UNIVERSITÉ DE TOURS

ÉCOLE DOCTORALE : SSBCV

ÉQUIPE de RECHERCHE : Plasticité Génomique et Expression Phénotypique

THÈSE présentée par :

Laura Helou

soutenue le : **22 Avril 2021**

pour obtenir le grade de : **Docteur de l'université de Tours**

Discipline/ Spécialité : Sciences de la Vie

Définition des cibles chromosomiques de PGBD5 dans les neurones pyramidaux du système nerveux central des vertébrés

THÈSE dirigée par :

Mr BIGOT Yves	Directeur de Recherche, CNRS, Tours

RAPPORTRICES :

Mme BÉTERMIER Mireille	Directrice de Recherche, CNRS, Paris-Saclay
Mme LESAGE Pascale	Directrice de Recherche, CNRS, Paris

JURY :

Mme BÉTERMIER Mireille	Directrice de Recherche, CNRS, Paris-Saclay
Mme LESAGE Pascale	Directrice de Recherche, CNRS, Paris
Mme ROULEUX-BONNIN Florence	Maître de Conférence, Université de Tours
Mr PONTAROTTI Pierre	Directeur de Recherche, CNRS, Marseille
Mr LECOMTE Thierry	Professeur des universités, Université de Tours
Mr BIGOT Yves	Directeur de Recherche, CNRS, Tours

Remerciements

Je souhaite tout d'abord remercier la Région Centre Val de Loire pour le financement de ces 3 années de thèse. Je remercie Florian Guillou et Thierry Magalon de m'avoir accueillie au sein de l'Unité Physiologie de la Reproduction et des Comportements (PRC) de l'INRAE de Nouzilly. Je remercie aussi l'Ecole Doctorale Sciences Santé Biologie et Chimie du Vivant pour les opportunités des formations suivies.

Je remercie les membres du jury qui m'ont fait l'honneur d'évaluer mon travail de thèse : Mireille Betermier et Pascale Lesage d'avoir accepté d'en être les rapportrices et Florence Rouleux-Bonnin, Thierry Lecomte et Pierre Pontarotti d'en être les examinateurs.

Je remercie également les membres de mon comité de thèse : Hugues Dardente et Philippe Monget pour leurs critiques et conseils avisés sur mon travail.

Je souhaite remercier mon Directeur de Thèse, Yves Bigot de m'avoir offert une telle opportunité : Celle de réaliser une thèse sur un sujet qui n'a cessé de me challenger. Je te suis reconnaissante de m'avoir encadrée et conseillée tout en me laissant libre de mener mes travaux. Merci pour ta grande disponibilité, ton écoute, ton humour, pour la richesse de nos discussions, qu'elles furent scientifiques ou humaines. Merci de m'avoir toujours encouragée. Merci de m'avoir montré ce que c'était d'être un chercheur.

Cette thèse ne se serait pas passée ainsi sans le soutien de Benoît Piégu. Tes conseils en bioinformatique m'ont été plus que précieux. Merci pour les bons moments partagés, nos discussions allant de la parallélisation aux vins de Touraine en passant par les expressions régulières !

Un grand merci à Linda Beauclair. Merci de m'avoir accueilli dans ton bureau, de m'avoir fait de la place aussi vite (j'ai encore peur de retrouver mon bureau en pièces détachées). Merci pour ta patience sur mes questions incessantes sur les plasmides ou les protocoles. Sans toi mon expérience à Tours n'aurait pas été tout aussi fleurie et colorée.

Je remercie également l'équipe de Marseille avec qui j'ai eu beaucoup de plaisir à échanger et partager sur les transposons domestiqués. Pierre, Louis, Célia, un grand merci pour votre chaleureux accueil !

Je remercie toutes les personnes de l'INRAE qui m'ont apporté sur le plan scientifique ou humain dont certains sont devenus de véritables amis avec le temps : Chantal, Thierry, Olivier, Jérémy, Pierre-Marie, Anthony, Shifa, Alice, Erika, Agostinho, Priscila, Luis ainsi que ce que j'ai malheureusement oublié de citer mais qui ne m'ont pas moins marqué. Je remercie

le personnel administratif, Sophie Mary, Ghislaine Ploux, Marie Jaros et Manon Binoit ainsi que les personnes de l'atelier.

La thèse, ce n'est pas une aventure en solitaire. S'il y avait bien quelque chose à laquelle je ne m'attendais pas en quittant le sud pour les champs époustouflants de colza, c'est bien d'y rencontrer mes compagnons de bataille : Alix & Namya. Je gage que si très vite on m'a présenté à toi Namya, ce n'était pas un hasard. Je n'aurais pu rêver meilleure « Doyenne Doctorante » que toi pour partager mes doutes et mes joies. Ou pour recevoir ma dose de coups de pied. Alix, tu as été mon binôme de galère. De coup de gueule. De rires aux éclats. Avec toi, j'aurais pu aller au bout du monde en marchant (bon, Marrakech en période préapocalyptique/covidienne c'était déjà pas mal). Merci pour tout. Je compte bien trouver de nouvelles pistes de randonnées à étrenner où que tu sois !

Nous sommes tous (presque) exilés aux quatre coins de la France. Ça ne nous a pas empêchés de nous retrouver (pas autant qu'on l'aurait souhaité !). Merci aux amis, Marianne, Jake, Gonché, Quentin, Cynthia, Hugo, Abdou, Florian. Nos rencontres ont chaque fois été de véritables bouffées d'air frais.

Je remercie ma famille, mes proches, sans vous, tout ça n'aurait pas été possible. Je remercie mon compagnon, Jérôme, pour ton soutien durant nos 3 années de thèse. Les épreuves, ce n'est pas ce qui nous a manqué ! Ta patience, ton amour ont été une véritable source d'énergie et d'apaisement. Nos longues discussions sur nos sujets respectifs, nos idées folles de publications, nos doutes, nos peines, nos joies… Rien n'aurait été possible sans toi. J'embrasse aussi Marie-Laure, Michel & Michel pour leurs encouragements et leurs soutiens à chacune de nos rencontres. Merci à Isabelle pour toutes tes petites (et grandes !) attentions à mon égard, merci à Valentin. Vous comptez beaucoup pour moi.

Je remercie infiniment mes parents. Vous m'avez toujours appris à me dépasser. C'est grâce à vos encouragements, vos critiques, votre soutien sans faille que j'ai pu avancer et être là aujourd'hui. Merci pour votre soutien malgré la distance. Merci Maman, pour tes bons petits plats, ton sourire, ta joie, la force que tu m'as transmise. Merci Papa, ma détermination et mon acharnement au travail, je le tiens de toi. Ce travail, c'est mon travail, mais c'est aussi le vôtre. Il est le résultat de vos sacrifices, de vos efforts. J'espère pouvoir à mon tour vous rendre un peu de tout ce que vous m'avez donné. Il est difficile d'exprimer la reconnaissance infinie que je vous porte. Merci à toi, Nathan, mon « petit » frère, toi au Sud et moi au « Nord ». Tous les moments partagés ont sans nul doute contribué à mon avancement. Merci pour ce que tu es.

Je dédie cette thèse à mes parents et à mes grands-parents qui ont fait tant de sacrifices pour que nous ayons une vie meilleure.

Résumé

Les éléments transposables sont des éléments génétiques mobiles présents dans presque tous les génomes eucaryotes et procaryotes. Ils co-évoluent avec leurs hôtes et ont participé au façonnage de leur environnement génomique au cours de leurs évolutions. Ils sont en outre à l'origine de nouveaux gènes dérivés de séquences codantes de transposon par domestication.

Le gène qui code la transposase domestiquée *piggyBac*, PGBD5, est très conservé depuis l'origine des vertébrés. Cette transposase est fonctionnelle et est capable de transposer un élément *piggyBac* d'insecte, *Ifp2*. Elle s'exprime spécifiquement dans le système nerveux central et plus précisément dans certains neurones de l'hippocampe, de l'amygdale et du bulbe olfactif qui sont respectivement impliqués dans les mémoires spatiale, contextuelle, émotionnelle et olfactive. La conservation élevée de PGBD5 chez les vertébrés suggère que sa fonction et ses cibles chromosomiques le seraient également.

Mon travail de thèse traite de l'origine et de l'activité de la transposase domestiquée PGBD5 dans le génome des vertébrés.

Dans un premier temps, nous avons étudié les capacités de transposition d'une transposase piggyBac sans le domaine CRD. En effet, PGBD5 peut être caractérisé comme étant une transposase ayant perdu son domaine CRD au cours de la domestication. Nous avons vérifié la nécessité de la présence d'un domaine C-terminal pour l'activité de transposase-recombinase.

Dans un second temps, nous avons étudié l'histoire évolutive de *pgbd5* et des *pgbd* en général. Nous avons identifié le *pble* phylogénétiquement le plus proche de *pgbd5* et ses capacités à être transposé par PGBD5.

Enfin, nous avons étudié les capacités de PGBD5 à exciser et réintégrer dans de nouveaux locus des *pble* humains afin d'évaluer la capacité de PGBD5 à mobiliser ces éléments présents dans le génome humain et d'affiner l'identification des cibles de la transposase.

Abstract

Transposable elements are mobile genetic elements present in almost all eukaryotic and prokaryotic genomes. They co-evolve with their hosts and have participated in the shaping of their genomic environment during their evolution. In addition, they are at the origin of new genes derived from transposon coding sequences by domestication.

The gene that encodes the domesticated piggyBac transposase, PGBD5, has been highly conserved since the origin of vertebrates. This transposase is functional and capable of transposing an insect piggyBac element, *Ifp2*. It is specifically expressed in the central nervous system and more precisely in certain neurons of the hippocampus, the amygdala and the olfactory bulb which are respectively involved in spatial, contextual, emotional and olfactory memories. The high conservation of PGBD5 in vertebrates suggests that its function and chromosomal targets would also be highly conserved.

My thesis work aims to study the origin and activity of the domesticated PGBD5 transposase in the vertebrate genome.

In the first chapter, we have studied the transposal capacities of a piggyBac transposase without the CRD domain. Indeed, PGBD5 can be characterized as a transposase having lost its CRD domain during domestication. We have verified the necessity of the presence of a C-terminal domain for the transposase-recombinase activity.

In the second chapter, we studied the evolutionary history of *pgbd5* and *pgbd* in general. We identified the phylogenetically closest *pble* to *pgbd5* and its ability to be transposed by PGBD5.

In the last chapter, we have studied the ability of PGBD5 to excise and reintegrate into new loci of human *pbles* in order to evaluate the ability of PGBD5 to mobilize these elements present in the human genome and to refine the identification of transposase targets.

