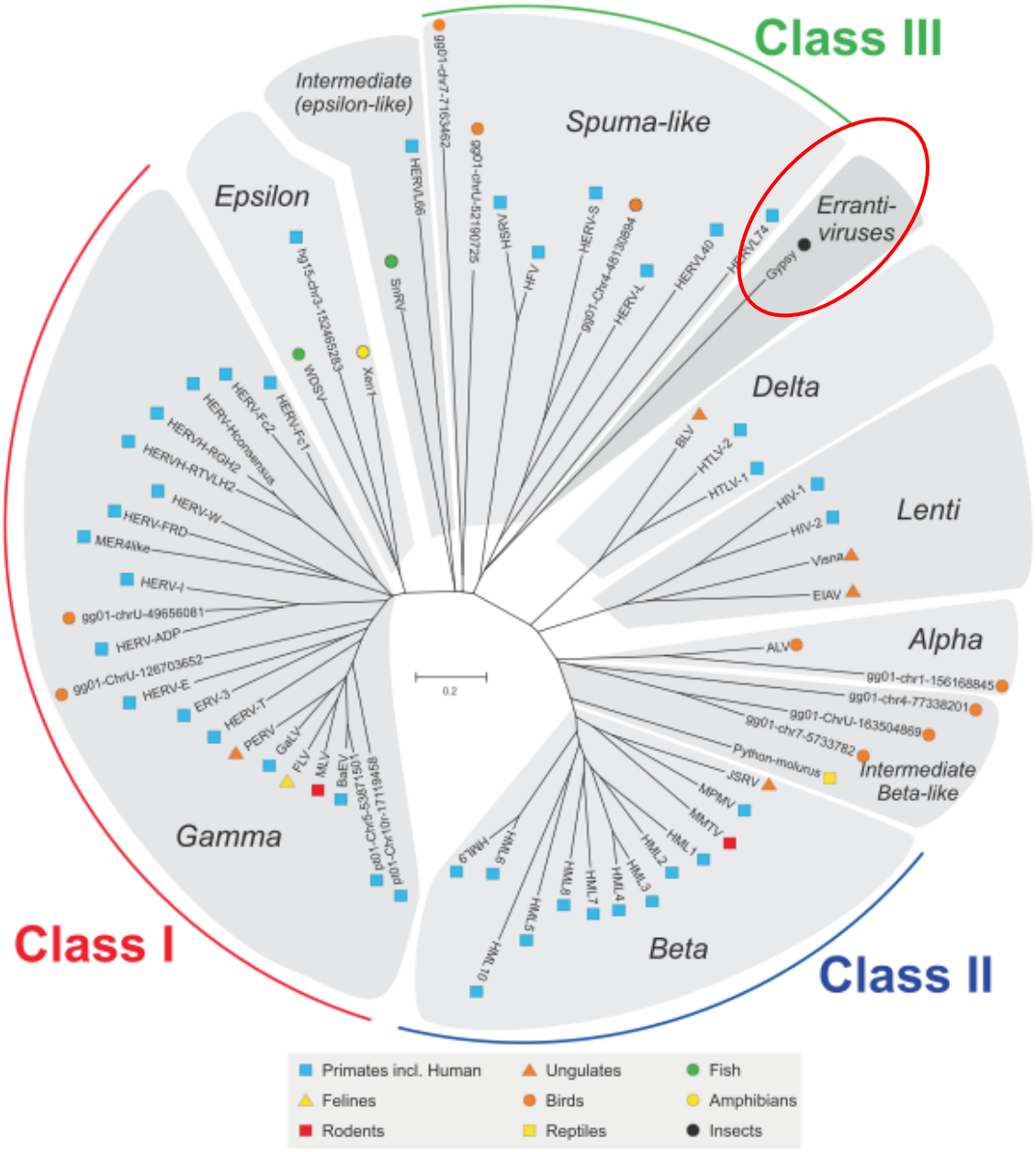


Figure 19 : Dendrogramme non raciné des trois classes d'ERV au sein de la famille des retroviridae réalisé par neighbor joining (NJ) sur la région *pol* (Jern et al., 2005). Sept genres sont représentés (en gras). Les catégories rétrovirales un peu plus vaguement définies sont indiqués dans la périphérie ('intermediate'). Chaque unité taxonomique des différentes espèces hôtes est indiquée par un symbole en couleur. La position des éléments Gypsy est entourée en rouge.

Chapitre VI. Perspectives

- Etudes des rétrotransposons chez les mollusques

Tout d'abord, il faudra finir l'étude des éléments Copia, Gypsy, BEL/Pao, DIRS et PLE dans le groupe des mollusques, notamment l'étude des génomes complets. Cette étude nous permettra de conclure, sur la dynamique des éléments GalEa dans un autre groupe de métazoaire, et également sur la diversité et la dynamique des Gypsy. Ensuite, nous pourrons nous intéresser aux éléments Hydra1-2, qui ont l'air bien présent dans ce taxon d'espèces, pour voir leur proportion au sein des génomes par rapport aux éléments GalEa, et leur part parmi les autres éléments Copia. Nous devrions étudier leur distribution plus largement au sein des métazoaires, voire même au sein des eucaryotes.

- Etudes des éléments Copia chez les champignons

Il reste à finir l'étude en détail des familles de Copia autres que GalEa que l'on retrouve au sein des pezizomycotina. Cela permettra de voir s'il y a présence d'une famille majoritaire de Copia au sein des pezizomycotina.

Une autre question sera de savoir s'il y a une famille de Copia bien présente au sein du phylum des champignons, comme les GalEa au sein des métazoaires. On pourrait réaliser cette étude car nous avons en notre possession beaucoup de génomes complètement séquencés d'ascomycètes (230) et de basidiomycètes (132), et 18 génomes d'autres champignons. Une telle étude, assez peu réalisable à cette échelle des champignons, devrait permettre de voir si l'on retrouve au sein des champignons, en plus des GalEa, d'autres familles décrites chez les métazoaires ou chez les plantes. En effet, une des principales questions est de savoir si une famille d'élément Copia est cantonné à un phylum. On a déjà une idée grâce à la découverte d'éléments GalEa au sein d'algues rouges et des champignons, alors qu'ils ont été décrits en premier chez des Métazoaires. Mais est-ce une exception des GalEa d'être présent au sein de 3 phyla, dû à un incroyable succès de ces éléments? Nous avons déjà un élément de réponse car lors de nos recherches des éléments GalEa au sein des pezizomycotina, nous avons retrouvé des éléments Tork et Oryco éléments décrits chez les plantes ; et Tricopia éléments décrits chez les métazoaires au sein des champignons. Ce qui nous amène effectivement à croire qu'une famille n'est pas restreinte à un phylum.

Pour compléter notre étude sur la comparaison de la dynamique des éléments Copia et Gypsy au sein des eucaryotes, nous aimerions étudier la distribution et la dynamique des éléments Gypsy en dehors des métazoaires. Comme nous avons réalisé une étude des éléments Copia, notamment des GalEa dans le sous-embranchement des pezizomycotina, nous pourrions étudier la distribution et la dynamique des Gypsy au sein de ce sous-embranchement. Cela nous permettrait par la suite de faire une étude comparative de la dynamique de ces éléments à une large échelle, celle d'un phylum: les champignons.

- Origine des éléments GalEa

Nous avons décrit des éléments GalEa chez de nombreux taxons de métazoaires, au sein des champignons, et d'algues rouges. Une question qui se pose est donc l'origine des éléments GalEa au sein des eucaryotes. Une des hypothèses est que les éléments GalEa étaient présents à la base des eucaryotes. Une analyse plus complète des GalEa présents chez les algues rouges est nécessaire pour savoir si cette présence est due à une origine commune avec les GalEa de champignons et de métazoaires, ou s'il est possible que la présence des GalEa au sein des algues rouges soit due à une acquisition secondaire par transferts horizontaux. Si tel est le cas, d'où proviennent ces transferts? En effet, les GalEa des algues rouges peuvent être dû à des transferts d'éléments GalEa de métazoaires ou de champignons. Nous savons qu'il existe des virus d'algues vertes qui infectent également les champignons et de ce fait brassent du matériel génétique entre ces 2 taxons (Blanc et al., 2010). Mais ces virus n'infectent apparemment pas les algues rouges. Peut-être existent-ils d'autres virus ayant le même comportement infectieux qui aurait permis ce transfert d'éléments GalEa. Pour cette étude nous disposons de 4 espèces avec le génome complètement séquencés (*Galdieria sulphuraria, Cyanidioschyzon merolae, Pyropia (Porphyra) yezoensis, Chondrus crispus*), ce qui peut nous donner des idées de diversité des éléments ainsi que leur part au sein des génomes.

- Etude des éléments Copia et Gypsy au sein des plantes

Il serait également intéressant de faire une étude comparative de la dynamique des éléments Gypsy et Copia au sein des plantes. L'étude des Copia est déjà réalisé au sein des métazoaires et en perspective au sein des champignons qui sont 2 phyla des unikontes, donc nous aimerions savoir si ce que nous voyons, se retrouve chez des bikontes, pour comprendre si les dynamiques des éléments sont dû à des effets phylum ou à des effets éléments. De plus nous

savons qu'il y a une variation du nombre de copies de Copia au sein des espèces de plantes avec des espèces avec un très grand nombre de copies et des espèces avec un faible nombre de copies. Les plantes sont très étudiées car elles ont un intérêt agronomique et l'on dispose donc là aussi de nombreux génomes séquencés. Nous aimerions connaitre la part des génomes correspondant à des éléments Copia et Gypsy chez les plantes, cela nous permettrait de faire un parallèle avec 2 études déjà réalisées chez les métazoaires (De la chaux et Wagner, 2011) et chez les champignons (Muszewska et al., 2011). A ma connaissance aucune synthèse sur les éléments Copia et Gypsy n'est disponible à une large échelle chez les plantes.

- Etude d'un autre couple de rétrotransposons, les DIRS et Ngaro au sein des eucaryotes

Indépendamment de ce que j'ai regardé pendant ma thèse, mais en gardant l'idée de la comparaison de dynamique de 2 rétrotransposons proches appartenant à la même superfamille, avec quelques différences structurales ainsi que des différences de séquences, nous pourrions réaliser une étude comparative des éléments DIRS et Ngaro au sein des eucaryotes. Les éléments DIRS ont déjà été étudiés à large échelle au sein des eucaryotes (Piednoël et al., 2011). Les éléments Ngaro qui sont proches des DIRS mais qui malgré tout présentent des différences (terminaisons différentes et absence de la MT) sont assez peu étudié. Ce couple de rétrotransposons a déjà été étudié au sein d'un phylum, celui des champignons, ce qui a permet de découvrir des Ngaro au sein d'espèces de champignons tel que les basidiomycètes et chez une espèce d'ascomycètes (Muszewska et al., 2013). Il reste donc à faire une étude des éléments Ngaro au sein des eucaryotes pour pouvoir comparer ce couple d'éléments au niveau de leur distribution et leur part au sein des génomes (grâce aux nombres de copies) pour ainsi comprendre leur dynamique. Nous pouvons réaliser cette étude grâce au logiciel ReDoSt, qui a été auparavant réalisé pour rechercher des éléments DIRS, que nous allons adapter grâce à des profils pour retrouver les éléments Ngaro. Les profils ont déjà été réalisés et testés.

Toutes ces perspectives sont des pistes à explorer, mais ne sont évidemment pas réalisables dans leur ensemble.

BIBLIOGRAPHIE

Bibliographie

Arkhipova, I.R. (2006). Distribution and phylogeny of Penelope-like elements in eukaryotes. Syst. Biol. *55*, 875–885.

Arkhipova, I.R., Pyatkov, K.I., Meselson, M., and Evgen'ev, M.B. (2003). Retroelements containing introns in diverse invertebrate taxa. Nat. Genet. *33*, 123–124.

Blanc, G., Duncan, G., Agarkova, I., Borodovsky, M., Gurnon, J., Kuo, A., Lindquist, E., Lucas, S., Pangilinan, J., Polle, J., et al. (2010). The Chlorella variabilis NC64A genome reveals adaptation to photosymbiosis, coevolution with viruses, and cryptic sex. Plant Cell *22*, 2943–2955.

Boeke, J.D., and Devine, S.E. (1998). Yeast retrotransposons: finding a nice quiet neighborhood. Cell *93*, 1087–1089.

Bonnivard, E., Catrice, O., Ravaux, J., Brown, S.C., and Higuet, D. (2009). Survey of genome size in 28 hydrothermal vent species covering 10 families. Genome Natl. Res. Counc. Can. Génome Cons. Natl. Rech. Can. *52*, 524–536.

Bui, Q.-T., Delaurière, L., Casse, N., Nicolas, V., Laulier, M., and Chénais, B. (2007). Molecular characterization and phylogenetic position of a new mariner-like element in the coastal crab, Pachygrapsus marmoratus. Gene *396*, 248–256.

Bui, Q.-T., Casse, N., Leignel, V., Nicolas, V., and Chénais, B. (2008). Widespread occurence of mariner transposons in coastal crabs. Mol. Phylogenet. Evol. *47*, 1181–1189.

Bureau, T.E., and Wessler, S.R. (1992). Tourist: a large family of small inverted repeat elements frequently associated with maize genes. Plant Cell *4*, 1283–1294.

Le Calvez, T., Burgaud, G., Mahé, S., Barbier, G., and Vandenkoornhuyse, P. (2009). Fungal diversity in deep-sea hydrothermal ecosystems. Appl. Environ. Microbiol. *75*, 6415–6421.

Cappello, J., Handelsman, K., and Lodish, H.F. (1985). Sequence of Dictyostelium DIRS-1: an apparent retrotransposon with inverted terminal repeats and an internal circle junction sequence. Cell *43*, 105–115.

Capy, P., Langin, T., Higuet, D., Maurer, P., and Bazin, C. (1997). Do the integrases of LTR-retrotransposons and class II element transposases have a common ancestor? Genetica *100*, 63–72.

Casse, N., Bui, Q.T., Nicolas, V., Renault, S., Bigot, Y., and Laulier, M. (2006). Species sympatry and horizontal transfers of Mariner transposons in marine crustacean genomes. Mol. Phylogenet. Evol. *40*, 609–619.

Cavalli, G., and Paro, R. (1998). Chromo-domain proteins: linking chromatin structure to epigenetic regulation. Curr. Opin. Cell Biol. *10*, 354–360.

De la Chaux, N., and Wagner, A. (2011). BEL/Pao retrotransposons in metazoan genomes.

BMC Evol. Biol. *11*, 154.

Chesney, M.A., Kidd, A.R., and Kimble, J. (2006). gon-14 functions with class B and class C synthetic multivulva genes to control larval growth in Caenorhabditis elegans. Genetics *172*, 915–928.

Clouaire, T., Roussigne, M., Ecochard, V., Mathe, C., Amalric, F., and Girard, J.-P. (2005). The THAP domain of THAP1 is a large C2CH module with zinc-dependent sequence-specific DNA-binding activity. Proc. Natl. Acad. Sci. U. S. A. *102*, 6907–6912.

Devos, K.M., Brown, J.K.M., and Bennetzen, J.L. (2002). Genome size reduction through illegitimate recombination counteracts genome expansion in Arabidopsis. Genome Res. *12*, 1075–1079.

Dewannieux, M., Esnault, C., and Heidmann, T. (2003). LINE-mediated retrotransposition of marked Alu sequences. Nat. Genet. *35*, 41–48.

Dhillon, B., Gill, N., Hamelin, R.C., and Goodwin, S.B. (2014). The landscape of transposable elements in the finished genome of the fungal wheat pathogen Mycosphaerella graminicola. BMC Genomics *15*, 1132.

Eickbush, T.H., and Jamburuthugoda, V.K. (2008). The diversity of retrotransposons and the properties of their reverse transcriptases. Virus Res. *134*, 221–234.

Ellinghaus, D., Kurtz, S., and Willhoeft, U. (2008). LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC Bioinformatics *9*, 18.

Evgen'ev, M.B., and Arkhipova, I.R. (2005). Penelope-like elements--a new class of retroelements: distribution, function and possible evolutionary significance. Cytogenet. Genome Res. *110*, 510–521.

Flutre, T., Duprat, E., Feuillet, C., and Quesneville, H. (2011). Considering Transposable Element Diversification in De Novo Annotation Approaches. PLoS ONE *6*, e16526.

Gao, X., Hou, Y., Ebina, H., Levin, H.L., and Voytas, D.F. (2008). Chromodomains direct integration of retrotransposons to heterochromatin. Genome Res. *18*, 359–369.

Goodwin, T.J., and Poulter, R.T. (2001). The DIRS1 group of retrotransposons. Mol. Biol. Evol. *18*, 2067–2082.

Goodwin, T.J.D., and Poulter, R.T.M. (2002). A group of deuterostome Ty3/ gypsy-like retrotransposons with Ty1/ copia-like pol-domain orders. Mol. Genet. Genomics MGG *267*, 481–491.

Goodwin, T.J.D., and Poulter, R.T.M. (2004). A new group of tyrosine recombinase-encoding retrotransposons. Mol. Biol. Evol. *21*, 746–759.

Goodwin, T.J.D., Butler, M.I., and Poulter, R.T.M. (2003). Cryptons: a group of tyrosine-recombinase-encoding DNA transposons from pathogenic fungi. Microbiol. Read.

Engl. *149*, 3099–3109.

Gregory, T.R. (2005). The C-value enigma in plants and animals: a review of parallels and an appeal for partnership. Ann. Bot. *95*, 133–146.

Hall, A.E., Kettler, G.C., and Preuss, D. (2006). Dynamic evolution at pericentromeres. Genome Res. *16*, 355–364.

Huang, S.-W., Lin, Y.-Y., You, E.-M., Liu, T.-T., Shu, H.-Y., Wu, K.-M., Tsai, S.-F., Lo, C.-F., Kou, G.-H., Ma, G.-C., et al. (2011). Fosmid library end sequencing reveals a rarely known genome structure of marine shrimp Penaeus monodon. BMC Genomics *12*, 242.

Jern, P., Sperber, G.O., and Blomberg, J. (2005). Use of endogenous retroviral sequences (ERVs) and structural markers for retroviral phylogenetic inference and taxonomy. Retrovirology *2*, 50.

Joly-Lopez, Z., and Bureau, T.E. (2014). Diversity and evolution of transposable elements in Arabidopsis. Chromosome Res. Int. J. Mol. Supramol. Evol. Asp. Chromosome Biol. *22*, 203–216.

Jordan, I.K., Rogozin, I.B., Glazko, G.V., and Koonin, E.V. (2003). Origin of a substantial fraction of human regulatory sequences from transposable elements. Trends Genet. TIG *19*, 68–72.

Kajikawa, M., and Okada, N. (2002). LINEs mobilize SINEs in the eel through a shared 3' sequence. Cell *111*, 433–444.

Kaminker, J.S., Bergman, C.M., Kronmiller, B., Carlson, J., Svirskas, R., Patel, S., Frise, E., Wheeler, D.A., Lewis, S.E., Rubin, G.M., et al. (2002). The transposable elements of the Drosophila melanogaster euchromatin: a genomics perspective. Genome Biol. *3*, RESEARCH0084.

Kapitonov, V.V., and Jurka, J. (2005). RAG1 core and V(D)J recombination signal sequences were derived from Transib transposons. PLoS Biol. *3*, e181.

Ke, N., Irwin, P.A., and Voytas, D.F. (1997). The pheromone response pathway activates transcription of Ty5 retrotransposons located within silent chromatin of Saccharomyces cerevisiae. EMBO J. *16*, 6272–6280.

Kim, J.M., Vanguri, S., Boeke, J.D., Gabriel, A., and Voytas, D.F. (1998). Transposable elements and genome organization: a comprehensive survey of retrotransposons revealed by the complete Saccharomyces cerevisiae genome sequence. Genome Res. *8*, 464–478.

Koonin, E.V., Zhou, S., and Lucchesi, J.C. (1995). The chromo superfamily: new members, duplication of the chromo domain and possible role in delivering transcription regulators to chromatin. Nucleic Acids Res. *23*, 4229–4233.

Kramerov, D.A., and Vassetzky, N.S. (2005). Short retroposons in eukaryotic genomes. Int. Rev. Cytol. *247*, 165–221.

Leebonoi, W., Sukthaworn, S., Panyim, S., and Udomkit, A. (2015). A novel gonad-specific Argonaute 4 serves as a defense against transposons in the black tiger shrimp Penaeus monodon. Fish Shellfish Immunol. *42*, 280–288.

Leitch, A.R., and Leitch, I.J. (2008). Genomic plasticity and the diversity of polyploid plants. Science *320*, 481–483.

Lim, J.K., and Simmons, M.J. (1994). Gross chromosome rearrangements mediated by transposable elements in Drosophila melanogaster. BioEssays News Rev. Mol. Cell. Dev. Biol. *16*, 269–275.

Llorens, C., Muñoz-Pomer, A., Bernad, L., Botella, H., and Moya, A. (2009). Network dynamics of eukaryotic LTR retroelements beyond phylogenetic trees. Biol. Direct *4*, 41.

Lorenzi, H.A., Robledo, G., and Levin, M.J. (2006). The VIPER elements of trypanosomes constitute a novel group of tyrosine recombinase-enconding retrotransposons. Mol. Biochem. Parasitol. *145*, 184–194.

Macfarlan, T., Kutney, S., Altman, B., Montross, R., Yu, J., and Chakravarti, D. (2005). Human THAP7 is a chromatin-associated, histone tail-binding protein that represses transcription via recruitment of HDAC3 and nuclear hormone receptor corepressor. J. Biol. Chem. *280*, 7346–7358.

Mackay, T.F.C., Richards, S., Stone, E.A., Barbadilla, A., Ayroles, J.F., Zhu, D., Casillas, S., Han, Y., Magwire, M.M., Cridland, J.M., et al. (2012). The Drosophila melanogaster Genetic Reference Panel. Nature *482*, 173–178.

Malik, H.S., and Eickbush, T.H. (1999). Modular evolution of the integrase domain in the Ty3/Gypsy class of LTR retrotransposons. J. Virol. *73*, 5186–5190.

Maumus, F., Rabinowicz, P., Bowler, C., and Rivarola, M. (2011). Stemming epigenetics in marine stramenopiles. Curr. Genomics *12*, 357–370.

Mikkelsen, T.S., Wakefield, M.J., Aken, B., Amemiya, C.T., Chang, J.L., Duke, S., Garber, M., Gentles, A.J., Goodstadt, L., Heger, A., et al. (2007). Genome of the marsupial Monodelphis domestica reveals innovation in non-coding sequences. Nature *447*, 167–177.

Muszewska, A., Hoffman-Sommer, M., and Grynberg, M. (2011). LTR retrotransposons in fungi. PloS One *6*, e29425.

Muszewska, A., Steczkiewicz, K., and Ginalski, K. (2013). DIRS and Ngaro Retrotransposons in Fungi. PLoS ONE *8*, e76319.

Navarro-Quezada, A., and Schoen, D.J. (2002). Sequence evolution and copy number of Ty1-copia retrotransposons in diverse plant genomes. Proc. Natl. Acad. Sci. U. S. A. *99*, 268–273.

Peddigari, S., Zhang, W., Sakai, M., Takechi, K., Takano, H., and Takio, S. (2008). A copia-like retrotransposon gene encoding gypsy-like integrase in a red alga, Porphyra

yezoensis. J. Mol. Evol. *66*, 72–79.

Piednoël, M., and Bonnivard, E. (2009). DIRS1-like retrotransposons are widely distributed among Decapoda and are particularly present in hydrothermal vent organisms. BMC Evol. Biol. *9*, 86.

Piednoël, M., Gonçalves, I.R., Higuet, D., and Bonnivard, E. (2011). Eukaryote DIRS1-like retrotransposons: an overview. BMC Genomics *12*, 621.

Piednoël, M., Donnart, T., Esnault, C., Graça, P., Higuet, D., and Bonnivard, E. (2013). LTR-Retrotransposons in R. exoculata and Other Crustaceans: The Outstanding Success of GalEa-Like Copia Elements. PLoS ONE *8*, e57675.

Poulter, R.T.M., and Goodwin, T.J.D. (2005). DIRS-1 and the other tyrosine recombinase retrotransposons. Cytogenet. Genome Res. *110*, 575–588.

Quesneville, H., Nouaud, D., and Anxolabehere, D. (2005). Recurrent recruitment of the THAP DNA-binding domain and molecular domestication of the P-transposable element. Mol. Biol. Evol. *22*, 741–746.

Rho, M., Schaack, S., Gao, X., Kim, S., Lynch, M., and Tang, H. (2010). LTR retroelements in the genome of Daphnia pulex. BMC Genomics *11*, 425.

Roussigne, M., Kossida, S., Lavigne, A.-C., Clouaire, T., Ecochard, V., Glories, A., Amalric, F., and Girard, J.-P. (2003). The THAP domain: a novel protein motif with similarity to the DNA-binding domain of P element transposase. Trends Biochem. Sci. *28*, 66–69.

SanMiguel, P., Tikhonov, A., Jin, Y.K., Motchoulskaia, N., Zakharov, D., Melake-Berhan, A., Springer, P.S., Edwards, K.J., Lee, M., Avramova, Z., et al. (1996). Nested retrotransposons in the intergenic regions of the maize genome. Science *274*, 765–768.

SanMiguel, P., Gaut, B.S., Tikhonov, A., Nakajima, Y., and Bennetzen, J.L. (1998). The paleontology of intergene retrotransposons of maize. Nat. Genet. *20*, 43–45.

Schaack, S., Choi, E., Lynch, M., and Pritham, E.J. (2010). DNA transposons and the role of recombination in mutation accumulation in Daphnia pulex. Genome Biol. *11*, R46.

Sinzelle, L., Izsvák, Z., and Ivics, Z. (2009). Molecular domestication of transposable elements: from detrimental parasites to useful host genes. Cell. Mol. Life Sci. CMLS *66*, 1073–1093.

Slotte, T., Hazzouri, K.M., Ågren, J.A., Koenig, D., Maumus, F., Guo, Y.-L., Steige, K., Platts, A.E., Escobar, J.S., Newman, L.K., et al. (2013). The Capsella rubella genome and the genomic consequences of rapid mating system evolution. Nat. Genet. *45*, 831–835.

Tarailo-Graovac, M., and Chen, N. (2009). Using RepeatMasker to identify repetitive elements in genomic sequences. Curr. Protoc. Bioinforma. Ed. Board Andreas Baxevanis Al *Chapter 4*, Unit 4.10.

Terrat, Y., Bonnivard, E., and Higuet, D. (2008). GalEa retrotransposons from galatheid squat lobsters (Decapoda, Anomura) define a new clade of Ty1/copia-like elements restricted to aquatic species. Mol. Genet. Genomics MGG *279*, 63–73.

Thomas, C.A. (1971). The Genetic Organization of Chromosomes. Annu. Rev. Genet. *5*, 237–256.

Varmus, H., and Brown, P. (1989). Retroviruses. In Mobile DNA, (Am Soc Microbiol.),.

De la Vega, E., Degnan, B.M., Hall, M.R., and Wilson, K.J. (2007). Differential expression of immune-related genes and transposable elements in black tiger shrimp (Penaeus monodon) exposed to a range of environmental stressors. Fish Shellfish Immunol. *23*, 1072–1088.

Villasante, A., Abad, J.P., Planelló, R., Méndez-Lago, M., Celniker, S.E., and de Pablos, B. (2007). Drosophila telomeric retrotransposons derived from an ancestral element that was recruited to replace telomerase. Genome Res. *17*, 1909–1918.