Liste des publications et communications

Liste des publications :

Helou L, Beauclair L, Dardente H, Arensburger P, Buisine N, Jaszczyszyn Y, Guillou F, Lecomte T, Kentsis A, Bigot Y. The C-terminal domain of piggyBac transposase is not required for DNA transposition. J Mol Biol. 2021 Jan 12:166805. doi: 10.1016/j.jmb.2020.166805. Epub ahead of print. PMID: 33450253.

Helou L, Beauclair L, Dardente H, Piégu B, Tsakou Ngouafo L, Lecomte T, Kentsis A, Pontarotti P, Bigot Y. The *piggyBac*-derived protein 5 (PGBD5) transposes both the closely and the distantly related *piggyBac*-like elements *Tcr-pble* and *Ifp2*. J Mol Biol. 2021. (Accepted)

Tsakou-Ngouafo L, Vicari C, **Helou L**, Keshri V, Das S, Bigot Y, Pontarotti P. (2020) DDE Transposon as Public Goods. In: Pontarotti P. (eds) Evolutionary Biology—A Transdisciplinary Approach. Springer, Cham. https://doi.org/10.1007/978-3-030-57246-4_14

Liste des communications orales :

Identification of more closely related *PiggyBac*-like elements of genes coding PiggyBac derived transposases among Vertebrates, using an evolutionary *in silico* approach. *23rd* Evolutionary Biology Meeting 2019, Marseille.

Caractérisation des sites de liaison de PGBD5, approche *in silico*. 32ème Colloque Biotechnocentre. 10 et 11 Octobre 2019. Domaine de Seillac, France. Prix Short Talk.

Caractérisation des sites de liaison de PGBD5 – Approche *in silico.* Journée Scientifique du Pôle BSIP, 12 Juin 2019, Tours.

Caractérisation des sites de liaison de PGBD5. 24ème Journée Thésards et 8ème Journée Post-Doctorants 2019, Unité PRC, UMR INRAE Tours.

Novelties about the evolution of PGBDs in vertebrates. PGBD-Day, 11 mars 2019, Hôtel Univers, Tours.

Définition des cibles chromosomiques de PGBD5 dans les neurones pyramidaux du système nerveux central des vertébrés. 23ème Journée Thésards et 7ème Journée Post-Doctorants, 18 Juin 2018, Unité PRC, UMR INRAE Tours.

Liste des présentations Poster :

Helou L, Beauclair L, Guillou F, Lecomte T, Bigot Yves. PGBD5, une transposase domestiquée ubiquitaire et très conservée chez les vertébrés : rôles dans les tissus sains et les tumeurs.

- 14ème Journée du Cancéropole Grand Ouest. 26 et 27 Octobre 2020.

Helou L, Beauclair L, Guillou F, Bigot Y. Identification of more closely related *PiggyBac*-like elements of genes coding PiggyBac derived transposases among Vertebrates, using an evolutionary *in silico* approach.

- *23rd* Evolutionary Biology Meeting 2019, Marseille.
- 32ème Colloque Biotechnocentre. 10 et 11 Octobre 2019. Domaine de Seillac. Prix Poster.

Helou L, Beauclair L, Guillou F, Piégu B, Bigot Yves. Définition des cibles chromosomiques de PGBD5 dans les neurones pyramidaux du système nerveux central des vertébrés.

- 31ème Colloque Biotechnocentre, 11 et 12 Octobre 2018, Domaine de Seillac.
- Non-Coding Genome, 7th Edition, 7 au 14 Février 2018, Institut Curie, Paris.

.

Table des matières

Remerciements 3
Résumé 5
Abstract 7
Liste des publications et communications 9
Table des matières 13
Table des tableaux 15
Table des figures 15
Liste des abréviations 16
Avant-propos 19
I. Éléments Transposables 25
1. Définition 25
2. Classification des ET chez les Eucaryotes 27
2.1. Une classification controversée 27
2.2. Les éléments de classe I 30
2.2.1) Les rétrotransposons à LTR 31
2.2.2) Les DIRS 31
2.2.3) Les PLE 32
2.2.4) Les LINE 32
2.2.5) Les SINE 32
2.3. Les éléments de classe II 33
2.3.1) Transposon TIR 33
2.3.2) Crypton 35
2.3.3) Hélitron 35
2.3.4) Polinton/Maverick 35
3. Dynamique des ET chez les eucaryotes 36
4. Impact des ET 38
II. *PiggyBac* 41
1. Le transposon *piggyBac* (*Ifp2*) 41
2. La Transposase *PiggyBac* (PB) 42
3. Mécanisme de transposition : excision et insertion 44
4. Devenir du site chromosomique donneur 45
5. Sites d'intégrations 45
6. Diversité des éléments de la famille piggyBac 47
III. Domestication de gènes contenus dans des éléments transposables. 51
1. DDE transposon as Public Goods 52
2. Domestication de gènes dérivés de transposons de la famille de piggyBac 75
2.1. Pgbd1 75
2.2. Pgbd2 77
2.3. Pgbd3 77
2.4. Pgbd4 78
2.5. Pgbd5 78
2.6. Pgbd6 79
2.7. NeoPgbd 80
Objectifs de la thèse 83
Résultats 87
Chapitre I. Le domaine C-terminal de la transposase piggyBac n'est pas nécessaire pour la transposition de l'ADN 89

1. Contexte et objectifs de l'article 89
2. Résumé des principaux résultats 89
Article 1: The C-terminal domain of piggyBac transposase is not required for DNA transposition 91
Chapitre II. La protéine dérivée de piggyBac 5 (PGBD5) transpose à la fois les éléments proches et éloignés de *piggyBac*, *Tcr-pble* et *Ifp2* 114
1. Contexte et objectifs de l'article 114
2. Résumé des principaux résultats 114
Article 2: The piggyBac-derived protein 5 (PGBD5) transposes both the closely and the distantly related piggyBac-like elements *Tcr-pble* and *Ifp2* 115
Chapitre III. Pble humains et pgbd5 177
1. Transposition des *pble* humains par PGBD5 177
2. Sites d'insertion de PGBD5 181
3. Sites de liaison de PGBD5 181
3.1. Souris 183
3.2. Homme 183
Chapitre IV. Outils informatiques mis en place 187
1. Recherche de *pble* 187
1.1. Démarche et sélection de candidats 187
1.2. G2tir 188
1.3. Estimation de la contrainte sélective sur les exons d'un gène 189
1.4. Phylogénie des transposases de pble et des pgbd 191
2. Pipeline d'analyse des sites d'insertions obtenus par LAM-PCR 192
3. Annotation des motifs IR dans un génome 193
4. Définition informatique des profils de liaison à l'ADN ou à la chromatine d'une protéine par analyse de données ChIPseq 195
Conclusion 197
Discussion et Perspectives 205
Références Bibliographiques 213

Table des tableaux

Tableau 1 - Distribution des *pble* humains dans hg38. 178
Tableau 2 - Conservation des séquences TIR (Terminal Inverted Repeat) et TSD (Target Site Duplication) des *pble* après événement de transposition par Mm523 dans les cellules HeLa. 180
Tableau 3 - Résultats du ChIP-séquençage à partir de différentes lignées cellulaires. . 182

Table des figures

Figure 1 - Comparaison et contenu de deux propositions pour la classification et l'annotation des ET eucaryotes. 26
Figure 2 - Représentation graphique des composants réplicatifs, intégratifs et structurels contribuant à la diversité des ET. 28
Figure 3 - Modification de l'expression d'un gène après l'insertion d'ET. 39
Figure 4 - Représentation schématique du transposon piggyBac 40
Figure 5 - Organisation des domaines de la transposase piggyBac. 42
Figure 6 - Mécanisme de transposition de piggyBac. 43
Figure 7 - Groupes de *pble* définis par Bouallègue et al, 2017 en fonction de l'organisation de leurs extrémités. 46
Figure 8 - Arbre phylogénétique de la superfamille piggyBac 48
Figure 9 - Répartition des *pgbd* chez les vertébrés. 74
Figure 10 - Caractéristiques des PGBD chez les vertébrés. 76
Figure 11 - Schéma récapitulatif des différentes étapes des tests de transposition dans les cellules G401. 176
Figure 12 - Résultats de transpositions dans les cellules G401 shGFP et shPGBD5 178
Figure 13 - Diagramme de Venn des différentes co-localisations des pics obtenus pour trois lignées cellulaires cancéreuses. 184
Figure 14 - Schéma du positionnement de la triade catalytique des transposases DDE en fonction des éléments de structure secondaire. 190