Volff, J.N., Körting, C., and Schartl, M. (2000). Multiple lineages of the non-LTR retrotransposon Rex1 with varying success in invading fish genomes. Mol. Biol. Evol. *17*, 1673–1684.

Wicker, T., and Keller, B. (2007). Genome-wide comparative analysis of copia retrotransposons in Triticeae, rice, and Arabidopsis reveals conserved ancient evolutionary lineages and distinct dynamics of individual copia families. Genome Res. *17*, 1072–1081.

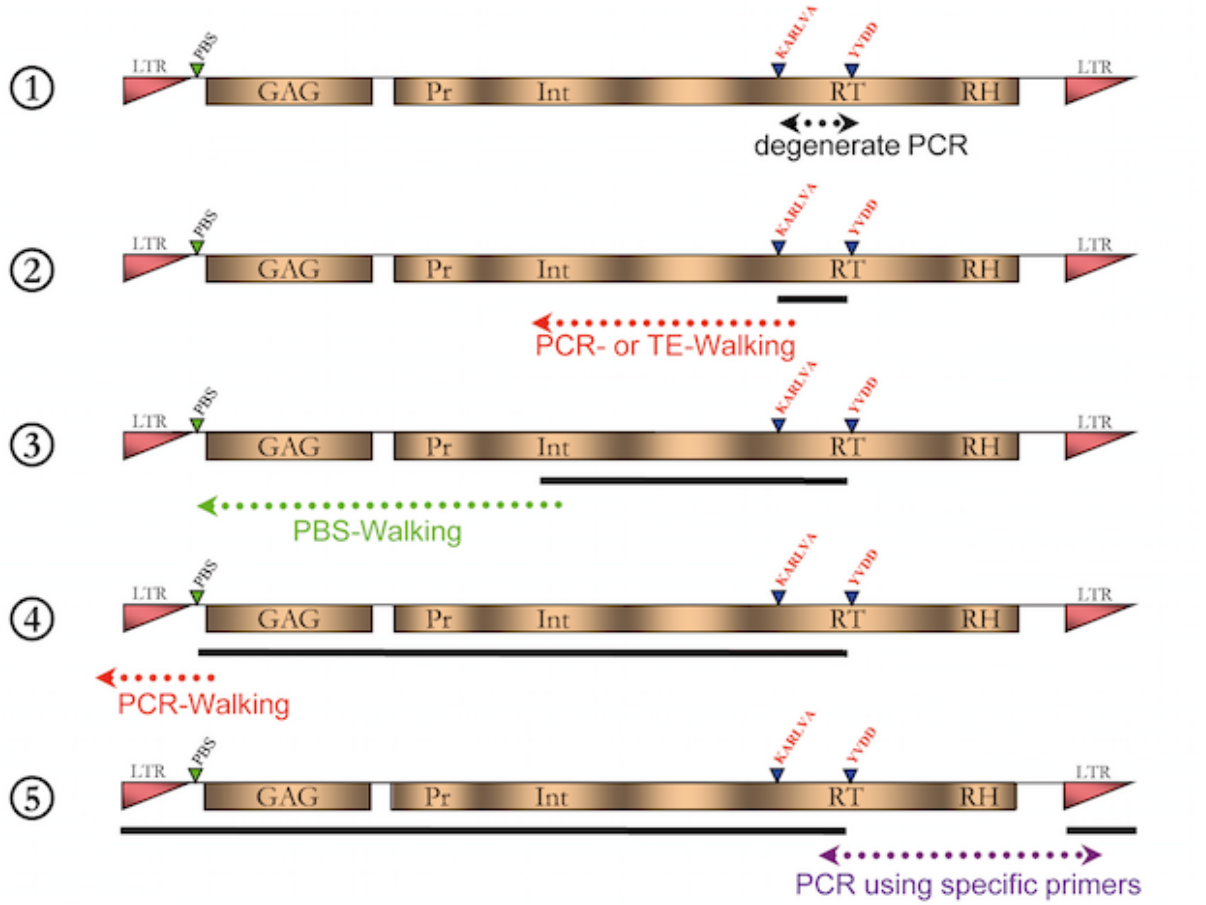
Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J.L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., et al. (2007). A unified classification system for eukaryotic transposable elements. Nat. Rev. Genet. *8*, 973–982.

Xiong, Y., and Eickbush, T.H. (1990). Origin and evolution of retroelements based upon their reverse transcriptase sequences. EMBO J. *9*, 3353–3362.

Zuker, C., Cappello, J., Lodish, H.F., George, P., and Chung, S. (1984). Dictyostelium transposable element DIRS-1 has 350-base-pair inverted terminal repeats that contain a heat shock promoter. Proc. Natl. Acad. Sci. U. S. A. *81*, 2660–2664.

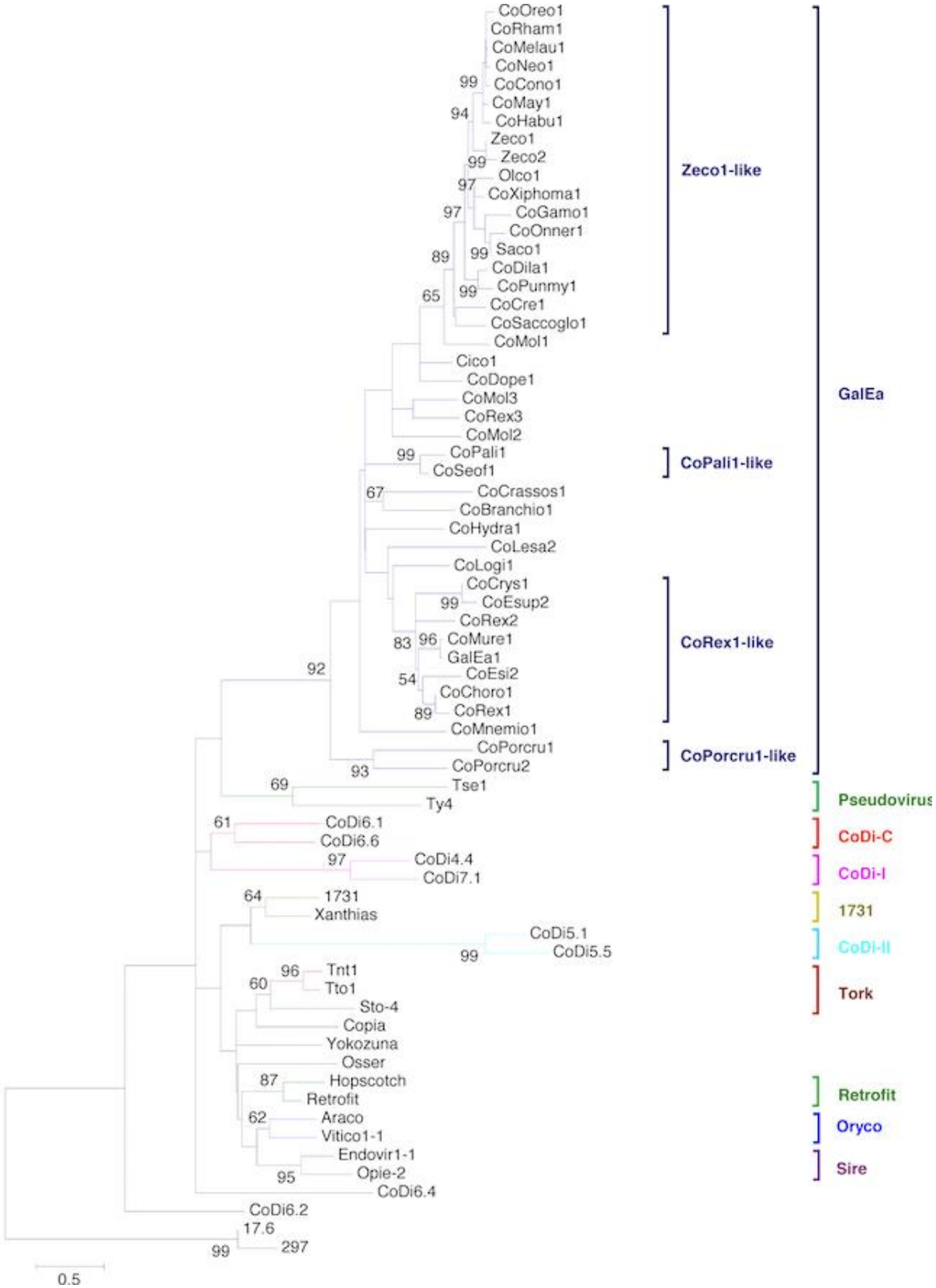
ANNEXES

Annexes 1: Données supplémentaires de l'article : LTR retrotransposons in crustaceans



Figure_S1

Characterization strategy of full-length LTR-retrotransposons. A copia retrotransposon is used as example. For each of the five steps, the known part of the element is represented by a full line whereas the walking part is indicated by colored dotted arrow: red, PCR or TE Walking; green, PBS Walking; purple: PCR using specific primers. The conserved domains used to design the degenerate primers and the PBS sequences are represented by blue and green triangles, respectively.


CoOreo1
CoRham1
CoMelau1
CoNeo1
99
CoCono1
CoMay1
94
CoHabu1
Zeco1
99
Zeco2
97
Olco1
CoXiphoma1
97
CoGamo1
CoOnner1
89
99
Saco1
CoDila1
65
99
CoPunmy1
CoCre1
CoSaccoglo1
CoMol1
Cico1
CoDope1
CoMol3
CoRex3
CoMol2
99
CoPali1
CoSeof1
67
CoCrassos1
CoBranchio1
CoHydra1
CoLesa2
CoLogi1
CoCrys1
99
CoEsup2
CoRex2
83
96
CoMure1
GalEa1
92
54
CoEsi2
CoChoro1
89
CoRex1
CoMnemio1
CoPorcru1
93
CoPorcru2
69
Tse1
Ty4
61
CoDi6.1
CoDi6.6
97
CoDi4.4
CoDi7.1
64
1731
Xanthias
CoDi5.1
99
CoDi5.5
96
Tnt1
60
Tto1
Sto-4
Copia
Yokozuna
Osser
87
Hopscotch
Retrofit
62
Araco
Vitico1-1
Endovir1-1
95
Opie-2
CoDi6.4
CoDi6.2
17.6
99
297
0.5
Zeco1-like
CoPali1-like
CoRex1-like
CoPorcru1-like
GalEa
Pseudovirus
CoDi-C
CoDi-I
1731
CoDi-II
Tork
Retrofit
Oryco
Sire

Figure_S2

Phylogenetic relationships among GalEa-like retrotransposons inferred from Neighbor-Joining analysis of RT/RH amino acid sequences. Statistical support (>50%) comes from non parametric bootstrapping using 100 replicates. Two to three representative elements of the other Copia clades are also included to the phylogeny. Gypsy sequences were used as outgroup.

PCR Approach

Species		Sequences	Element	Length (bp)	Accession number	PCR approaches	Primer Forward	Primer Reverse
	Gypsy							
Rimicaris exoculata		GyRex1a	GyRex1	446	HF548771	PCR walking	WAP2	GyRex1_92-
		GyRex1b	GyRex1	791	HF548772	PCR walking	WAP2	GyRex1_160-
		GyRex1c	GyRex1	1350	HF548773	PCR walking	WAP2	GyRex1_239-
		GyRex1d	GyRex1	651	HF548774	PCR walking	WAP2	GyRex1_109-
		GyRex1e	GyRex1	275	HF548775	Degenerate primers	GD1	GD2
		GyRex1f	GyRex1	866	HF548776	TE walking	GyRex1_198+	GD3
		GyRex1g	GyRex1	1111	HF548777	PCR walking	GyRex1_1296+	WAP2
		GyRex1h	GyRex1	564	HF548778	PCR walking	GyRex1_2233+	WAP2
		GyRex2a	GyRex2	1438	HF548779	Specific primers	GyRex2_3658+	GyRex2_110-
		GyRex2b	GyRex2	500	HF548780	PCR walking	WAP2	GyRex2_207-
		GyRex2c	GyRex2	1075	HF548781	PBS walking	GyRex2 PBS+	GyRex2_60-
		GyRex2d	GyRex2	858	HF548782	TE walking	GD0	GyRex2_56-
		GyRex2e	GyRex2	524	HF548783	Degenerate primers	GD1	GD2
		GyRex2f	GyRex2	420	HF548784	TE walking	GyRex2_360+	GD3
		GyRex2g	GyRex2	482	HF548785	PCR walking	GyRex2_1663+	WAP2
		GyRex2h	GyRex2	413	HF548786	PCR walking	GyRex2_2070+	WAP2
		GyRex2i	GyRex2	1221	HF548787	Specific primers	GyRex2_3658+	GyRex2_250-
		GyRex2j	GyRex2	684	HF548788	Specific primers	GyRex2_3658+	GyRex2_110-
		GyRex2k	GyRex2	729	HF548789	PCR walking	GyRex2_5033+	WAP2
		GyRex3a	GyRex3	233	HF548790	PCR walking	WAP2	GyRex3_74-
		GyRex3b	GyRex3	472	HF548791	PCR walking	WAP2	GyRex3_64-
		GyRex3c	GyRex3	515	HF548792	Degenerate primers	GD1	GD2
		GyRex3d	GyRex3	423	HF548793	TE walking	GyRex3_422+	GD3
		GyRex3e	GyRex3	426	HF548794	PCR walking	GyRex3_1199+	WAP2
		GyRex3f	GyRex3	1086	HF548795	PCR walking	GyRex3 1491+	WAP2
		GyRex4a	GyRex4	622	HF548796	Degenerate primers	GD3	GD3
		GyRex4b	GyRex4	546	HF548797	TE walking	GD1	GyRex5_270-
		GyRex5a	GyRex5	671	HF548798	PCR walking	-	WAP2
		GyRex6a	GyRex6	1160	HF548799	PCR walking	WAP2	Gyrex2_60-
Agononida laurentae		GyAla1a	GyAla1	526	HF548816	Degenerate primers	GD1bis	rcDD4soft
		GyAla2a	GyAla2	592	HF548817	TE walking	GD1	GyAla2_169-
		GyAla2b	GyAla2	619	HF548818	Degenerate primers	GD3	GD3
Alvinocaris markensis		GyAlma2a	GyAlma2	999	HF548823	Specific primers	GyRex2_769+	GyRex2_1767-
		GyAlma2b	GyAlma2	999	HF548824	Specific primers	GyRex2_769+	GyRex2_1767-
Bythograea thermydron		GyBy1a	GyBy1	524	HF548800	Degenerate primers	GD1	GD2
		GyBy1b	GyBy1	388	HF548801	TE walking	GyBy1_452+	GD3
Chorocaris chacei		GyChoro1a	GyChoro1	525	HF548810	Degenerate primers	GD1	GD2
		GyChoro1b	GyChoro1	357	HF548811	TE walking	GyChoro1_510+	GD3
		GyChoro2a	GyChoro2	340	HF548812	TE walking	GD1	GyChoro2_117-
		GyChoro2b	GyChoro2	620	HF548813	Degenerate primers	GD3	GD3
Litopenaeus vanameii		GyLiva6b	GyLiva6	830	HF548815	TE walking	GyLiva6_463+	GD3
Maja squinado		GyMaja1a	GyMaja1	569	HF548821	Degenerate primers	GD1	GD2
		GyMaja2a	GyMaja2	623	HF548822	Degenerate primers	GD3	GD3
Mirocaris fortunata		GyMiro2a	GyMiro2	950	HF548814	Specific primers	GyRex2_769+	GyRex2_1767-
Munidopsis recta		GyMur1a	GyMur1	862	HF548809	Specific primer	GyEa1+	GyEa1-
Necora puber		GyNeco1a	GyNeco1	521	HF548802	Degenerate primers	GD1	GD2
		GyNeco1b	GyNeco1	494	HF548803	TE walking	GyNeco1_353+	GD3
Nematocarcinus burukovskii		GyNem1a	GyNem1	526	HF548804	Degenerate primers	GD1bis	rcDD4soft
		GyNem1b	GyNem1	515	HF548805	TE walking	GyNem1_331+	GD3
Orconectes limosus		GyOrli1a	GyOrli1	866	HF548806	Degenerate primers	GD1	GD3
Penaeus monodon		GyPemo1b	GyPemo1	730	HF548819	TE walking	GyPemo1_294+	GD3
		GyPemo2b	GyPemo2	430	HF548820	TE walking	GyPemo2_329+	GD3
Sacculina carcini		GySac1a	GySac1	887	HF548807	Degenerate primers	GD1	GD3
		GySac2a	GySac2	895	HF548808	Degenerate primers	GD1	GD3