Liste des abréviations

aa : Acide aminé

ADN : Acide Désoxyribonucléique

ARN : Acide Ribonucléique

BWA : *Burrows-Wheeler Aligner*

Cas1 : *CRISPR-associated protein 1*

ChIP-Seq : *Chromatin ImmunoPrecipitation Sequencing*

CRD : *Cysteine-Rich Domain*

CRISPR : *Clustered Regularly Interspaced Short Palindromic Repeats*

DDBD : *Dimerization and DNA-binding domain*

DIRS : *Dictyostelium Intermediate Repeat Sequence*

DR : *Direct Repeat*

EGM : Éléments génétiques mobiles

EN : Endonucléase

ERV : *Endogenous retrovirus*

ET : Élément transposable

GAG : *Group-specific antigen*

GFP : *Green Fluorescent Protein*

HD : Hippocampe Dorsal

HT : *Horizontal transfert*

HTT : *Horizontal transfer of transposable elements*

HV : Hippocampe Ventral

Ifp2 : *Insertion in Few Polyhedral mutants*

IIR : *Inverted Intern Repeat*

INT : Intégrase

IR : *Inverted repeat*

IS : *Insertion sequence*

kDa : kilodalton

kpb : kilo paires de bases

LAD : *Lamina-Associated Domains*

LAM-PCR : *Linear-amplification mediated PCR*

LINE : *Long Interspersed Nuclear Elements*

LTR : *Long terminal repeats*

NeoR : Néomycine Résistant

NLS : *Nuclear Localization Sequence*

NTD : *N-Terminal Domain*

ORF : *Open Reading Frame*

pb : paires de bases

PB : *PiggyBac transposase*

pble : *piggyBac-like element*

PCR : *Polymerase Chain Reaction*

PGBD ou *pgbd* : *"piggyBac derived transposase" protein or gene*

PLE : *Penelope-like element*

Pol : Polymérase

RAG : *Recombinaison Activating Gene*

RH : Ribonucléase H

RT : *Reverse transcriptase*

SINE : *Short Interspersed Nuclear Element*

SNC : Système Nerveux Central

SR : *Serine recombinase*

STIR : *Sub-terminal Inverted Repeats*

SV40 : *Simian virus 40*

TIR : *Terminal Inverted Repeats*

TSD : *Target Site Duplication*

UTR : *Untranslated Region*

YR : *Tyrosine recombinase*

Avant-propos

Il y a encore deux siècles, nous n'avions aucune connaissance sur le support de l'information génétique. S'il est vrai qu'à l'époque des philosophes grecs, Aristote propose que les traits acquis durant toute la vie d'un organisme puissent être communiqués à leur progéniture. C'est en 1857 que Gregor Mendel, considéré comme « le père de la génétique », démontre par ses expériences la manière dont les caractères se transmettent de génération en génération. Quelques années plus tard, en 1869, Friedrich Miescher découvre dans le noyau des cellules une substance riche en phosphate appelée nucléine. Il faut attendre 1953 pour que la structure de la nucléine soit élucidée. La structure en double hélice de l'ADN composée de deux bases puriques et deux bases pyrimidiques est décrite par Rosalind Franklin, Francis Crick et James Watson. Une décennie plus tard, Robert W. Holley, Har Gobind Khorana et Marshall W. Nirenberg reçoivent le prix Nobel pour leurs travaux sur le processus de traduction du matériel génétique en une protéine, et ce par le biais de l'ARN. En 1977, la caractérisation de l'ADN permet à Frederick Sanger, Allan Maxam et Walter Gilbert de développer des méthodes de séquençage de l'ADN. Ces nouvelles technologies ont par ailleurs permis de séquencer, la même année, pour la première fois un génome complet, celui du virus bactériophage φX 1741. Ces technologies sont complétées par Kary Mullis qui, en 1983, invente la réaction en chaine par polymérase (PCR), une méthode de biologie moléculaire d'amplification génique *in vitro*.

Depuis cette époque, les technologies de séquençage sont en plein essor. Il est devenu possible de séquencer les génomes plus rapidement et à moindre coût. Le projet de séquençage du génome humain débutait en 1989, 13 ans et un investissement total de 2,7 milliards de dollars ont été nécessaires pour produire un génome complet. Avec l'avènement des technologies de séquençage à Haut Débit (HTS ; High-Throughput Sequencing ou encore NGS ; Next Generation Sequencing), il est possible de séquencer ce même génome en une semaine pour un coût d'environ 1000 euros. Ces nouvelles technologies ont permis de rendre possible l'étude de problématiques biologiques restées longtemps impossibles à aborder comme l'analyse comparative des insertions d'éléments transposables (ET). Elles ont notamment révélé la sous-estimation du contenu en ET des génomes.

L'annotation du génome humain montre qu'il est constitué de 23 000 gènes codant des protéines. Les ET représentent plus d'un tiers du génome humain (environ 41 % dans hg38) sachant que les exons ne représentent que 2 % de la totalité du génome, 98 % du génome correspond à l'ADN non codant. Pendant plusieurs années, on qualifiait cet ADN non codant d'ADN poubelle. Aujourd'hui, nous savons qu'il est constitué de séquences répétées (en tandem ou dispersées), d'introns et d'ADN non codant unique (non répété). L'analyse approfondie de ces éléments nous permet de ne plus les considérer comme « non fonctionnels »

mais, au contraire, comme des séquences jouant un rôle prédominant dans le fonctionnement d'un organisme.

Mon travail de thèse traite de l'origine et de l'activité de la transposase domestiquée PGBD5 dans le génome des Vertébrés. La première partie de mon manuscrit concerne une revue bibliographique générale sur les ET, le transposon *piggyBac* et les éléments de sa famille qui lui sont apparentés, et la domestication des ET. À la fin de cette partie sont posées les questions abordées au cours de ma thèse. La seconde partie est consacrée à la présentation des résultats selon trois volets correspondants à nos trois axes d'étude. Dans la troisième partie, les résultats les plus marquants sont synthétisés et discutés.

Introduction

I. Éléments Transposables

1. Définition

Pendant très longtemps, le génome d'un organisme procaryote ou eucaryote était considéré comme stable. Cependant en 1944, les travaux de Barbara McClintock portant sur les instabilités phénotypiques chez le maïs dévoilent le contraire (McClintock, 1950). Elle a observé que la couleur des grains de maïs change dans différents croisements de sélection. Elle a interprété ses résultats en matière de régulation de l'activité des gènes comme étant dus à une sorte d'« éléments de contrôle » qui ont la capacité de se déplacer d'un endroit à l'autre dans les chromosomes. Ces éléments se déplaçant lorsque les cellules se divisent et prolifèrent, ils ne mutent les gènes que dans certaines cellules. Les changements de couleur des grains de maïs sont ici dus à l'impact de l'excision et/ou de l'insertion de ces éléments sur les gènes de pigmentation. Ces travaux ont donc été les premiers à démontrer que l'instabilité phénotypique observée chez le maïs est le reflet de réarrangements dû à la mobilité de séquences d'ADN (McClintock, 1950).

Il a fallu attendre plusieurs années pour que l'intérêt pour ces éléments se développe. Une autre découverte au début des années 1970 entraine un nouvel intérêt pour ces éléments mobiles. Il s'agit de la découverte des séquences d'insertion (IS) chez *Escherichia coli* par l'équipe Szybalski (Malamy, Fiandt and Szybalski, 1972). Dans leur article, ils indiquent que « Les IS sont présents dans des organismes supérieurs et qu'ils sont facilement transposés d'un site à un autre dans le génome. Ces éléments IS sont analogues à ceux identifiés par McClintock dans le maïs. ». Cette seconde découverte souligne l'importance des travaux pionniers et originaux de McClintock auprès de la communauté génétique.

En même temps que Szybalski « redécouvrait » les ET, Susumu Ohno a popularisé le terme « junk DNA » pour ADN poubelle qui a influencé le domaine de la génomique pendant des décennies [(Ohno, 1972) ; bien que le terme lui-même fut déjà utilisé avant (Aronson *et al.*, 1960; Ehret and De Haller, 1963). Ohno faisait référence aux séquences dites non codantes, à tout morceau d'ADN qui ne code pas de protéine. L'image négative des éléments transposables a commencé à changer au début des années 1990 quand certains chercheurs ont remarqué la valeur évolutive de ces éléments (Brosius, 1991; Makałowski, Mitchell and Labuda, 1994). Par exemple l'une des plus vastes enquêtes bibliographiques sur la contribution des ET à la régulation des gènes de l'hôte a identifié environ 80 cas où les éléments régulateurs des gènes de vertébrés sont dérivés d'ET (Brosius, 1999).

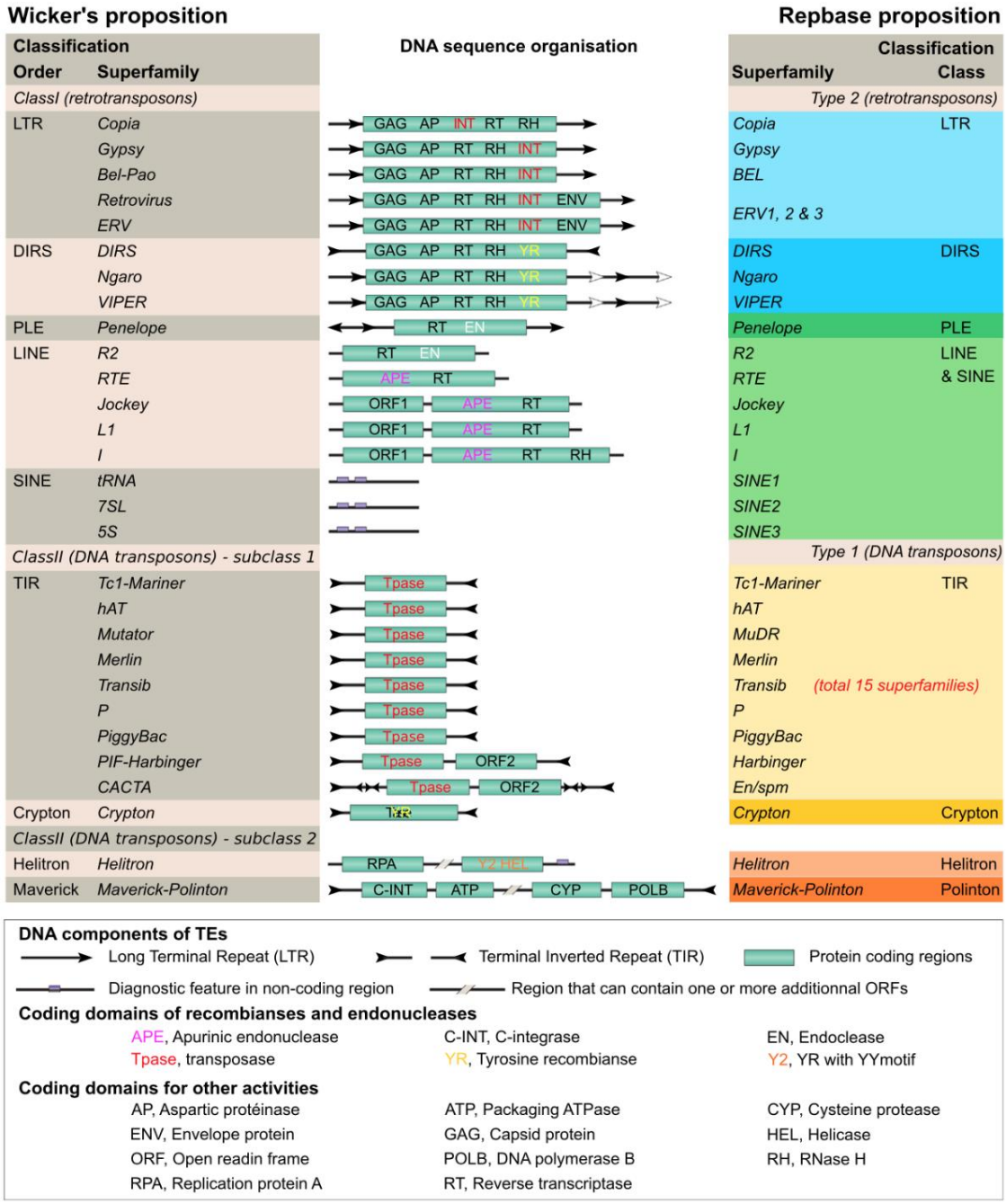


Figure 1 - Comparaison et contenu de deux propositions pour la classification et l'annotation des ET eucaryotes.

La proposition Repbase est présentée sur la droite (Jurka *et al.*, 2005; Kapitonov and Jurka, 2008) et la proposition Wicker sur la gauche (Wicker *et al.*, 2007). Ces propositions sont basées sur les caractéristiques de la séquence ADN et d'acides aminés. Les deux propositions divisent les ET en deux groupes, les rétrotransposons et les transposons d'ADN. Cette division basale est appelée le niveau « type » dans la proposition Repbase et un niveau « classe » dans la proposition Wicker. Chacun de ces deux types ou classes est ensuite subdivisé en « classes » dans la proposition Repbase ou en « ordres » dans la proposition Wicker. Dans l'ensemble, les « classes » de Repbase et les « ordres » de Wicker sont très similaires et chaque groupe contient les mêmes superfamilles d'ET. Une représentation schématique de l'organisation des séquences d'ADN au sein de chaque superfamille d'ET est fournie au milieu de la figure. Une légende de symboles est fournie au bas de la figure. [Source : (Piégu *et al.*, 2015)]

Les ET sont des éléments génétiques mobiles présents dans presque tous les génomes eucaryotes et procaryotes. Ces éléments sont capables de transposer, c'est-à-dire de se déplacer d'un locus à un autre au sein du génome qui les contient. La mobilité des ET peut induire l'apparition de mutations délétères, de réarrangements chromosomiques, mais la domestication de certains transposons apporte un avantage sélectif pour l'espèce. Les activités mutationnelles des ET contribuent donc à la diversité génétique des organismes (Chénais *et al.*, 2012). Les ET sont des moteurs majeurs de l'évolution des génomes en étant notamment une source de diversification fonctionnelle. Ils sont de différents types et sont en nombre variables dans les génomes selon les espèces.

À partir des années 2000, les programmes de séquençage à grande échelle des génomes ont permis de montrer que les ET constituaient une part importante du génome chez la plupart des eucaryotes. Ils constituent 44 % du génome humain (Lander *et al.*, 2001), 60 % chez le maïs *Zea mays* et jusqu'à 77 % dans le génome de la grenouille *Rana esculenta* (Biémont and Vieira, 2006). Cette quantité d'ET peut cependant être moindre dans certains organismes comme la drosophile *Drosophila melanogaster* dont la composition varie de 15 à 20 % du génome, et chez la levure *Saccharomyces cerevisiae*, 3 à 5 %. D'abord essentiellement étudiés du point de vue de leur annotation, de leurs mécanismes de mobilité et de leur évolution au sein des génomes, les éléments transposables sont de plus en plus étudiés pour l'impact qu'ils ont sur les génomes. Avec l'avancée de la génomique, la recherche sur les ET se concentre sur le rôle de ses éléments dans l'évolution de la régulation des gènes.

2. Classification des ET chez les Eucaryotes

2.1. *Une classification controversée*

Plusieurs systèmes de classification des ET ont été proposés dans la littérature pour les eucaryotes. Deux sont principalement utilisées dans la littérature scientifique. Le premier système de classification a été proposé par Finnegan en 1989 (Finnegan, 1989). Il distingue deux classes d'ET caractérisés par leur intermédiaire de transposition : ARN (classe I ou rétrotransposons) ou ADN (classe II ou transposons à ADN). Le mécanisme de transposition de la classe I est communément appelé « copier-coller » et celui de la classe II est appelé « couper-coller ». En 2007, un deuxième système de classification est présenté par Wicker et ses collaborateurs (Wicker *et al.*, 2007). Ils proposent une classification hiérarchique basée sur les caractéristiques structurelles et les modes de réplication des ET (cependant, ceux-ci à l'époque sont encore très mal connus, voire non élucidés pour la grande majorité des ET).

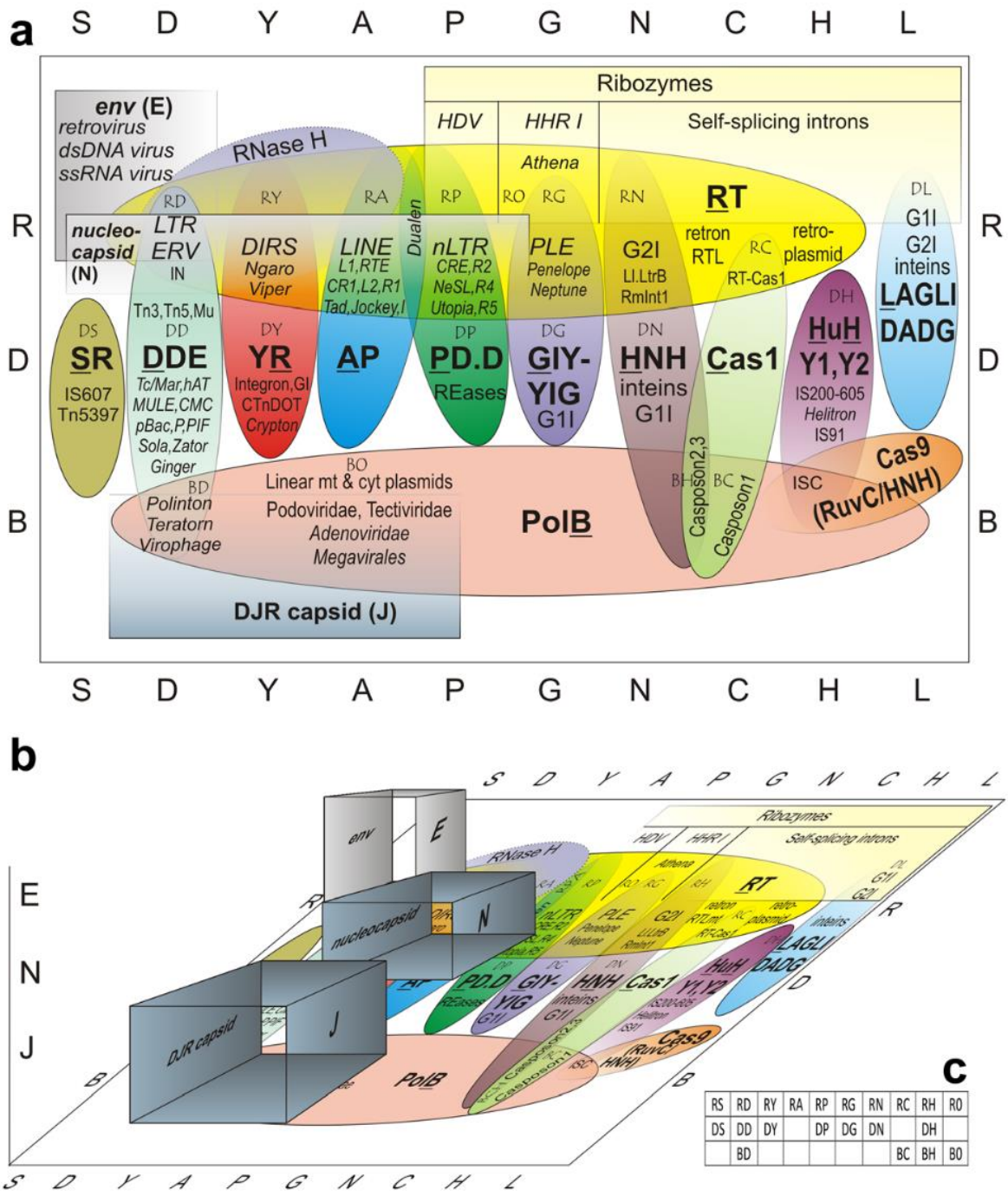


Figure 2 - Représentation graphique des composants réplicatifs, intégratifs et structurels contribuant à la diversité des ET.

(a) Diversité des combinaisons de polymérase-phosphotransférase dans les éléments mobiles. Les principaux types de polymérases et d'endonucléases sont en gras, et sont également représentés par des codes à lettre unique le long des deux axes respectifs. Les combinaisons à deux lettres sont indiquées pour chaque type d'ET aux intersections. **(b)** Même chose, avec l'ajout de composants structurels dans la troisième dimension. **(c)** Une grille 2-D énumérant les combinaisons actuellement connues de polymérases et d'endonucléases. Les quelques autres types d'endonucléases que l'on ne trouve que dans les introns du groupe I ne sont pas indiqués pour les combinaisons simples de polymérases et d'endonucléases. Source : (Arkhipova, 2017).

Il est à noter que les classifications de Wicker et celle proposée par le GIRI (propriétaire de Repbase) et l'Institut of System Biology (ISB) (Kapitonov and Jurka, 2008) présentent les mêmes critères de classification (Kapitonov and Jurka, 2008) ; (Piégu *et al.*, 2015)].

Malgré la volonté de réaliser une classification unifiée des éléments transposables, ces dernières restent néanmoins incomplètes. Plusieurs limites à ces classifications ont pu être mises en avant ces dernières années (Piégu *et al.*, 2015; Arensburger, Piégu and Bigot, 2016; Arkhipova, 2017). Ces classifications ont été définies essentiellement par des scientifiques issus du monde des eucaryotes, elles ne prennent pas en compte l'ensemble des mécanismes de transpositions et sont axées principalement sur les caractéristiques structurales. Plusieurs études (Piégu *et al.*, 2015; Arensburger, Piégu and Bigot, 2016; Arkhipova, 2017) soulignent l'importance d'inclure d'autres caractéristiques afin de classifier les ET, notamment en se basant sur les trois caractéristiques d'un ET : son mécanisme d'intégration, son mécanisme de réplication, l'usage de protéines structurelles généralement impliquées dans le trafic intra et intercellulaire de ces éléments (**Figure 2**). En plus de ces caractéristiques, les ET peuvent coder d'autres enzymes ou protéines de structure qui peuvent affecter l'efficacité de la prolifération des ET ou le degré de contrôle de l'hôte. Les ET peuvent transporter des gènes qui peuvent être utiles à l'hôte (ex. : résistance aux antibiotiques). Mais ces gènes ne sont pas nécessaires pour les fonctions de mobilité et peuvent être définis comme moins pertinents pour la classification car ils peuvent apparaitre et disparaitre épisodiquement (Arkhipova, 2017). Lorsque l'on s'intéresse aux fonctions réplicatives, intégratives et structurelles (**Figure 2**), on peut se rendre compte de la difficulté à établir des groupes. Par exemple, le chevauchement des gènes Pol (transcriptase inverse) et Int (intégrase) chez certains rétroéléments, c'est-à-dire des gènes codant respectivement les enzymes de réplication et d'intégration, ou leur absence, créent potentiellement des catégories d'ET distinctes.