	Copia							
Rimicaris exoculata		CoRex1a	CoRex1	357	HF548730	PCR walking	WAP2	CoRex1 169-
		CoRex1b	CoRex1	2032	HF548731	PBS walking	PBSmet	CoRex1_136-
		CoRex1c	CoRex1	535	HF548732	PCR walking	WAP2	CoRex1_49-
		CoRex1d	CoRex1	803	HF548733	PCR walking	WAP2	CoRex1_106-
		CoRex1e	CoRex1	367	HF548734	Degenerate primers	CD1	CD2
		CoRex1f	CoRex1	907	HF548735	TE walking	CoRex1_356+	CD3
		CoRex1g	CoRex1	339	HF548736	PCR walking	CoRex1_4396+	WAP2
		CoRex1h	CoRex1	259	HF548737	Specific primer	CoRex1_5500+	CoRex1-LTR
		CoRex2a	CoRex2	873	HF548738	PCR walking	WAP2	CoRex2_260-
		CoRex2b	CoRex2	2075	HF548739	PBS walking	PBSmet	CoRex2_210-
		CoRex2c	CoRex2	993	HF548740	PCR walking	WAP2	CoRex2_44-
		CoRex2d	CoRex2	289	HF548741	PCR walking	WAP2	CoRex2_75-
		CoRex2e	CoRex2	335	HF548742	Degenerate primers	CD1	CD2
		CoRex2f	CoRex2	1135	HF548743	TE walking	CoRex2_285+	WAP2
		CoRex2g	CoRex2	606	HF548744	TE walking	CoRex2 2368+	WAP2
		CoRex3a	CoRex3	1770	HF548745	PBS walking	PBSmet	CoRex3_213-
		CoRex3b	CoRex3	1403	HF548746	PCR walking	WAP2	CoRex3 47-
		CoRex3c	CoRex3	637	HF548747	PCR walking	WAP2	CoRex3_96-
		CoRex3d	CoRex3	370	HF548748	Degenerate primers	CD1	CD2
		CoRex3e	CoRex3	420	HF548749	PCR walking	CoRex3_274+	WAP2
		CoRex1-cDNAa	CoRex1	319	HF548722	Specific primers	CoRex1_56+	CoRex1_365-
		CoRex1-cDNAb	CoRex1	319	HF548723	Specific primers	CoRex1_56+	CoRex1_365-
		CoRex1-cDNAc	CoRex1	319	HF548724	Specific primers	CoRex1_56+	CoRex1_365-
		CoRex2 cDNAa	CoRex2	297	HF548725	Specific primers	CoRex2_110+	CoRex2_369-
		CoRex2-cDNAb	CoRex2	297	HF548726	Specific primers	CoRex2_110+	CoRex2_369-
		CoRex2-cDNAc	CoRex2	297	HF548727	Specific primers	CoRex2_110+	CoRex2_369-
		CoRex2-cDNAd	CoRex3	289	HF548728	Specific primers	CoRex2_33+	CoRex2_311-
		CoRex2-cDNAe	CoRex4	288	HF548729	Specific primers	CoRex2_33+	CoRex2_311-
Alvinocaris lusca		CoAllu1a	CoAllu1	1118	HF548756	Specific primers	CoRex1_56+	CoRex1_4396-
		CoAllu1b	CoAllu1	1116	HF548757	Specific primers	CoRex1_56+	CoRex1_4396-
		CoAllu2a	CoAllu2	1247	HF548758	Specific primers	CoRex2_3022+	CoRex2_1191-
		CoAllu2b	CoAllu2	1246	HF548759	Specific primers	CoRex2_3022+	CoRex2_1191-
Alvinocaris markensis		CoAlma1a	CoAlma1	907	HF548760	Specific primers	CoRex1_49+	CoRex1_237-
		CoAlma1b	CoAlma1	898	HF548761	Specific primers	CoRex1_49+	CoRex1_237-
		CoAlma2a	CoAlma2	1082	HF548762	Specific primers	CoRex2_3022+	CoRex2_1191-
		CoAlma3a	CoAlma3	280	HF548763	Specific primers	CoRex3_80+	CoRex3_372-
		CoAlma3b	CoAlma3	308	HF548764	Specific primers	CoRex3_80+	CoRex3_372-
Chorocaris chacei		CoChoro1a	CoChoro1	369	HF548754	Degenerate primers	CD1	CD2
		CoChoro1b	CoChoro1	687	HF548755	TE walking	CoChoro1_244+	CD5GalEa
Maja squinado		CoMaja1a	CoMaja1	2101	HF548750	Degenerate primers	CD1	CD2
		CoMaja1b	CoMaja1	401	HF548751	PBS walking	PBS-Met	CoMaja1_117-
Mirocaris fortunata		CoMiro2a	CoMiro2	1235	HF548765	Specific primers	CoRex2_3022+	CoRex2_1191-
		CoMiro2b	CoMiro2	1245	HF548766	Specific primers	CoRex2_3022+	CoRex2_1191-
		CoMiro3a	CoMiro3	309	HF548767	Specific primers	CoRex3_80+	CoRex3_372-
Munidopsis recta		CoMur1a	CoMur1	350	HF548752	Degenerate primers	CD1	CD2
		CoMur1b	CoMur1	816	HF548753	TE walking	CoMure1_99+	CD5GalEa
	LINE							
Rimicaris exoculata		Lirex1a	LiRex1	354	HF548769	PCR walking	WAP2	GyRex3_178-
		Lirex2a	LiRex2	592	HF548770	PCR walking	GyRex2_2070+	WAP2
	Transposon							
Rimicaris exoculata		T-rex1a	T-Rex1	675	HF548768	Degenerate primers	CD1	CD2

In silico Approach

GyPemo1a	GyPemo1	415	*Penaeus monodon*	MGID126613 est_p_monodon329	http://www.marinegenomics.org/
GyPemo2a	GyPemo2	444	*Penaeus monodon*	MGID126443 est_p_monodon136	
GyLiva1a	GyLiva1	837	*Litopenaeus vanameii*	MGID352304 CK591125.1	
GyLiva2a	GyLiva2	651	*Litopenaeus vanameii*	MGID410110 LV_NC_RA12J07f	
GyLiva3a	GyLiva3	787	*Litopenaeus vanameii*	MGID410111 LV_NC_RA12J07r	
GyLiva 4a	GyLiva4	471	*Litopenaeus vanameii*	MGID408910 LV_NC_RA09J01f	
GyLiva 5a	GyLiva5	453	*Litopenaeus vanameii*	MGID410477 LV_NC_RA13E21f	
GyLiva6a	GyLiva6	695	*Litopenaeus vanameii*	MGID353368	
GyLiva7a	GyLiva7	291	*Litopenaeus vanameii*	MGID372863 LV_HC_RA060G07r	
GyEcrys1a	GyEcrys1	2261	*Euphausia crystallorophias*	comp152260_c0_seq1	
GyHoam1a	GyHoam1	860	*Homarus americanus*	MGID100769 est_h_americanus2358	
GyHoam2a	GyHoam2	722	*Homarus americanus*	MGID98322 est_h_americanus5401	
GyHoam3a	GyHoam3	870	*Homarus americanus*	MGID98312 est_h_americanus5412	
GyPaha1a	GyPaha1	1562	*Parhyale hawaiensis*	741_1-4_AFHB	http://genome.jgi-psf.org/parha/parha.download.ftp.html
GyPaha2a	GyPaha2	732	*Parhyale hawaiensis*	741_0_AFHB26204.b2	
GyPaha3a	GyPaha3	800	*Parhyale hawaiensis*	5864_0_AFHB13272.b2_AFHB	
GyPaha4a	GyPaha4	689	*Parhyale hawaiensis*	10574_0_AFHB4423.x1_AFHB	
GyPaha5a	GyPaha5	722	*Parhyale hawaiensis*	0_0_AFHB14058.b2_AFHB	
GyPaha6a	GyPaha6	673	*Parhyale hawaiensis*	10480_0_AFHB40998.g3_AFHB	
GyPaha7a	GyPaha7	733	*Parhyale hawaiensis*	3094_3_AFHB	
GyPaha8a	GyPaha8	688	*Parhyale hawaiensis*	3697_1_AFHB	
CoEcrys1a	CoEcrys1	1119	*Euphausia crystallorophias*	comp11338_c0_seq1	
CoEsup1a	CoEsup1	1509	*Euphausia superba*	contig00846	
CoEsup2a	CoEsup2	913	*Euphausia superba*	contig02254	
CoEsup3a	CoEsup3	1030	*Euphausia superba*	contig19342	
CoEsup4a	CoEsup4	329	*Euphausia superba*	gb|ES543757.1|	
CoPeci1a	CoPeci1	1418	*Petrolisthes cinctipes*	Contig5495	
CoPeci2a	CoPeci2	1405	*Petrolisthes cinctipes*	Contig574	
CoPeci3a	CoPeci3	636	*Petrolisthes cinctipes*	CCAG43292.G1	
CoCepu1a	CoCepu1	468	*Celuca pugilator*	gb|DW176640.1|	
CoPaha1a	CoPaha1	842	*Parhyale hawaiensis*	2612_0_AFHB15258.b2_AFHB	
CoPaha2a	CoPaha2	956	*Parhyale hawaiensis*	1594_1_AFHB	
CoPaha 3a	CoPaha3	794	*Parhyale hawaiensis*	2612_0_AFHB24143.b2_AFHB	
CoPaha4a	CoPaha4	1093	*Parhyale hawaiensis*	2988-1	
CoLesa1a	CoLesa1	2501	*Lepeophtheirus salmonis*	ADND02013164.1	
CoLesa2a	CoLesa2	3801	*Lepeophtheirus salmonis*	ADND02074815.1	
CoLesa3a	CoLesa3	2142	*Lepeophtheirus salmonis*	ADND02104953.1	
CoLesa4a	CoLesa4	2301	*Lepeophtheirus salmonis*	ADND02043341.1	
Codapu1-C12	Codapu1	5452	*Daphnia pulex*	EU528615.1	
Codapu2-C1	Codapu2	5600	*Daphnia pulex*	EU528614.1	
Copia-10_DPu-I	Codapu3	5336	*Daphnia pulex*	repbase	

Table_S1

Report of the sequences obtained from PCR approaches. For each element, the host species, name, length and accession number are given, as well as the PCR methodology and primers used.

Higher Taxon	Species	Element	Accession Numbers	Method	GalEa Best Hit	E-Value	Non-GalEa Best Hit	E-Value
Chordata	*Branchiostoma floridae*	CoBranchio1	ABEP02005391	Figure S2				
	Branchiostoma lanceolatum	CoBranlan1	JT876534.1	BlastP	RT_GalEa1	4E-33	RT_Retrofit	3E-16
	Ciona intestinalis	CiCo1	DQ913003.2	Terrat and al., 2008				
	Molgula tectiformis	CoMol1	CJ341574.1	Figure S2				
		CoMol2	CJ419811.1	Figure S2				
		CoMol3	CJ420990	Figure S2				
Cnidaria	*Hydractinia symbiolongicarpus*	CoHydra1	AC234860	Figure S2				
Crustacea	*Lepeophtheirus salmonis*	CoLesa2	ADND02074815.1	Figure 4 and Figure S2				
		CoLesa3	ADND02104953.1	Figure 4				
	Rimicaris exoculata	CoRex1	HF548730 - HF548737	Figure 4 and Figure S2				
		CoRex2	HF548738 - HF548744	Figure 4 and Figure S2				
		CoRex3	HF548745 - HF548749	Figure 4 and Figure S2				
	Chorocaris chacei	CoChoro1	HF5487546 - HF5487547	Figure 4 and Figure S2				
	Eumunida sternomaculata	GalEa1	DQ912987.1	Terrat and al., 2008				
	Eumunida annulosa	GalEa1	EU097705.1	Terrat and al., 2008				
		GalEa2	DQ912993.1	Terrat and al., 2008				
		GalEa3	DQ912994.1	Terrat and al., 2008				
	Agonida laurentae	GalEa1	DQ912986.1	Terrat and al., 2008				
		GalEa2	DQ912989.1	Terrat and al., 2008				
		GalEa3	DQ912990.1	Terrat and al., 2008				
	Munidopsis recta	CoMure1	HF548752 - HF548753	Figure 4 and Figure S2				
	Galathea squamifera	GalEa1	DQ912988.1	Terrat and al., 2008				
	Munida gregaria	GalEa1	DQ912983.1	Terrat and al., 2008				
	Munida zebra	GalEa1	DQ912984.1	Terrat and al., 2008				
	Petrolisthes cinctipes	CoPeci1	Tagmount et al., 2010	BlastP	INT_GalEa1	6E-96	INT_SPM	2E-18
		CoPeci2	Tagmount et al., 2010	BlastP	INT_Olco1	1E-45	INT_Tor2	4E-12
		CoPeci3	CCAG43292.G1; FE831172.1	BlastP	RNaseH_Cico1	4E-21	RNaseH_Melmoth	1E-04
	Maja squinado	CoMaja1	HF548750 - HF548751	BlastP	INT_GalEa1	2E-93	INT_ASSBSV	3E-19
	Celuca pugilator	CoCepu1	DW176640.1	BlastP	RNaseH_GalEa1	6E-34	RNaseH_Araco	7E-04
	Eriocheir sinensis	CoEsi1	JR775724.1	BlastP	RT_GalEa1	3E-42	RT_Araco	4E-11
		CoEsi2	JR773210.1	Figure S2				
	Euphausia superba	CoEsup1	Toullec personal communication	BlastP	INT_GalEa1	2E-71	INT_ASSBSV	2E-16
		CoEsup2	Toullec personal communication	Figure 4 and Figure S2				
		CoEsup3	Toullec personal communication	Figure 4				
		CoEsup4	ES543757.1	BlastP	RNaseH_Zeco2	2E-15	RNaseH_CoDi6.2	1E-04
	Euphausia crystallorophias	CoEcrys1	Toullec personal communication	Figure 4 and Figure S2				
	Parhyale hawaiensis	CoPaha1	Jgi: 2612_0_AFHB15258.b2_AFHB	Figure 4				
		CoPaha2	Jgi:1594_1_AFHB	BlastP	RNaseH_GalEa1	3E-33	RNaseH_pCal	0.13
		CoPaha3	Jgi: 2612_0_AFHB24143.b2_AFHB	BlastP	INT_GalEa1	2E-19	INT_Tor2	6E-05
Ctenophora	*Mnemiopsis leidyi*	CoMnemio	AGCP01024244.1; AGCP01005248.1	Figure S2				
Echinoderma	*Paracentrotus lividus*	CoPali1	AM551615.1	Figure S2				
Actinopterygii	*Cyprinus carpio*	CoCycar1	HN151375	BlastP	INT_Zeco2	E-113	INT_MdEV	6E-13
	Danio rerio	Zeco2	DQ913002.1	Terrat et al., 2008				
		Zeco1	DQ913001.2	Terrat et al., 2008				
	Dicentrarchus labrax	CoDila1	CABK01032996.1	Figure S2				
	Gadus morhua	CoGamo1	CAEA01533636.1	Fisure S2				
	Haplochromis burtoni	CoHabu1	NZ01055413.1	Figure S2				
	Maylandia zebra	CoMay1	AGTA01055500.1	Figure S2				
	Mchenga conophoros	CoCono1	ABPJ01012473.1	Figure S2				
	Melanochromis auratus	CoMelau1	ABPL01008581.1	Figure S2				
	Neolamprologus brichardi	CoNeo1	AFNY01022989.1	Figure S2				
	Oncorhynchus mykiss	CoOnmy1	EZ792763.1	BlastP	INT_Zeco1	3E-46	INT_Cer1	4E-09
		CoOnmy2	EZ885234.1	BlastX	RT_Olco1	5E-35	RT_Araco	2E-04
		CoOnmy3	EZ876093.1	BlastX	RT_Zeco1	2E-37	RT_CoDi6.2	0.37
	Oncorhynchus nerka	CoOnner1	EZ837805.1	Figure S2				
	Oreochromis niloticus	CoOreo1	AERX01057963.2	Figure S2				
	Oryzias latipes	Olco1	DQ913000.2	Terrat et al., 2008				
	Pundamilia nyererei	CoPunmy1	AFNX01060828.1	Figure S2				
	Rhamphochromis esox	CoRham1	ABPN01008849.1	Figure S2				
	Salmo salar	Saco1	AGKD01016086.1	Figure S2				
	Sparus aurata	CoSpar1	FR554682	BlastX	INT_Olco1	4E-72	INT_REV	4E-12
	Xiphophorus maculatus	CoXiphoma1	AGAJ01021299.1	Figure S2				
Hemichordata	*Saccoglossus kowalevskii*	CoSaccoglo1	ACQM01121239.1	Figure S2				
Mollusca	*Crassostrea gigas*	CoCrassos1	CU999318.1	Figure S2				
	Crepidula fornicata	CoCre1	EZ566921.1	Figure S2				
	Doryteuthis pealeii	CoDope1	JK336837.1	Figure S2				
	Euprymna scolopes	CoEusco1	DW268340.1	BlastP	INT_Zeco1	1E-43	INT_Kabuki	1E-10
	Idiosepius paradoxus	CoIdio1	DB918837.1	BlastP	RT_GalEa1	4E-36	RT_Araco	2E-08
	Lottia gigantea	CoLogi1	FC693657.1	Figure S2				
	Sepia officinalis	CoSeof1	FO161620.1; FO178035.1; FO188867.1	Figure S2				
Rhodophyta	*Porphyridium cruentum*	CoPorcru1	HS681201.1; HS723886.1; HS648682.1; HS760695.1; HS723886.1	Figure S2				
		CoPorcru2	HS908486.1; HS868349.1; HS811949.1; HS809312.1	Figure S2				

Species	Element	Accession Numbers	Method	GalEa Best Hit	E-Value	Non-GalEa Best Hit	E-Value
Daphnia pulex	Copia-12_Dpu-I	Repbase: Copia-12_Dpu-I	Figure 4				
	Copia-1_Dpu-I	Repbase: Copia-1_Dpu-I	Figure 4				
	Copia-10_Dpu-I	Repbase: Copia-10_Dpu-I	Figure 4				
Lepeophtheirus salmonis	CoLesa1	ADND02013164.1	Figure 4				
	CoLesa4	ADND02043341.1	Figure 4				
Parhyale hawaiensis	CoPaha4	Jgi: contig 2988-1	BlastP	INT_Zeco1	0.077	INT_Hydra1-2	9,00E-50

Species	Element	Accession Numbers	
Daphnia pulex	Gypsy-15_DPu-I	Gypsy-15_DPu-I	http://www.girinst.org/repbase/ update/index.html
	Gypsy-22_DPu-I	Gypsy-22_DPu-I	http://www.girinst.org/repbase/ update/index.html
	Gypsy-25_DPu-I	Gypsy-25_DPu-I	http://www.girinst.org/repbase/ update/index.html
	Gypsy-26_DPu-I	Gypsy-26_DPu-I	http://www.girinst.org/repbase/ update/index.html
	Gypsy-27_DPu-I	Gypsy-27_DPu-I	http://www.girinst.org/repbase/ update/index.html
Lepeophtheirus salmonis	GyLesa1	ADND02072466.1	NCBI_wgs
	GyLesa2	ADND02000013.1	NCBI_wgs
	GyLesa3	ADND02000013.1	NCBI_wgs
	GyLesa5	ADND02049798.1	NCBI_wgs
Sacculina carcini	GySac1	HF548807	
	GySac2	HF548808	
Penaeus monodon	GyPemo1	MGID126613 est_p_monodon329, HF548819 - HF548819	http://www.marinegenomics.org/
	GyPemo2	MGID126443 est_p_monodon136	http://www.marinegenomics.org/
Litopenaeus vanameii	GyLiva1	MGID352304 CK591125.1	http://www.marinegenomics.org/
	GyLiva3	MGID410111 LV_NC_RA12J07r	http://www.marinegenomics.org/
	GyLiva4	MGID408910 LV_NC_RA09J01f	http://www.marinegenomics.org/
	GyLiva5	MGID410477 LV_NC_RA13E21f	http://www.marinegenomics.org/
	GyLiva6	MGID353368	http://www.marinegenomics.org/
Rimicaris exoculata	GyRex1	HF548771 - HF548778	
	GyRex2	HF548779 - HF54889	
	GyRex3	HF548790 - HF548795	
	GyRex4	HF548796 - HF548797	
	GyRex5	HF548798	
	GyRex6	HF548799	
Nematocarcinus burukovskii	GyNem1	HF548804 - HF548805	
Chorocaris chacei	GyChoro1	HF548810 - HF548811	
	GyChoro2	HF548812 - HF548813	
Homarus americanus	GyHoam1	MGID100769 est_h_americanus2358	http://www.marinegenomics.org/
	GyHoam2	MGID98322 est_h_americanus5401	http://www.marinegenomics.org/
	GyHoam3	MGID98312 est_h_americanus5412	http://www.marinegenomics.org/
Orconectes limosus	GyOrli1	HF548806	
Agonida laurentae	GyAla1	HF548816	
	GyAla2	HF548817 - HF548818	
Munidopsis recta	GyMure1	HF548809	
Bythograea thermydron	GyBy1	HF548800 - HF548801	
Necora puber	GyNeco1	HF548802 - HF548803	
Maja squinado	GyMaja1	HF548821	
	GyMaja2	HF548822	
Euphausia crystallorophias	GyEcrys1	Toullec personal communication	
Parhyale hawaiensis	GyPaha1	741_1-4_AFHB	http://genome.jgi-psf.org/parha/parha.download.ftp.html
	GyPaha2	741_0_AFHB26204.b2	http://genome.jgi-psf.org/parha/parha.download.ftp.html
	GyPaha3	5864_0_AFHB13272.b2_AFHB	http://genome.jgi-psf.org/parha/parha.download.ftp.html
	GyPaha4	10574_0_AFHB4423.x1_AFHB	http://genome.jgi-psf.org/parha/parha.download.ftp.html
	GyPaha5	0_0_AFHB14058.b2_AFHB	http://genome.jgi-psf.org/parha/parha.download.ftp.html
	GyPaha6	10480_0_AFHB40998.g3_AFHB	http://genome.jgi-psf.org/parha/parha.download.ftp.html
	GyPaha7	3094_3_AFHB	http://genome.jgi-psf.org/parha/parha.download.ftp.html

Table_S2

List of GalEa-like retrotransposons identified. For each element, the corresponding host species and the accession number(s) are indicated. The GalEa nature of the elements was determined following different classification methods: Figure B and SupData E correspond to the phylogenetic analyses; BlastP to the BLAST-based classification method, for which the best GalEa and non-GalEa hits are given with the corresponding E-values.

Species	Elements	Nucleotidic sequence	Proteic sequence
	primer CD1	ARR GCN MGN YTN GTN GC	KARLVA
Maja squinado	CoMaja1	AGG GCG AGG CUG GUG GC	RARLVA
Euphausiacea superba	CoEsup 3	AAA GCC CGG CUU GUU GC	KARLVA
Parhyale hawaiensis	CoPaha1	AAA GCG CGA UUG GUU GC	KARLVA
Lepeophtheirus salmonis	CoLesa2	AAA UCA AGA CUG GUA GC	KSRLVA
Lepeophtheirus salmonis	CoLesa3	AAG ACA AGA UUA GUA GU	KTRLVV
Lepeophtheirus salmonis	CoLesa4	AAA GAA AGA UUA GUA UU	KERLVL
Dapnhia pulex	Copia-1_Dpu-I	AAA GCA AGA CUA GUA GC	KARLVA
Rimicaris exoculata	CoRex1	AAA GUG CAU UUA GUG GC	KVHLVA
Rimicaris exoculata	CoRex2	AAG GCC AGG CUU GUA GC	KARLVA
Drosophila simulans	Copia	AAA GCU AGA UUG GUU GC	KARLVA
Drosophila melanogaster	1731	AAG GCA CGA CUA GUA GC	KARLVA
Nicotiana tabacum	Tnt1	AAA GCU CGA UUG GUG GU	KARLVV
Ciona intestinalis	Cico1	AAA GCC AGG CUA GUG GC	KARLVA
Saccharomyces cerevisae	Ty4	AAG GCU AGG AUA GUC UG	KARIVC
Kluyveromyces marxianus	Tkm1	AAG UGC AGA CUA GUU GC	KCRLVA
	primer CD2	TAY GTN GAY GAY NTN NT	YVDDLF
Munidopsis recta	CoMure1	UAC GUU GAC GAU CUU UU	YVDDLF
Chorocharis chacei	CoChoro1	UAU GUU GAC GAU UUC UU	YVDDFL
Euphausiacea cristallo	CoCrys1	UAU GUG GAU GAC UUC UU	YVDDFL
Euphausiacea superba	CoEsup 3	UAC GUA GAC GAC UUC CU	YVDDFL
Parhyale hawaiensis	CoPaha1	CAU GUG GAU GAU UUC UA	HVDDFY
Lepeophtheirus salmonis	CoLesa1	UAU GUA GAU GAU AUU UU	YVDDIL
Lepeophtheirus salmonis	CoLesa2	CAU GUU GAU AAU UUG UU	HVDNLL
Lepeophtheirus salmonis	CoLesa4	UAU GUU GUC AAU AUU UU	YVVNIL
Dapnhia pulex	Copia-1_Dpu-I	UGG GUG GAU GAC GGU CU	WVDDGL
Rimicaris exoculata	CoRex2	UAC AUG GAU GAC UUC UU	YMDDFL
Rimicaris exoculata	CoRex3	CAU GUG GAU GAU UAU UU	HVDDYF
Drosophila simulans	Copia	UAU GUA GAU GAU GUG GU	YVDDVV
Drosophila melanogaster	1731	UAU GUU GAU GAU UUA AU	YVDDLI
Nicotiana tabacum	Tnt1	UAU GUG GAU GAC AUG CU	YVDDML
Ciona intestinalis	Cico1	UAU GUG GAU GAC AUU UU	YVDDIL
Saccharomyces cerevisae	Ty4	UAU GUU GAU GAC UGC GU	YVDDCV
Kluyveromyces marxianus	Tkm1	UUU GUC GAU GAC AUG CU	FVDDML