En raison de l'absence de domaines protéiques conservés dans 100 % des transposases codées par des transposons à ADN, la classification de nombre de ces éléments peut être considérée comme étant moins bien établie que celle des rétrotransposons. Le mécanisme de transposition est le socle de la classification des ET. En général, ce mécanisme est étroitement lié aux enzymes codées par les ET. Si l'on considère ensemble les ET eucaryotes et procaryotes, Kojima classe les transposases codées par les transposons ADN en six types (Kojima, 2018) : DDD/E transposase, DEDD transposase, tyrosine recombinase (YR), sérine recombinase (SR), nucléase HUH et Cas1 nucléase (Siguier *et al.*, 2006; Chandler *et al.*, 2013; Krupovic *et al.*, 2014). Parmi ceux-ci, les DEDD transposases, les SR et les protéines affiliées à la nucléase

Cas1 n'ont pas été identifiées dans les ET d'eucaryote. Par exemple, les casposons partagent plusieurs caractéristiques avec les transposons ADN eucaryotes Polinton/Maverick comme les TIR et les ADN polymérase de la famille B. Mais, contrairement à tout autre élément mobile connu, les Casposons utilisent une endonucléase Cas1 (CRISPR-associated protein 1) pour s'intégrer dans le génome de l'hôte via un mécanisme similaire à l'intégration de nouveaux spacers dans les locus CRISPR (Krupovic *et al.*, 2014). Ceci a pour conséquence qu'ils ne trouvent pas leur place dans la classification de Wicker. Par ailleurs, d'autres éléments mobiles, comme les rétrozymes ou les Zisupton ne trouvent pas leur place dans les classifications existantes. Les premiers ne codent pas d'enzyme protéique, juste une molécule d'ARN ayant des propriétés enzymatiques, probablement nucléasiques (de la Peña and Cervera, 2017). Les seconds codent des enzymes qui ne sont pas connues pour être impliquées dans la transposition, le clivage ou la recombinaison de fragment d'ADN (Bohne *et al.*, 2012).

La difficulté d'établir une classification représentant au mieux la réalité et résolvant les lacunes des classifications existantes montre la complexité et la diversité des ET. Le défi d'établir un recensement des ET est un challenge complexe malgré les initiatives pour établir un nouveau consensus reposant sur l'utilisation de critères de classifications prenant en compte les mécanismes de transposition ainsi que les enzymes utilisées au cours de la transposition (Piégu *et al.*, 2015).

Malgré les limites soulignées précédemment, je présenterai dans la suite du manuscrit les ET selon la classification la plus utilisée dans la littérature scientifique, celle dite de Wicker (Wicker *et al.*, 2007).

2.2. *Les éléments de classe I*

Les ET de classe I transposent en utilisant un intermédiaire ARN. L'ARN intermédiaire est transcrit à partir de la copie d'un ET dans l'ADN génomique par une ARN polymérase de la cellule hôte, puis il est reverse-transcrit en ADN par une transcriptase reverse (RT) codée par ces ET. Ce second intermédiaire ADN est ensuite intégré à un autre locus chromosomique du génome. Chaque événement de mobilité produit une nouvelle copie par duplication, ce qui pourrait expliquer pourquoi les éléments de classe I sont les principaux contributeurs à la fraction de séquences répétées dans les grands génomes. Dans la classification de Wicker, les rétrotransposons sont divisés en cinq ordres : les rétrotransposons à LTR, les éléments de types DIRS (*Dictyostelium* intermediate repeat sequences), les éléments de type Penelope (PLE), les LINE (éléments longs dispersés) et les SINE (éléments courts dispersés).

2.2.1) Les rétrotransposons à LTR

Les rétrotransposons à LTR sont caractérisés par la présence de longues répétitions terminales (LTR) dont la taille va de plusieurs centaines à plusieurs milliers de paires de bases. Les rétrovirus exogènes et les rétrotransposons à LTR contiennent un gène *gag* qui code une protéine avec un domaine conservé de nucléocapside. Les protéines GAG forment des particules de type viral dans le cytoplasme à l'intérieur duquel se produit la transcription inverse. Le gène *Pol* produit une polyprotéine contenant une transcriptase inverse (RT), une ribonucléase H (RH) et une intégrase (INT) qui fournissent la machinerie enzymatique pour la transcription inverse et l'intégration dans le génome de l'hôte. Contrairement aux rétrotransposons à LTR, les rétrovirus exogènes contiennent un gène *env* qui code une protéine d'enveloppe qui permet leur excrétion et facilite leur migration vers d'autres cellules. Certains rétrotransposons à LTR peuvent contenir des restes d'un gène *env* non fonctionnel. Leurs capacités d'insertion sont limitées au génome de la cellule qui les héberge (Kazazian, 2004). Cependant, les rétrotransposons à LTR peuvent réacquérir le gène *env* par recombinaison et redevenir des agents infectieux (Malik, Henikoff and Eickbush, 2000). Actuellement, la plupart des séquences LTR (85 %) du génome humain ne se trouvent que sous forme de LTR isolés, les régions internes du rétrotransposon ayant été perdues par recombinaison homologue entre LTR adjacents (Leib-Mösch *et al.*, 1993). Il est intéressant de noter que les rétrotransposons à LTR se réinsèrent préférentiellement dans des régions génomiques spécifiques, souvent autour des gènes, avec un impact fonctionnel potentiellement important sur l'expression des gènes de l'hôte (Kazazian, 2004).

2.2.2) Les DIRS

L'ordre des DIRS (*Dictyostelium* intermediate repeat sequences) regroupe des transposons structurellement divergents qui possèdent un gène de la tyrosine recombinase (YR) au lieu d'une intégrase (INT). Leur intégration ne produit pas de duplication du site d'insertion (aussi appelé TSD [target site duplication]). Leurs extrémités ressemblent soit à des répétitions directes, soit à des répétitions inversées. Ces caractéristiques indiquent un mécanisme d'intégration différent des autres éléments mobiles de classe I. Les DIRS ont été découverts dans le génome de la moisissure visqueuse, *Dictyostelium discoideum*, au début des années 1980 et ils sont présents dans toutes les lignées d'eucaryotes, y compris les vertébrés (Goodwin and Poulter, 2001).

2.2.3) Les PLE

L'ordre des éléments de type Penelope (PLE) possède un seul ORF qui code pour une protéine contenant des domaines de transcriptase inverse (RT) et d'endonucléase (EN). En fonction du PLE, les domaines EN peuvent avoir des origines endonucléases différentes. Certains membres des PLE ont des séquences de type LTR qui peuvent avoir une orientation inverse et posséder un intron fonctionnel. Les PLE ont une distribution large, mais inégale, allant des amibes et des champignons, aux vertébrés avec un nombre de copies pouvant atteindre des milliers par génome (Evgen'ev and Arkhipova, 2005). Il est intéressant de noter qu'aucune séquence PLE n'a été trouvée dans les génomes des mammifères, et qu'elle a apparemment été perdue dans le génome de *Caenorhabditis elegans*. Bien que des PLE avec un ORF intact aient été trouvées dans plusieurs génomes, y compris *Ciona intestinalis* et *Danio rerio*, le seul représentant transcriptionnellement actif, Penelope, est connu de *Drosophila virilis*.

2.2.4) Les LINE

Les LINE forment un groupe de rétrotransposons sans LTR très répandus dans le génome de nombreux eucaryotes (Singer, 1982; Jurka, 1998). D'après les caractéristiques structurelles et la phylogénie de leur enzyme clé, la transcriptase inverse (RT), les LINE sont divisés en cinq groupes principaux, les *R2*, *RTE*, *Jockey*, *L1* et *I*. Tous les LINE codent pour au moins une protéine qui contient un domaine RT et un domaine d'endonucléase (EN), soit un APE N-terminal, soit un RLE C-terminal (rarement les deux). Un domaine de ribonucléase H est présent dans la superfamille *I*. À l'exception des anciennes superfamilles *R2* et *RTE*, les LINE codent généralement pour une autre protéine appelée ORF1, une protéine de liaison à l'ARN non spécifique qui contient des motifs en doigt de zinc et de leucine « zipper ». Bien que les LINE génèrent en général un TSD lors de leur insertion, la troncature de leur extrémité 5' au moment de l'insertion les rend difficiles à localiser (Eickbush and Malik, 2002). Ces « troncatures » résultent de l'arrêt prématuré de la transcription inverse de l'intermédiaire ARN. À l'extrémité 3' d'un LINE, on peut retrouver soit une queue poly (A), soit des répétitions en tandem, soit simplement une région riche en A (Goodier, 2016).

2.2.5) Les SINE

Les SINE forment un groupe de rétrotransposons non autonomes et d'origines diverses. En d'autres termes, les SINE présents dans une espèce ou dans une lignée d'espèces peuvent ne pas être apparentés. Entre lignées d'espèces, les SINE sont en général non apparentées. Leur

regroupement sous forme d'un ordre pourrait donc être excessif puisqu'un ordre est censé regrouper des entités biologiques ayant un ancêtre commun.

Les SINE identifiées à ce jour ont chacun une taille qui, entre « espèces de SINE », varie de 80 à 1000 pb. L'origine de chaque SINE provient de la rétrotransposition accidentelle d'un des divers transcrits de l'ARN polymérase III [Pol III ; (Kramerov and Vassetzky, 2005)]. Chaque « espèce » de SINE utilise la machinerie enzymatique d'une espèce de LINE pour transposer. Certains SINE ont un partenaire LINE unique et obligatoire (Dewannieux, Esnault and Heidmann, 2003) alors que d'autres peuvent utiliser la machinerie de plusieurs LINE (Rowold and Herrera, 2000). Les SINE sont petits (80-500 pb) et dupliquent un TSD (5-15 pb) au moment de leur insertion. La « tête » qui abrite le promoteur Pol III, définit les superfamilles de SINE et révèle leur origine : ARNt, ARN 7SL, ARN 5S, ou chimère de plusieurs éléments [Ex. : SVA chez l'humain (Hancks and Kazazian, 2010)]. Les régions internes du SINE (50-200 pb) sont spécifiques à une famille et d'origine variable, dérivant parfois de la dimérisation ou de la trimérisation du SINE. L'origine de la région 3', bien qu'elle puisse parfois venir d'un LINE, est généralement obscure. Elle peut être riche en A ou AT, abriter des répétitions en tandem de 3 à 5 pb, ou contenir une queue poly (T), le signal de terminaison de la Pol III. Les éléments Alu chez l'humain sont les plus connus, on en retrouve au moins 1 000 000 exemplaires dans le génome humain (Hancks and Kazazian, 2012).

2.3. *Les éléments de classe II*

Les ET de classe II, comme ceux de classe I, sont présents dans presque tous les génomes eucaryotes. Les éléments de classe II se retrouvent également chez les procaryotes sous des formes simples appelées séquences d'insertion (IS). Wicker définit deux sous-classes qui se distinguent par le nombre de brins d'ADN qui sont coupés lors de la transposition. En raison de l'absence de domaines protéiques conservés parmi les transposons ADN, la classification des transposons est moins largement acceptée que celle des rétrotransposons. Pour rappel ; le mécanisme de transposition est le cadre pour la classification des ET. En général, celui-ci est étroitement lié à la composition des domaines protéiques codés par les ET (Kojima, 2018).