													ANN	ANR	TCR	TCN	ACR	TA	
	CD1	ARR	GCN	MGN	YTN	GTN	GC	KARLVA			CD2	TAY	GTN	GAY	GAY	NTN	NT	YVDDL
Maja squinado	Comaja1	AGG	GCG	AGG	CUG	GUG	GC	RARLVA		*Munidopsis recta*	Comur1	UAC	GUU	GAC	GAU	CUU	UU	YVDDLF
Euphausiacea superba	Coesup 3	AAA	GCC	CGG	CUU	GUU	GC	KARLVA		*Chorocharis*	Cochoro1	UAU	GUU	GAC	GAU	UUC	UU	YVDDFL
Parhyale hawaiensis	Copaha1	AAA	GCG	CGA	UUG	GUU	GC	KARLVA		*euphausiacea cristallo*	Coecris1	UAU	GUG	GAU	GAC	UUC	UU	YVDDFL
Lepeophtheirus salmonis	Colesa1	AAG	GCA	CGG	CUA	GUA	GC	KARLVA		*Euphausiacea superba*	Coesup 3	UAC	GUA	GAC	GAC	UUC	CU	YVDDFL
Lepeophtheirus salmonis	Colesa2	AAA	UCA	AGA	CUG	GUA	GC	KSRLVA		*Parhyale hawaiensis*	Copaha1	CAU	GUG	GAU	GAU	UUC	UA	HVDDFY
Lepeophtheirus salmonis	Colesa3	AAG	ACA	AGA	UUA	GUA	GU	KTRLVV		*Lepeophtheirus salmonis*	Colesa1	UAU	GUA	GAU	GAU	AUU	UU	YVDDIL
Lepeophtheirus salmonis	Colesa4	AAA	GAA	AGA	UUA	GUA	UU	KERLVL		*Lepeophtheirus salmonis*	Colesa2	CAU	GUU	GAU	AAU	UUG	UU	HVDNLL
Dapnhia pulex	Codapu1	AAA	GCA	AGA	CUA	GUA	GC	KARLVA		*Lepeophtheirus salmonis*	Colesa4	UAU	GUU	GUC	AAU	AUU	UU	YVVNIL
Rimicaris exoculata	corex1	AAA	GUG	CAU	UUA	GUG	GC	KVHLVA		*Dapnhia pulex*	Codapu1	UGG	GUG	GAU	GAC	GGU	CU	WVDDGL
Rimicaris exoculata	corex2	AAG	GCC	AGG	CUU	GUA	GC	KARLVA		*Rimicaris exoculata*	corex1	UAU	GUU	GAC	GAU	UUC	UU	YVDDFL
Rimicaris exoculata	corex3	AAA	GCU	AGA	CUA	GUC	GC	KARLVA		*Rimicaris exoculata*	corex2	UAC	AUG	GAU	GAC	UUC	UU	YMDDFL
										Rimicaris exoculata	corex3	CAU	GUG	GAU	GAU	UAU	UU	HVDDYF
Drosophila simulans	Copia	AAA	GCU	AGA	UUG	GUU	GC	KARLVA		Drosophila simulans	Copia	UAU	GUA	GAU	GAU	GUG	GU	YVDDVV
Drosophila melanogaster	1731	AAG	GCA	CGA	CUA	GUA	GC	KARLVA		Drosophila melanogaster	1731	UAU	GUU	GAU	GAU	UUA	AU	YVDDLI
Nicotiana tabacum	Tnt1	AAA	GCU	CGA	UUG	GUG	GU	KARLVV		Nicotiana tabacum	Tnt1	UAU	GUG	GAU	GAC	AUG	CU	YVDDML
Ciona intestinalis	Cico1	AAA	GCC	AGG	CUA	GUG	GC	KARLVA		Ciona intestinalis	Cico1	UAU	GUG	GAU	GAC	AUU	UU	YVDDIL
Saccharomyces cerevisae	Ty4	AAG	GCU	AGG	AUA	GUC	UG	KARIVC		Saccharomyces cerevisae	Ty4	UAU	GUU	GAU	GAC	UGC	GU	YVDDCV
Kluyveromyces marxianus	Tkm1	AAG	UGC	AGA	CUA	GUU	GC	KCRLVA		Kluyveromyces marxianus	Tkm1	UUU	GUC	GAU	GAC	AUG	CU	FVDDML

Table_S3.
Comparison of CD1 and CD2 primers with Copia sequences.
Dissimilarities at nucleic or amino-acid levels are indicated in red.

Annexes 2: Données supplémentaires de l'article : GalEa retrotransposons in Fungi

Element	Host				ripped copy
CoAuhe1	*Aulographum hederae v2.0*	Dothideomycetes	Capnodiales	scaffold_15_152119_157936	
CoLeflu1	*Lentithecium fluviatile v1.0*	Dothideomycetes	Pleosporales	scaffold_19	x
CoLeflu2	*Lentithecium fluviatile v1.1*	Dothideomycetes	Pleosporales	scaffold_16	x
CoLeflu3	*Lentithecium fluviatile v1.2*	Dothideomycetes	Pleosporales	scaffold_1	x
CoLeflu4	*Lentithecium fluviatile v1.3*	Dothideomycetes	Pleosporales	scaffold_7	x
CoPyte1	*Pyrenophora teres*	Dothideomycetes	Pleosporales	XM_003306218.1	
CoPytri1	*Pyrenophora tritici-repentis*	Dothideomycetes	Pleosporales	XM_001940316.1|	
CoPytri2	*Pyrenophora tritici-repentis*	Dothideomycetes	Pleosporales	Supercontig_1_9_1403633_1409.1	
CoPytri3	*Pyrenophora tritici-repentis*	Dothideomycetes	Pleosporales	Supercontig_1_9_532949_53904.1	
Cegeo1	*Cenococcum geophilum 1.58 v1.0*	Dothideomycetes	incertae sedis	scaffold_137	
Cegeo2	*Cenococcum geophilum 1.58 v1.1*	Dothideomycetes	incertae sedis	scaffold_1	x
Cegeo3	*Cenococcum geophilum 1.58 v1.2*	Dothideomycetes	incertae sedis	scaffold_2	x
Cegeo4	*Cenococcum geophilum 1.58 v1.3*	Dothideomycetes	incertae sedis	scaffold_77	x
CoNefi1	*Aspergillus (=Neosartorya) fischeri*	Eurotiomycetes	Eurotiales	AAKE03000013.1	
CoAsbra1	*Aspergillus brasiliensis*	Eurotiomycetes	Eurotiales	scaffold_15 : 696707 - 697435	
CoAsni1	*Aspergillus niger*	Eurotiomycetes	Eurotiales	Supercontig 1	x
CoTama1	*Talaromyces marneffei*	Eurotiomycetes	Eurotiales	XM_002148536.1	
CoTasti1	*Talaromyces stipitatus*	Eurotiomycetes	Eurotiales	XM_002339988.1	
CoTasti2	*Talaromyces stipitatus*	Eurotiomycetes	Eurotiales	sc_1105507295549_193648_200133	
CoTasti3	*Talaromyces stipitatus*	Eurotiomycetes	Eurotiales	sc_1105507295523_313780_319716	
CoTasti4a	*Talaromyces stipitatus*	Eurotiomycetes	Eurotiales	sc_1105507295517_362599_368879	
CoTasti4b	*Talaromyces stipitatus*	Eurotiomycetes	Eurotiales	sc_1105507295487_93890_100172	
CoTasti4c	*Talaromyces stipitatus*	Eurotiomycetes	Eurotiales	sc_1105507295487_162315_168593	
CoBlugra1	*Blumeria graminis f. sp. hordei*	Leotiomycetes	Erysiphales	ABSB02007977.1	
CoBlugra2	*Blumeria graminis f. sp. hordei*	Leotiomycetes	Erysiphales	HF_943531_792850_797984	
CoBlugra3	*Blumeria graminis f. sp. hordei*	Leotiomycetes	Erysiphales	HF_943549_1008260_1014140	
CoBlugra4	*Blumeria graminis f. sp. hordei*	Leotiomycetes	Erysiphales	HF_943532_1404457_1410335	
CoBlugra5a	*Blumeria graminis f. sp. hordei*	Leotiomycetes	Erysiphales	HF_943526_1689139_1694600	
CoBlugra5b	*Blumeria graminis f. sp. hordei*	Leotiomycetes	Erysiphales	HF_943526_601961_607599	
CoBlugra5c	*Blumeria graminis f. sp. hordei*	Leotiomycetes	Erysiphales	HF_943505_788007_793645	
CoBlugra5d	*Blumeria graminis f. sp. hordei*	Leotiomycetes	Erysiphales	HF_943534_628614_634253	
CoBlugra6	*Blumeria graminis f. sp. hordei*	Leotiomycetes	Erysiphales	HF_943541_769408_775061	
CoBlugra7	*Blumeria graminis f. sp. hordei*	Leotiomycetes	Erysiphales	HF_943563_536750_542621	
CoBlugra8	*Blumeria graminis f. sp. hordei*	Leotiomycetes	Erysiphales	HF_943548_3763495_3768937	
CoBlugra9	*Blumeria graminis f. sp. hordei*	Leotiomycetes	Erysiphales	HF_943505_1470619_1478969	
CoBlugra10	*Blumeria graminis f. sp. hordei*	Leotiomycetes	Erysiphales	HF_943514_476282_481794	
CoErypi1	*Erysiphe pisi*	Leotiomycetes	Erysiphales	CACN01010277.1	x
CoErypi2	*Erysiphe pisi*	Leotiomycetes	Erysiphales	gi323830023_c336_3150_8871	
CoAssa1	*Ascocoryne sarcoides NRRL50072*	Leotiomycetes	Helotiales	AIAA01000105.1	
CoBofu2a	*Botryotinia fuckeliana (= Botrytis cinerea)*	Leotiomycetes	Helotiales	Supercontig_1.92	
CoBofu2b	*Botryotinia fuckeliana (= Botrytis cinerea)*	Leotiomycetes	Helotiales	Supercontig_1.78	
CoBofu3	*Botrvotinia fuckeliana (= Botrvtis cinerea)*	Leotiomvcetes	Helotiales		
CoScle1	*Sclerotinia sclerotiorum*	tiomycetes	Helotiales	XM_001586087.1	
CoSclehol	*Sclerotinia homoeocarpa*	Leotiomycetes	Helotiales	AKKO01018210.1	
CoMabru1	*Marssonina brunnea*	Leotiomycetes	Helotiales	AFXC01000436.1	
CoMeva1	*Meliniomyces variabilis F v1.0*	Leotiomycetes	Helotiales	scaffold_9 : 2138960 - 2140486	
CoMeva2	*Meliniomyces variabilis F v1.1*	Leotiomycetes	Helotiales	scaffold_26 : 170937 - 172448	x
CoOima1a	*Oidiodendron maius*	Leotiomycetes	incertae sedis	scaffold_6	
CoOima1b	*Oidiodendron maius*	Leotiomycetes	incertae sedis	scaffold_13	
CoOima1c	*Oidiodendron maius*	Leotiomycetes	incertae sedis	scaffold_24	
CoOima2	*Oidiodendron maius*	Leotiomycetes	incertae sedis		x
CoColhi1	*Colletotrichum higginsianum*	Sordariomycetes	Glomerellales	Supercontig 6656: 21-1589	
CoColhi2	*Colletotrichum higginsianum*	Sordariomycetes	Glomerellales	Supercontig 6901: 627-2249	x
CoCogra1	*Colletotrichum(=Glomerella) graminicola M1.001*	Sordariomycetes	Glomerellales	Supercontig 164: 17635-18660	x
CoCogra2	*Colletotrichum(=Glomerella) graminicola M1.002*	Sordariomycetes	Glomerellales	Supercontig 114: 38957-40603	x
CoGloac1	*Glomerella acutata (Colletotrichum fiorinae MH 18) v1.0*	Sordariomycetes	Glomerellales	scaffold_75 : 22666 - 24234	x
CoVaa1	*Verticillium albo-atrum VaMs.102*	Sordariomycetes	Glomerellales	Supercontig 7: 1266361-1267446	x
CoVerda1	*Verticillium dahliae VdLs.17*	Sordariomycetes	Glomerellales	Supercontig_1.1 : 439389 - 440798	x
CoVerda2	*Verticillium dahliae VdLs.18*	Sordariomycetes	Glomerellales	Supercontig_1.8 : 881481 - 882878	
CoActa1	*Aciculosporiumt take*	Sordariomycetes	Hypocreales	AFQZ01002400.1	
CoClafu1	*Claviceps fusiformis*	Sordariomycetes	Hypocreales	AFRA01000981.1	x
CoClapa1	*Claviceps paspali*	Sordariomycetes	Hypocreales	AFRC01000028.1	x
CoClapu1	*Claviceps purpurea*	Sordariomycetes	Hypocreales	CAGA01000008.1	
CoCormi1	*Cordyceps militaris*	Sordariomycetes	Hypocreales	AEVU01000216.1	x
CoEpia1	*Epichloe amarillans*	Sordariomycetes	Hypocreales	AFRF01000421.1	x
CoEpia2	*Epichloe amarillans*	Sordariomycetes	Hypocreales	AFRF01000421.1	x
CoEpib1	*Epichloe brachyelytri*	Sordariomycetes	Hypocreales	AFRB01000288.1	x
CoEpib2	*Epichloe brachyelytri*	Sordariomycetes	Hypocreales	AFRB01000450.1	x
CoEpifes1	*Epichloe festucae*	Sordariomycetes	Hypocreales	ADFL02000356.1	
CoEpifes2	*Epichloe festucae*	Sordariomycetes	Hypocreales	AFRX01000183.1	x
CoEpigly1	*Epichloe glyceriae*	Sordariomycetes	Hypocreales	AFRG01000153.1	
CoEpity1	*Epichloe typhina*	Sordariomycetes	Hypocreales	AMDI01000045.1	x
CoEpity2	*Epichloe typhina*	Sordariomycetes	Hypocreales	AMDI01001101.1	x
CoMetani1	*Metarhizium anisopliae*	Sordariomycetes	Hypocreales	ADNJ01000231.1	
CoMetani2	*Metarhizium anisopliae*	Sordariomycetes	Hypocreales		
CoNeoga1	*Metarhizium robertsii ARSEF 23*	Sordariomycetes	Hypocreales	AFRE01000996.1	x
CoNeoga2	*Neotyphodium gansuense*	Sordariomycetes	Hypocreales	AFRE01000437.1	x
CoMagor1	*Magnaporthe oryzae=grisae*	Sordariomycetes	Magnaporthales	AHZT01001345.1	
CoMagpo1	*Magnaporthe poae*	Sordariomycetes	Magnaporthales	Supercontig 1: 5937689-5939056	
CoMagpo2	*Magnaporthe poae*	Sordariomycetes	Magnaporthales	Supercontig 11: 635231-636607	x
CoMagpo3	*Magnaporthe poae*	Sordariomycetes	Magnaporthales	Supercontig 119: 3148-4722	x
CoMagpo4	*Magnaporthe poae*	Sordariomycetes	Magnaporthales	ADBL01000449.1	
CoOpno1	*Ophiostoma novo-ulmi subsp*	Sordariomycetes	Ophiostomatales	AMZD01000009.1	x

CoChaglo1a	*Chaetomium globosum*	Sordariomycetes	Sordariales	XM_001230268.1	
CoChaglo1b	*Chaetomium globosum*	Sordariomycetes	Sordariales	Supercontig_1.1	
CoChaglo1c	*Chaetomium globosum*	Sordariomycetes	Sordariales	Supercontig_1.2	
CoChaglo1d	*Chaetomium globosum*	Sordariomycetes	Sordariales	Supercontig_1.7	
CoChaglo1e	*Chaetomium globosum*	Sordariomycetes	Sordariales	Supercontig_1.1	
CoDaes1	*Daldinia eschscholzii*	Sordariomycetes	Xylariales	scaffold_49 : 1352 - 3016	x
CoHypo(CO27)1	*Hypoxylon sp. CO27-5*	Sordariomycetes	Xylariales	scaffold_373 : 5832 - 7400	x
CoHypo(CO27)2	*Hypoxylon sp. CO27-6*	Sordariomycetes	Xylariales	scaffold_49 : 1352 - 3016	x
CoHypo(EC38)1	*Hypoxylon sp. EC38*	Sordariomycetes	Xylariales	scaffold_278	x
CoHypo(EC38)2	*Hypoxylon sp. EC39*	Sordariomycetes	Xylariales	scaffold_143	x

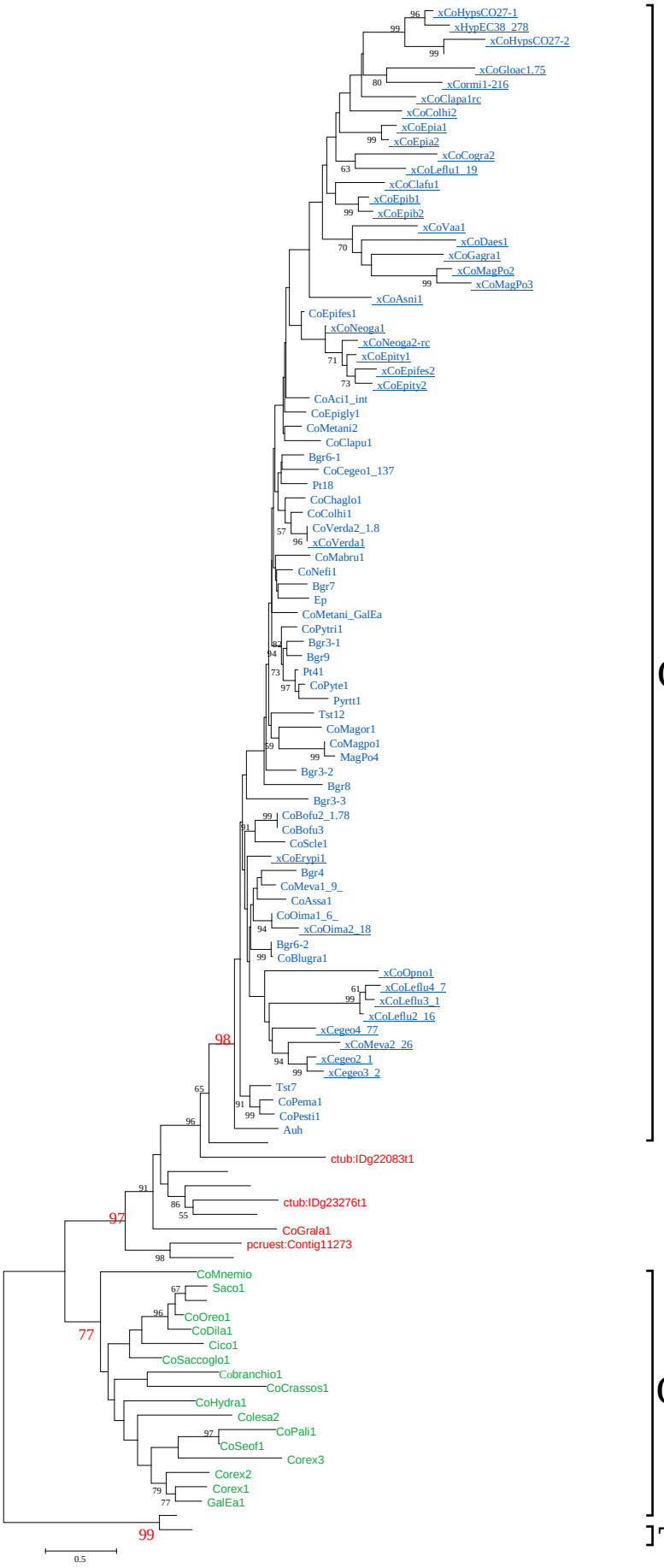
Supdata 1. List of GalEa retrotransposons from Fungi. For each element, the corresponding host species and the accession number are indicated, in bold for sequences newly deposited. The seven annotated elements are highlighted in red. The different copies from a same element have the same name but are distinguished by a letter at their end. An element is considered as ripped when his AT-content is higher than 70%.

	Subphylum name	Class name	Order name	No GalEa detected in sequenced genome Galea detected in sequenced genome povided by JGI (or other intitutes) GalEa detected on NCBI database	genome size (Mb)	CoGaFu sequences detected
Ascomycota	Pezizomycotina	Dothideomycetes	Botryosphaeriales	*Aplosporella prunicola CBS 121.167 v1.0*		
				Botryosphaeria dothidea		
				Macrophomina phaseolina MS6		
				Neofusicoccum parvum UCRNP2		
				Saccharata proteae CBS 121410 v1.0		
				Rhytidhysteron rufulum		
Ascomycota	Pezizomycotina	Dothideomycetes	Capnodiales	*Aulographum hederae v2.0*	31.98	
				Baudoinia compniacensis UAMH 10762 (4089826) v1.0		
				Cercospora zeae-maydis v1.0		
				Cladosporium fulvum v1.0		
				Dissoconium aciculare v1.0		
				Dothistroma septosporum		
				Mycosphaerella fijiensis		
				Mycosphaerella graminicola		
				Pleurata hortae v1.1	16.94	only two sequences of 878 and 803 bp
				Polychaeton citri v1.0		
				Septoria musiva		
				Septoria populicola v1.0		
				Zasmidium cellar		
Ascomycota	Pezizomycotina	Dothideomycetes	Dothideales	*Aureobasidium pullulans*		
Ascomycota	Pezizomycotina	Dothideomycetes	Hysteriales	*Hysterium pulicare*		
Ascomycota	Pezizomycotina	Dothideomycetes	Mytilinidiales	*Lophium mytilinum CBS 269.34 Standard Draft*		
Ascomycota	Pezizomycotina	Dothideomycetes	Patellariales	*Patellaria atrata v1.0*		
Ascomycota	Pezizomycotina	Dothideomycetes	Pleosporales	*Aaosphaeria arxii CBS 175.79 v1.0*		
				Alternaria brassicicola		
				Amniculicola lignicola CBS 123094 v1.0	49.57	
				Byssothecium circinans CBS 675.92 Standard Draft	49.29	
				Cochliobolus carbonum		
				Cochliobolus heterostrophus		
				Cochliobolus lunatus		
				Cochliobolus miyabeanus		
				Cochliobolus sativus ND90Pr v1.0		
				Cochliobolus victoriae		
				Corynespora cassiicola Philippines Standard Draft		
				Cucurbitaria berberidis CBS 394.84 v1.0		
				Delitschia confertaspora ATCC 74209 Standard Draft		
				Didymella exigua		
				Didymella exigua CBS 183.55 v1.0		
				Dothidotthia symphoricarpi v1.0		
				Karstenula rhodostoma CBS 690.94 v1.0		
				Lentithecium fluviatile v1.0	54.69	
				Leptosphaeria maculans		
				Lophiostoma macrostomum v1.0		
				Macroventuria anomochaeta CBS 525.71 v1.0		
				Myriangium duriaei	25.69	only one sequence of 487 bp
				Ophiobolus disseminans CBS 113818 Standard Draft	41.68	only two sequences of 427 and 606 bp
				Phoma tracheiphila IPT5 v1.0	34.24	only two sequences of 781 and 1098 bp
				Pleomassaria siparia v1.0	43.17	
				Pyrenochaeta lycopersici		
				Pyrenophora teres	33.58	
				Pyrenophora tritici-repentis	37.84	
				Setosphaeria turcica		
				Sporormia fimetaria v1.0	25.90	
				Stagonospora (Phaeosphaeria) nodorum		
				Westerdykella ornata CBS 379.55 Standard Draft		
				Zopfia rhizophila v1.0	152.78	
Ascomycota	Pezizomycotina	Dothideomycetes	Trypetheliales	*Trypethelium eluteriae v1.0*		
Ascomycota	Pezizomycotina	Dothideomycetes	incertae sedis	*Acidomyces richmondensis BFW*		
				Cenococcum geophilum 1.58 v1.0	177.57	
				Lepidopterella palustris v1.0		
				Trematosphaeria pertusa CBS 122368 v1.0		
				Venturia pyrina		
Ascomycota	Pezizomycotina	Eurotiomycetes	Eurotiales	*Aspergillus (=Neosartorya) fischeri*	32.55	
				Aspergillus brasiliensis	35.81	only two sequences of 728 and 1669 bp
				Aspergillus carbonarius		
				Aspergillus clavatus		
				Aspergillus flavus		
				Aspergillus fumigatus		
				Aspergillus glaucus		
				Aspergillus nidulans		
				Aspergillus niger	34.85	only one sequence of 1398 bp
				Aspergillus oryzae		
				Aspergillus terreus		
				Aspergillus acidus v1.0		
				Aspergillus aculeatus ATCC 16872		
				Aspergillus kawachii IFO 4308		
				Aspergillus zydowii v1.0		
				Aspergillus tubingensis v1.0		
				Aspergillus versicolor v1.0		
				Aspergillus wentii v1.0		
				Aspergillus zonatus v1.0		
				Eurotium rubrum v1.0		
				Penicillium bilaiae ATCC 20851 v1.0	37.54	only one sequence of 533 bp
				Penicillium brevicompactum AgRF18 v1.0		
				Penicillium brevicompactum v2.0		
				Penicillium canescens ATCC 10419 v1.0		
				Penicillium chrysogenum		
				Penicillium digitatum PHI26		
				Penicillium expansum ATCC 24692 v1.0		
				Penicillium fellutanum ATCC 48694 v1.0		
				Penicillium glabrum DAOM 239074 v1.0		
				Penicillium janthinellum ATCC 10455 v1.0		
				Penicillium lanosocoeruleum ATCC 48919 v1.0		