2.3.1) **Transposon TIR**

Ils sont caractérisés par la présence de répétitions terminales inversées (TIR) de longueur variable à leurs extrémités. Ils codent une transposase qui se lie à ces répétitions inversées et assure la catalyse de toutes les étapes de la transposition. Dans la classification de

Wicker, neuf superfamilles sont décrites et elles se distinguent par la séquence des TIR, la taille des TSD et la phylogénie des transposase (Wicker *et al.*, 2007). Ces TIR transposons peuvent augmenter leur nombre de copies en transposant juste après le passage de la fourche de réplication chromosomique, en s'excisant d'une position déjà répliquée dans une autre, avant que la fourche de réplication ne passe. Ils peuvent également augmenter leur nombre de copies en exploitant la réparation par recombinaison homologue du locus excisé par le locus non excisé juste après le passage de la fourche de réplication (Nassif *et al.*, 1994; Capy *et al.*, 1998).

La **superfamille *Tc1-mariner***, ubiquitaire chez les eucaryotes, possède une structure simple de deux TIR et d'une ORF de transposase. Ils s'insèrent préférentiellement près d'un TA, générant alors des TSD « TA ». Les membres de la **superfamille *hAT*** ont des TSD de 8bp, des TIR plutôt courts allant de 5 à 27 pb et une taille totale inférieure à 4 kb. La **superfamille *Mutator*** est aussi présente chez tous les eucaryotes. Bien que ses TIR puissent s'étendre à plusieurs centaines de paires de bases, ils sont parfois très courts ou indétectables. Les éléments *Mutator* produisent des TSD de 9 à 11 pb. Dans la **superfamille *Merlin***, les TIR s'étendent de quelques dizaines à plusieurs centaines de paires de bases, ils sont flanqués par des TSD de 8-9 pb. Ces éléments n'ont été décrits que chez les animaux et les eubactéries. La **superfamille *Transib*** possède une transposase à motif DDE, elle est liée à la protéine RAG1 impliquée dans la recombinaison V (D) J des gènes des immunoglobulines et des récepteurs des cellules T (Kapitonov and Jurka, 2005). Des ET *transib* ont été retrouvés chez *D. melanogaster* et chez les moustiques. La **superfamille *P***, initialement présente dans les génomes d'insectes, est maintenant connue pour être aussi présente chez les métazoaires. Ces éléments génèrent des TSD de 8 pb. La **superfamille *piggyBac*** s'insère préférentiellement près d'un TTAA et ces membres sont présents chez tous les eucaryotes. Cette superfamille sera présentée plus en détail dans la section (II. PiggyBac). La **superfamille *PIF-Harbinger*** possède un site d'insertion préférentiel de séquence « TAA ». Ces ET possèdent deux ORF, une codant pour une transposase DDE et une codant pour une protéine de liaison à l'ADN. Les éléments de la **superfamille des *CACTA*** possèdent une transposase et un ORF dont la fonction est encore inconnue (Wicker *et al.*, 2003). Chez les plantes, les TIR courts se terminent par des motifs cAcTA (parfois cAcTG) hautement conservés. Ils sont bordés par des TSD de 3 pb, tandis que chez les animaux et les champignons, un motif ccc remplace le motif cAcTA à l'extrémité des TIR et des TSD de 2 pb sont générés (Bureau and Wessler, 1994; Wicker *et al.*, 2007).

2.3.2) **Crypton**

Wicker inclut les éléments *Crypton* dans la sous-classe I, mais dans un second ordre, ils ont été identifiés chez les animaux, les champignons, les oomycètes et les diatomées (Kojima and Jurka, 2011). Ils possèdent un ORF codant une tyrosine recombinase, mais il leur manque un domaine RT, suggérant qu'ils se transposent via un intermédiaire ADN. Ils ne possèdent pas de TIR, mais semblent générer des TSD suite à leur intégration.

2.3.3) **Hélitron**

La sous-classe II comprend deux ordres d'ET qui, comme ceux de la sous-classe I, ne forment pas d'ARN intermédiaire. Cependant, contrairement aux transposons ADN « classiques », ils se répliquent en utilisant une molécule d'ADN simple brin comme intermédiaire.

Les *hélitrons* se répliquent en utilisant un mécanisme de cercle roulant et leur insertion n'entraine pas la duplication du site d'insertion (Kapitonov and Jurka, 2001). Ils codent une tyrosine recombinase avec d'autres protéines impliquées entre autres dans la réplication. Les *hélitrons* ont été décrits pour la première fois dans des plantes, mais ils sont présents dans d'autres phyla comme les champignons et les animaux (Hood, 2005; Pritham and Feschotte, 2007).

2.3.4) **Polinton/Maverick**

Les transposons *Polinton/Maverick* sont de grands transposons qui ont été trouvés dans des lignées eucaryotes à l'exception des plantes (Pritham, Putliwala and Feschotte, 2007). Leurs analyses génomiques et phylogénétiques montrent qu'ils seraient des formes endogènes probablement chimériques de plusieurs familles de virus (Krupovic and Koonin, 2015; Krupovic, Yutin and Koonin, 2016). Ils codent plusieurs protéines dont une ADN polymérase B et une intégrase. Kapitonov et Jurka ont suggéré que leur cycle de vie comprend une excision simple brin suivi d'une réplication extrachromosomique et d'une réintégration dans un nouveau locus (Kapitonov and Jurka, 2006). Cette hypothèse n'est actuellement pas confirmée par des données moléculaires, et la synthèse disponible sur ce sujet montre qu'ils pourraient utiliser plusieurs mécanismes d'intégration (Krupovic and Forterre, 2015).

3. Dynamique des ET chez les eucaryotes

Les ET sont des constituants ubiquitaires des génomes eucaryotes. Ils peuvent se multiplier et coloniser le génome, les copies d'ET sont héritées de manière verticale, des parents à la descendance. Ils peuvent aussi être transmis d'une espèce à une autre, par transfert horizontal (HT) pouvant impliquer un organisme vecteur comme un virus, une bactérie ou un parasite hématophage (Gilbert and Feschotte, 2018). Le transfert horizontal d'ET (HTT pour Horizontal Transfer of Transposable Element) est souvent mis en évidence à l'aide de trois critères : incongruités phylogénétiques entre les phylogénies des hôtes et celles des ET, répartitions inégales des ET dans les phylogénies et fortes similarités entre les ET d'espèces éloignées.

La dynamique ou le cycle des ET peut être divisé en trois phases : l'invasion, la maturité et la sénescence (Kidwell and Lisch, 2001). La dynamique la plus commune est l'apparition d'un (ou plusieurs) cycle invasion-régression, généralement suivi d'un événement de domestication, avant la perte définitive de la famille (Le Rouzic, Boutin and Capy, 2007). La première phase est nécessairement l'invasion de la population après l'arrivée d'une nouvelle séquence d'ET active dans le génome d'une espèce, soit par réactivation d'une copie précédemment inactive ou, plus vraisemblablement, par transfert horizontal (Kidwell, 1992; Sánchez-Gracia, Maside and Charlesworth, 2005). Dans une telle situation, un élément doit être immédiatement actif et efficace sur le plan de la transposition ; sinon, sa perte est probable en raison des mutations et de la sélection naturelle. Cette étape d'invasion a déjà été observée expérimentalement chez des espèces modèles telles que *Drosophila melanogaster* avec l'élément P (Anxolabéhère, Kidwell and Periquet, 1988; Biémont *et al.*, 1994; Kimura and Kidwell, 1994). Certains éléments sont donc capables d'envahir efficacement une nouvelle espèce.

C'est lors de la seconde phase que les mécanismes de régulation et l'apparition de mutation limitent l'invasion. Ainsi, lors de l'invasion d'un génome par une copie d'ET, si elle transpose à un taux constant peu élevé, elle risque de ne pas se maintenir au sein du génome, notamment à cause de la dérive génétique. La dérive génétique est un phénomène qui entraine la variation stochastique des fréquences alléliques dans une population (Le Rouzic and Capy, 2005). Cependant, si l'ET est trop actif, il peut induire des dégâts dans le génome de l'hôte pouvant entrainer la stérilité de l'organisme hôte et donc la perte de l'ET. Seuls les ET dont la transposition est régulée peuvent se maintenir dans un génome. En effet, la transposition des

ET peut s'avérer néfaste pour les génomes et la régulation de leur activité est essentielle pour empêcher d'importants remaniements du génome. Les cellules hôtes peuvent réguler le taux de transcription de la transposase afin d'agir sur le niveau de quantité active de la protéine ou en agissant sur la protéine elle-même. La méthylation de l'ADN au niveau des cytosines est l'un des contrôles épigénétiques les plus répandus chez les eucaryotes (Bird, 2002). Chez les mammifères, la méthylation est une des réponses majeures à l'envahissement par tout ADN étranger (parasites génétiques viraux ou ET, mais aussi vecteurs de transgénèse non viraux) notamment en intervenant au niveau des promoteurs des ET (Yoder, Walsh and Bestor, 1997) ou en éteignant globalement l'activité de l'élément (Barau *et al.*, 2016). Il a été montré chez la souris que la méthyltransférase DNMT3C n'agissait qu'au cours de la spermatogenèse fœtale en reconnaissant et méthylant sélectivement les transposons les plus actifs (et donc potentiellement les plus dangereux) afin de maintenir la fertilité (Barau *et al.*, 2016).

On retrouve aussi chez les plantes le même phénomène avec la méthylation des TIR des éléments *Activator* et *Mutator* qui entraine une inhibition de la transposition (Chandler and Walbot, 1986; Chomet, Wessler and Dellaporta, 1987; Slotkin, Freeling and Lisch, 2003). On peut également observer une inhibition par la surproduction de la transposase. Ceci a notamment été décrit chez *D. melanogaster* avec l'élément non autonome *peach* excisé par la transposase de *Mos1* (Lohe and Hartl, 1996). Si *Mos1* est surexprimé alors, la fréquence d'excision de l'élément *peach* diminue. Cela est expliqué par le fait qu'une trop grande quantité de transposase dans la cellule entraine la production d'oligomères inactifs (ou moins actifs) qui vont saturer et diminuer l'efficacité du processus de transposition par l'inhibition compétitive entre des éléments actifs et inactifs pour l'accès aux TIR bloquant ainsi l'excision des copies actives.