				Penicillium oxalicum 114-2		
				Penicillium raistrickii ATCC 10490 v1.0		
				Penicillium roqueforti		
				Talaromyces aculeatus ATCC 10409 v1.0	37.27	
				Talaromyces marneffei	28.64	
				Talaromyces stipitatus	35.69	
				Thermoascus aurantiacus		
Ascomycota	Pezizomycotina	Eurotiomycetes	Onygenales	*Arthroderma benhamiae*		
				Ascosphaera apis (Beebase)		
				Blastomyces (=Ajellomyces) dermatitidis SLH14081 (Broad)		
				Coccidioides immitis RS		
				Coccidioides posadasii Silveira (Broad)		
				Gymnascella aurantiaca v1.0		
				Gymnascella citrina v1.1		
				Histoplasma capsulatum NAm1		
				Microsporum canis		
				Microsporum gypseum (Broad)		
				Paracoccitioides brasiliensis Pb01		
				Trichophyton equinum (Broad)		
				Trichophyton rubrum		
				Trichophyton tonsurans (Broad)		
				Trichophyton verrucosum		
				Uncinocarpus reesii		
Ascomycota	Pezizomycotina	Eurotiomycetes	Verrucariales	*Endocarpon pusillum*		
	Pezizomycotina	Eurotiomycetes	incertae sedis	*Monascus purpureus v1.0*		
				monascus ruber		
Ascomycota	Pezizomycotina	Leotiomycetes	Erysiphales	*Blumeria graminis f. sp. hordei DH14*	118.73	
				Erysiphe pisi (Max Planck Institues)	49.38	
				Golovinomyces orontii (Max Planck Institues)		
Ascomycota	Pezizomycotina	Leotiomycetes	Helotiales	*Ascocoryne sarcoides NRRL50072*		
				Botryotinia fuckeliana (= Botrytis cinerea)	42.66	
				Botrytis tulipae Bt9901		
				Cadophora sp. DSE1049 v1.0		
				Chalara longipes BDJ v1.0		
				Glarea lozoyensis ATCC 20868		
				Marssonina brunnea		
				Melinitomyces bicolor E v2.0	82.38	
				Melinitomyces variabilis F v1.0	55.86	
				Sclerotinia borealis		
				Sclerotinia sclerotiorum	38.33	
				Sclerotinia homoeocarpa		
Ascomycota	Pezizomycotina	Leotiomycetes	incertae sedis	*Oidiodendron maius*	46.43	
				Amorphotheca resinae v1.0		
Ascomycota	Pezizomycotina	Lecanoromycetes		*Cladonia grayi Cgr/DA2myc/ss v1.0*		
				Xanthoria parietina 46-1 v1.0		
Ascomycota	Pezizomycotina	Orbiliomycetes		*Arthrobotrys oligospora ATCC 24927*		
				Monacrosporium haptotylum CBS 200.50		
Ascomycota	Pezizomycotina	Pezizomycetes		*Ascobolus immersus RN42 v1.0*		
				Choiromyces venosus 120613-1 v1.0		
				Morchella conica CCBAS932 v1.0		
				Pyronema confluens CBS100304		
				Terfezia boudieri S1 v1.0		
				Tuber melanosporum from Genoscope		
				Wilcoxina mikolae CBS 423.85 v1.0		
Ascomycota	Pezizomycotina	Sordariomycetes	Calosphaeriales	*Phaeoacremonium aleophilum UCRPA7*		
Ascomycota	Pezizomycotina	Sordariomycetes	Coniochaetales	*Coniochaeta ligniaria CBS 111746*		
Ascomycota	Pezizomycotina	Sordariomycetes	Diaporthales	*Cryphonectria parasitica*		
				Ophiognomonia clavigignenti-juglandacearum		
Ascomycota	Pezizomycotina	Sordariomycetes	Glomerellales	*Acremonium alcalophilum v*		
				Colletotrichum cereale		
				Colletotrichum higginsianum	49.08	only to sequence of 1623 & 2723 bp
				Colletotrichum graminicola M1.001	51.60	
				Colletotrichum acutata (Colletotrichum florinae MH 18) v1.0	50.04	
				Glomerella cingulata		
				Sodiomyces alkalinus v1.0		
				Verticillium albo-atrum VaMs.102	32.83	
				Verticillium dahliae VdLs.17	33.83	
Ascomycota	Pezizomycotina	Sordariomycetes	Hypocreales	*Aciculosporium take*		
				Beauveria bassiana ARSEF 2860	33.69	
				Claviceps fusiformis		
				Claviceps paspali		
				Claviceps purpurea		
				Cordyceps militaris	32.27	
				Epichloe amarillans		
				Epichloe brachyelytri		
				Epichloe festucae		
				Epichloe glyceriae		
				Epichloe typhina		
				Fusarium fujikuroi IMI 58289		
				Fusarium graminearum		
				Fusarium oxysporum		
				Fusarium verticillioides =(Gibberella moniliformis)		
				Ilyonectria sp.		
				Metacordyceps chlamydosporia		
				Metarhizium acridum CQMa 102		
				Metarhizium anisopliae		
				Metarhizium robertsii ARSEF 23	39.14	
				Nectria haematococca (=fusarium solani)		
				Neotyphodium gansuense		
				Trichoderma asperellum CBS 433.97 v1.0		
				Trichoderma atroviride		
				Trichoderma citrinoviride		
				Trichoderma harzianum CBS 226.95 v1.0		
				Trichoderma longibrachiatum ATCC 18648 v3.0		
				Trichoderma reesei		
				Trichoderma virens		
				Villosiclava virens		

Ascomycota	Pezizomycotina	Sordariomycetes	Magnaporthales	*Gaeumannomyces graminis*	43.62	
				Magnaporthe oryzae=grisae	41.03	only one sequence of 500 bp
				Magnaporthe poae	39.50	
Ascomycota	Pezizomycotina	Sordariomycetes	Ophiostomatales	*Grosmannia clavigera kw1407*		
				Ophiostoma novo-ulmi subsp		
				Ophiostoma piceae UAMH 11346	32.84	
Ascomycota	Pezizomycotina	Sordariomycetes	Sordariales	*Chaetomium globosum*	34.9	
				Myceliophthora (Sporotrichum) thermophila		
				Neurospora crassa OR74A		
				Neurospora discreta		
				Neurospora tetrasperma		
				Podospora anserina		
				Sordaria macrospora		
				Sporotrichum thermophile		
				Thielavia antarctica CBS 123565 v1.0		
				Thielavia appendiculata CBS 731.68 v1.0		
				Thielavia arenaria CBS 508.74 v1.0		
				Thielavia hyrcaniae CBS 757.83 v1.0		
				Thielavia terrestris		
Ascomycota	Pezizomycotina	Sordariomycetes	Xylariales	*Anthostoma avocetta NRRL 3190 v1.0*	56.23	
				Daldinia eschscholzii	37.55	
				Eutypa lata UCREL1		
				Hypoxylon sp. CI-4A v1.0	37.70	only one sequence of 309 bp
				Hypoxylon sp. CO27-5	46.59	
				Hypoxylon sp. EC38	47.30	
Ascomycota	Pezizomycotina	Sordariomycetes	incertae sedis	*Apiospora montagnei*		
Ascomycota	Pezizomycotina	Xylonomycetes		*Xylona heveae TC161 v1.0*		
Ascomycota	Saccharomycotina			*Ascoidea rubescens DSM 1968*		
				Ashbya gossypii ATCC 10895		
				Babjeviella inositovora NRRL Y-12698 v1.0		
				Candida arabinofermentans NRRL YB-2248 v1.0		
				Candida caseinolytica Y-17796 v1.0		
				Candida tanzawaensis NRRL Y-17324 v1.0		
				Candida tenuis NRRL Y-1498 v1.0		
				Cyberlindnera jadinii NRRL Y-1542 v1.0		
				Debaryomyces hansenii		
				Dekkera bruxellensis CBS 2499 v2.0		
				Hanseniaspora valbyensis NRRL Y-1626 v1.1		
				Hansenula polymorpha NCYC 495 leu1.1 (ATCC MYA-335)		
				Hyphopichia burtonii NRRL Y-1933 v1.0		
				Kluyveromyces lactis		
				Lipomyces starkeyi NRRL Y-11557 v1.0		
				Metschnikowia bicuspidata NRRL YB-4993 v1.0		
				Nadsonia fulvescens var. elongata DSM 6958 v1.0		
				Pachysolen tannophilus NRRL Y-2460		
				Pichia membranifaciens v2.0		
				Pichia pastoris		
				Pichia stipitis v2.0		
				Saccharomyces cerevisiae M3707 Dikaryon		
				Saccharomyces cerevisiae M3836 v1.0		
				Saccharomyces cerevisiae M3838 v1.0		
				Saccharomyces cerevisiae M3839 v1.0		
				Saccharomyces cerevisiae S288C		
				Saccharomyces cerevisiae YB210 Standard Draft		
				Spathaspora passalidarum NRRL Y-27907 v2.0		
				Wickerhamomyces anomalus NRRL Y-366-8 v1.0		
				Yarrowia lipolytica (strain CLIB122)		
Ascomycota	Taphrinomycotina			*Pneumocystis jirovecii*		
				Saitoella complicata NRRL Y-17804 v1.0		
				Schizosaccharomyces cryophilus OY26		
				Schizosaccharomyces japonicus yFS275		
				Schizosaccharomyces octosporus yFS286		
				Schizosaccharomyces pombe		
				Taphrina deformans		

Supdata 2. List of Fungi species harboring GalEa retrotransposons and of genomes tested. For each species concerned the acronym used during the study and the data source website are indicated. Classification was redrawn from MycoCosm. In green: assembled genomes harboring GalEa elements. In red: assembled genomes devoid of GalEa element. In orange: other species harboring GalEa elements.



GalEa of Pezizomycotina

GalEa of Metazoa

Tork

Supdata 3. Phylogenetic relationships among GalEa retrotransposons, including ripped elements. The tree is constructed using the Neighbor-Joining method on RT/RH amino acid sequences. Elements are colored according to their host: Fungi in bold blue, with ripped sequences being underlined; rhodophyta in Red; and metazoan in green. Statistical support (>50%) comes from non-parametric bootstrapping using 100 replicates. Retrotransposons of the Tork clade sequences were used as outgroup. (PPT)

Element	Host	Size (bp)	LTR				putative	GagPol	conserved motifs								
			Size	start	end		PolyPurine Tract	fusion		*pol*							
						Primer Binding Site			*gag*	protease	integrase	Reverse Transcriptase			RnaseH		
Corex1	*Rimicaris Exoculata*	4949	217	TGTtag	tatCA	TGGTAGCAGAGC	AAAGAAAGAGGACGAG	No	C(2)C(4)H	DTGC	D(53)D(35)E	KVHLVA	DVKAAY	YVDD	TRPDI	SDAS	ADCLTK
Corex2	*Rimicaris Exoculata*	4875	133	TGTtag	tatCA	TGGTAGCAGAGC	AAGAGAAAAAGGAGGGGGAGA		C(2)C(4)H(4	DTGC	D(53)D(35)E	KARLVA	DVKAAY	YMDD	TRPDV	SDAS	ANALTK
Galea1	*Eumunida annulosa*	4669	124	TGTgtg	tatCA	TGGTAGCAGAGC	GAAGAAATGGA	Yes	C(2)C(4)H(4	DSGC	D(49)D(35)E	KARLVA	DVKAAY	YVDD	TRPDI	SDAS	ADCLTK
Cico1	*Ciona intestinalis*	4513	323	TGTtaa	caaCA	TGGTAGCAGAGC	AAAAGAACAGAA	Yes	C(2)C(4)H(4	DTAC	D(54)D(35)E	KARLVA	DVKTAF	YVDD	TRPDI	ADAA	AVKSTK
Zeco1	*Danio rerio*	4807	192	TGTtaa	ttaCA	TGGTAGCAGAGC	AGGGAGA		C(2)C(4)H(4	DTAC	D(54)D(35)E	KARLVA	DIKSAF	HVDD	TRPDI	SDAS	ADCLTK
Olco1	*Oryzias latipes*	4760	187	TGTtaa	ttaCA	TGGTAGCAGAGC	AAAAAAAAAAAAA	Yes	C(2)C(4)H(4	DTAC	nd	KARLVA	DIKAAF	HVDD	SRPDV	SDAS	ADCLTK
						Conserved Hairpin Site											
CoTasti4	*Talaromyces stipitatus*	6282	219	TGTtgg	ccaCA	CTGATCAGT	AAAAGAGA	Yes	C(2)C(4)C(4	DTGA	D(54)D(35)E	KSRLVI	DITQAY	QTDD	CQPEA	TDGS	ADAFTK
CoTasti2	*Talaromyces stipitatus*	6485	218	TGTtgg	ccaCA	CTGATCAGT	AAAAGAGA	Yes	C(2)C(4)C(4	DTGA	D(54)D(35)E	KSRLVI	DITQAY	QTDD	CQPEA	TDGS	ADAFTK
CoChaglo1	*Chaetomium globosum*	6137	254	TGTcgg	ctaCA	CTGATCAGT	AAAAGAAGAGAC	Yes	C(2)C(4)C(4	DSGA	D(54)D(35)E	KSRLVV	DITQAY	QTDD	CQPEA	VDGS	ADSFTK
CoBofu2	*Botryotinia fuckeliana*	6385	163	TGTcgg	caaCA	CTGATCAGT	AGGAAGGAAGG	Yes	C(2)C(4)C(4	DTAA	D(55)D(35)E	KSRLVI	DITQAY	QTDD	CQPEA	TDAS	ADAMTK
CoOima1	*Oidiodendron malus*	5408	178	TGTata	aaaCA	CTGATCAGT	GAGAGAAGGGAGA		C(2)C(4)C(4	DTGA	D(54)D(35)E	KSRLVV	DITQAY	QTDD	SQPEA	TDAS	ADAMTK
CoAuhe1	*Aulographum hederae*	5757	213	TGTggc	cgaCA	CTGATCAGT	AAGGGGCAGA	Yes	C(2)C(4)C(4	DTGA	D(55)D(35)E	KSRLVV	DISQAY	QTDD	SQPEA	TDSS	ADAMTK
CoErypi2	*Erysiphe pisi*	5786	195	TGTatg	caaCA	CTGATCAGT	AAAAAAAAAAAAAAA	Yes	C(2)C(4)C(4	DTGA	D(55)D(35)E	KSRLVI	DISQAY	QTDD	CQPEA	VDAS	GDAMTK
CoBlugra5	*Blumeria graminis*	5640	203	TGTggg	caaCA	CTGATCAGT	AAAGAAAAGA		C(2)C(4)C(4	DTGA	D(54)D(35)E	KSRLVI	DITQAY	QTDD	CQPEA	TDGS	ADALTK

Supdata 4. Annotation of seven GalEa retrotransposons from Fungi. Element features and conserved motif sequences are compared to those of previously annotated metazoan GalEa retrotransposons. nd: not determined. Each newly identified element will be submitted to Repbase.