Enfin, il existe une phase de sénescence durant laquelle on peut observer une forte diminution du nombre de copies jusqu'à l'élimination complète de la famille. Cela peut être dû à l'accumulation de mutations conduisant à la perte totale d'activité des ET et de leur impossibilité d'être mobilisé en trans (Lohe *et al.*, 1995; Lohe, De Aguiar and Hartl, 1997; Le Rouzic and Capy, 2005), soit à des événement de recombinaisons ectopiques entre plusieurs copies d'ET engendrant des remaniements chromosomiques (Biémont *et al.*, 1997; Kalendar *et al.*, 2011) ou soit par la perte stochastique de copies par dérive génétique (Capy, David and Hartl, 1992; Lohe *et al.*, 1995). Chez les mammifères, l'extinction des LINE-1 a été observée chez les lignées de rhinocéros, de rongeurs, de chauves-souris, d'insectivores et d'Afrothériens (Ivancevic *et al.*, 2016; Sookdeo, Hepp and Boissinot, 2018; Blumenstiel, 2019).

La présence d'ET dans le génome dépend donc de l'équilibre entre la transposition et le taux d'accumulation des mutations dégénératives. Si le taux de transposition est inférieur au taux de mutation qui rend un élément inactif, il est probable que cet ET va à terme disparaître du génome de l'hôte (Nuzhdin, 1999). Cependant, des copies de ET peuvent s'insérer dans des régions où elles induisent une augmentation de la "fitness" en faveur des individus qui les portent (insertion adaptative), ce qui induit leur fixation dans le génome.

4. Impact des ET

Les ET ont un rôle majeur dans l'évolution des génomes. Les recombinaisons entre les éléments situés sur un même chromosome peuvent générer des variants structuraux tels que des duplications, les délétions ou les inversions de séquences dans le génome de l'hôte. Les ET peuvent avoir une multitude d'effets lorsqu'ils s'insèrent dans de nouveaux locus. Ces effets varient en fonction de l'endroit où l'ET s'insère. L'insertion d'ET peut aussi être à l'origine de phénomène d'inactivation de gène ou encore de variation d'expression des gènes environnants. Lorsqu'un ET s'insère dans une séquence codante, cela peut entrainer l'inactivation du gène dans lequel l'ET s'est inséré. Au moins 65 insertions *de novo* d'un rétrotransposon sans LTR sont à l'origine de maladies humaines (Belancio, Hedges and Deininger, 2008). Ils peuvent aussi faire varier le niveau d'expression des gènes environnants en s'insérant dans des régions intergéniques. Cela peut être dû soit à la structure de l'ET soit aux conséquences de son inactivation par les systèmes de défense (régulation) de l'hôte. De nombreux ET possèdent des structures de régulation de l'expression comme des promoteurs et terminateurs de transcription ou encore des signaux de polyadénilation (Saedler *et al.*, 1974; Charlier, Piette and Glansdorff, 1982; Jaurin and Normark, 1983). Par exemple, des rétrotransposons sans LTR possèdent des signaux de polyadénilation qui, s'ils sont insérés entre un gène et son promoteur, peuvent empêcher la transcription des gènes en aval (Cordaux and Batzer, 2009).

Les ET ont participé à l'évolution des gènes en étant une source de motifs cis-régulateurs de l'expression des gènes de leurs hôtes. Il existe plusieurs cas d'insertion d'ET ayant eu un impact positif. Dans ce cas, la copie d'ET ou la partie induisant l'effet bénéfique sera maintenue dans le génome.

La transposition d'ET à proximité des gènes codants pour des protéines peut déclencher de nouveaux phénotypes adaptatifs. Les ET nouvellement insérés (ou excisés) peuvent affecter ces gènes en modifiant l'environnement règlementaire ou codant à un endroit donné. Ainsi, l'insertion dans des régions introniques promotrices ou régulatrices peut modifier les schémas

d'expression des gènes en aval. Un des exemples les plus marquants est celui de la phalène du bouleau avec le phénotype adaptatif au « mélanisme industriel » (**Figure 3**). L'insertion d'un ET dans un intron a entrainé l'augmentation de la transcription du gène *cortex* (Van't Hof *et al.*, 2016).

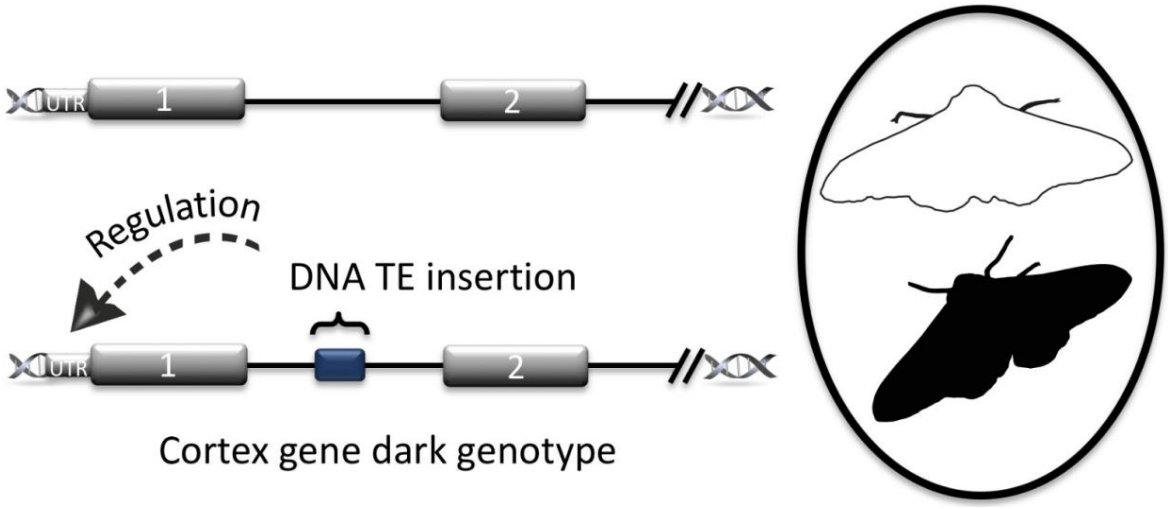


Figure 3 - Modification de l'expression d'un gène après l'insertion d'ET.
Cas du phalène du bouleau. [Source : (Schrader and Schmitz, 2018)].

Les insertions d'ET peuvent également conférer une résistance aux pesticides en affectant les profils d'expression des gènes métaboliques (Rostant, Wedell and Hosken, 2012; Guio, Barrón and González, 2014; Mateo, Ullastres and González, 2014; Le Goff and Hilliou, 2017). Ces insertions peuvent aussi entrainer des modifications au niveau épigénétique qui peuvent affecter l'expression des gènes voisins (Horvath and Slotte, 2017). Chez *E. coli*, il a été démontré que l'insertion d'un seul ET entraine la transition adaptative d'un mode de vie commensal à un mode de vie pathogène (Proença, Barral and Gordo, 2017).

Au cours de l'évolution, les ET peuvent être « détournés » et devenir des éléments constitutifs du génome. C'est ce qu'on appelle l'exaptation ou domestication moléculaire. Contrairement aux éléments transposables "sauvages" autonomes, les ET domestiqués ne se comportent plus comme des séquences mobiles. Alors que les ET classiques sont présents en plusieurs copies et fréquemment présents à différentes positions dans le génome d'espèces divergentes, les éléments domestiqués sont présents en une seule copie et détectables à des loci orthologues dans différents organismes. Certaines séquences nécessaires à la transposition, par exemple les LTR, sont généralement non fonctionnelles ou manquantes, ce qui empêche une transposition autonome. La distinction avec les ET simplement défectueux inactivés par des

mutations est faite par le fait que les séquences codantes domestiquées montrent la conservation d'un cadre de lecture ouvert fonctionnel dans différentes espèces. Fréquemment, un excès de substitutions nucléotidiques synonymes (sans effet sur la séquence protéique) par rapport aux substitutions non synonymes (appelé aussi *dN/dS*) indique que le gène domestiqué a évolué sous sélection purificatrice (négative) et ne correspond pas à un pseudogène. Comme une telle préservation du potentiel codant est indépendante de la transposition, on peut faire l'hypothèse que l'ancien gène de l'ET remplit une nouvelle fonction. (Britten, 1996; Kidwell and Lisch, 1997; Miller *et al.*, 1999; Kidwell, 2002; Volff, 2006; Sinzelle, Izsvák and Ivics, 2009; Joly-Lopez *et al.*, 2016). Cette partie sera développée plus en détail dans la partie III.

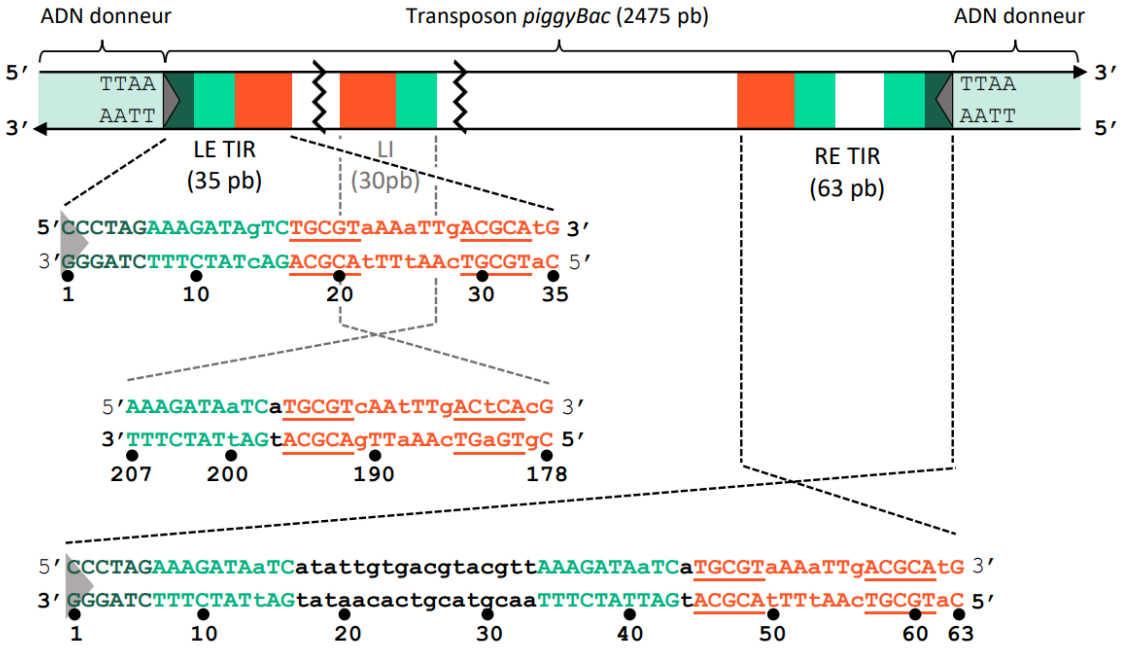


Figure 4 - Représentation schématique du transposon piggyBac

Le transposon *piggyBac* est flanqué de chaque côté par TTAA. Les extrémités gauche (LE TIR) et droite (RE TIR) de *piggyBac* sont constituées d'une répétition terminale inversée (vert) et d'une répétition interne inversée (orange). Extrémité gauche (LE), Interne gauche (LI) et extrémité droite (RE). Les rectangles oranges désignent la séquence ADN ciblée par PB et en vert clair, la séquence d'ADN qui interagit avec PB(1-558) d'après (Morellet *et al.*, 2018). Les séquences d'ADN sont numérotées à partir de la séquence cible TTAA qui flanque chaque extrémité du transposon. [Adapté de : (Morellet *et al.*, 2018; Chen *et al.*, 2020)].

II. *PiggyBac*

Parmi l'ensemble des familles de transposons, nous nous sommes plus particulièrement intéressés à la famille de transposon *piggyBac*.

1. Le transposon *piggyBac* (*Ifp2*)

Le transposon *piggyBac* (appelé aussi *Ifp2* pour *Insertion in Few Polyhedral mutants*) est un élément typique de classe II découvert par M. J. Fraser (Fraser, Smith and Summers, 1983) à la suite de l'observation de baculovirus mutant (forme polyédrique) au cours de la propagation de ces virus dans une lignée cellulaire de *Trichoplusia ni*, un lépidoptère appelé communément la fausse-arpenteuse du chou (Fraser, Smith and Summers, 1983; Hwei-gene Heidi Wang, Fraser and Cary, 1989). L'analyse de chacun des virus mutants a révélé qu'un fragment d'ADN de la cellule hôte s'était inséré dans chacun des génomes viraux, entrainant l'interruption du gène codant une protéine virale de 25 kDa, la polyhédrine (Fraser, Smith and Summers, 1983; Yusa, 2014). La taille des fragments ADN de l'hôte inséré dans ce gène viral variait entre 0,5 et 2,8 kb. L'un des fragments les plus fréquemment insérés avait une taille de 2,5 kb. Ce fragment était présent sous forme de copies multiples dans le génome de *Trichoplusia ni*. Ces observations confirment que les insertions observées dans le génome viral sont des éléments mobiles présents dans le génome de *T. ni*. Le séquençage du fragment d'ADN de 2,5 kb inséré a révélé que son site d'intégration présentait les signatures typiques de transposon d'ADN, telles que la duplication parfaite du site d'insertion de 4 pb (TTAA) et les répétitions terminales inversées (TIR) aux extrémités (Fraser *et al.*, 1985). C'est en 1989 que la séquence complète de l'ADN inséré a été déterminée, révélant pour la première fois la structure du transposon *Ifp2* (Cary *et al.*, 1989).

Ifp2 est un transposon de 2475 pb dont la mobilité se fait via un intermédiaire ADN. Il a un site d'insertion spécifique, le tétranucléotide TTAA (Fraser, Smith and Summers, 1983; Mitra, Fain-Thornton and Craig, 2008). À ses extrémités, l'élément présente (**Figure 4**) des TIR qui sont asymétriques. Le TIR de l'extrémité gauche (LE TIR, 5'TIR) et le TIR de l'extrémité droite (RE TIR ; 3'TIR) ont une taille respective de 35 et 63 pb. Le LE TIR et le RE TIR se composent d'une répétition terminale inversée de 13 pb (gris clair/bleu) et d'une répétition inversée interne de 19 pb (blanc/violet) appelée inversions subterminales inversées STIR). Ils sont séparés respectivement par un espace de 3 et 31bp (Elick, Lobo and Fraser, 1997; Morellet *et al.*, 2018; Chen *et al.*, 2020). Les STIR contiennent deux motifs palindromiques de 5 pb (TGCGT.....ACGCA).

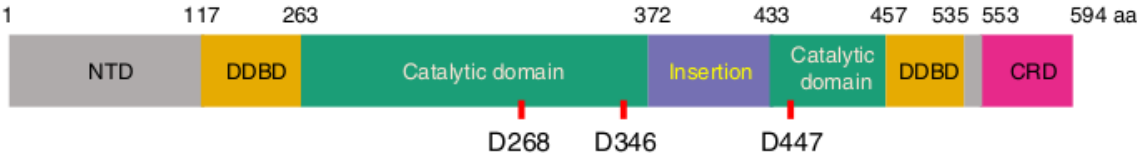


Figure 5 - Organisation des domaines de la transposase piggyBac.

Le domaine catalytique contient le motif conservé DDD (en rouge). **NTD** : Domaine N-terminal. **DDBD** : Domaine de dimérisation et de liaison à l'ADN. **CRD** : Domaine C-terminal riche en cystéine. [Source : (Chen *et al.*, 2020)].

Entre les STIR est présent un cadre de lecture ouvert (ORF) de 1,8 kb qui code une transposase de 594 acides aminés (**Figure 5**). Il a été montré que la transposase de *piggyBac* est aussi capable de mobiliser en *trans* pour la transposition (excision et intégration à un autre locus) des éléments *piggyBac* non autonomes (Elick, 1996). Ces éléments *piggyBac* non autonomes appelés *pble* (*piggyBac-like elements*) appartiennent à la superfamille piggyBac. Il s'intègre au niveau d'une séquence TTAA et l'excision précise des *pble* est délimitée par les deux TIR.

2. La Transposase *PiggyBac* (PB)

Comme beaucoup d'autres transposases et intégrases virales, la transposase de *piggyBac* possède des domaines très conservés. Elle est composée de plusieurs domaines : un domaine catalytique qui contient le motif conservé DDD (en rouge **Figure 5**), un domaine N-terminal (NTD), un domaine de dimérisation et de liaison à l'ADN (DDBD) et un domaine C-terminal riche en cystéine (CRD).

La transposase *piggyBac* fait partie de la superfamille des RnaseH-like ou intégrase virale. De nombreuses transposases et intégrases rétrovirales contiennent un domaine DD[E/D] qui comprend deux acides aspartiques (D) hautement conservés et un autre acide aspartique ou un acide glutamique (E). Le domaine DD[E/D] fixe des ions métalliques divalents tels que Mg2+, qui sont nécessaires à l'activité de transposition/intégration. La mutation de l'un de ces acides aminés supprime complètement l'activité de transposition. L'alignement des transposases de types *piggyBac* trouvées dans diverses espèces (allant des insectes aux primates en passant par les tuniciers) montre qu'elles contiennent plusieurs blocs très conservés d'acides aminés contenant plusieurs D et E dans la région centrale entre les positions 130 et 552 (Sarkar *et al.*, 2003; Keith, Fraser and Fraser, 2008; Mitra, Fain-Thornton and Craig, 2008).

Le domaine catalytique est constitué de trois résidus d'acide aspartique appelée aussi triade catalytique DDD : Aspartate D268, D346 et D447 qui assurent les réactions de clivage de

l'ADN de l'hôte et son intégration (Sarkar *et al.*, 2003; Keith, Fraser and Fraser, 2008; Yusa *et al.*, 2011). Il existe une autre aspartate en position D450 mais son rôle n'a pas encore été démontré. En effet, il peut être remplacé par un glutamate (E) sans modifier l'activité de la transposase (Keith, Fraser and Fraser, 2008).

Le domaine DDBD interagit avec les deux TIR du transposon, la plupart des interactions protéine/ADN sont en *trans*, de sorte que les interactions d'un monomère avec un TIR dirigent son extrémité dans le site actif de l'autre monomère. Ces transformations sont courantes dans les assemblages de transpososomes (Davies *et al.*, 2000; Richardson *et al.*, 2009; Montaño, Pigli and Rice, 2012; Liu, Yang and Schatz, 2019; Chen *et al.*, 2020).

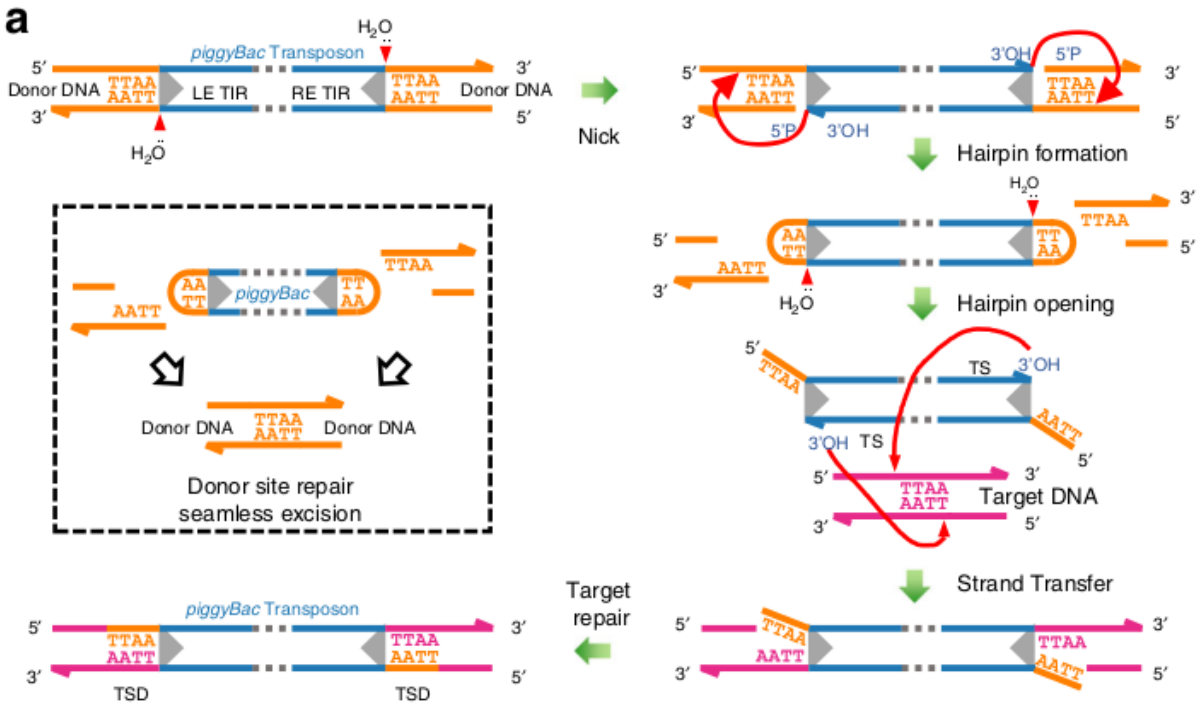


Figure 6 - Mécanisme de transposition de piggyBac.
L'hydrolyse par la transposase piggyBac libère le 3'— OH sur le brin d'ADN qui sera intégré. Elle est suivie d'une transestérification dans laquelle ce 3'— OH attaque l'ADN flanquant quatre nucléotides de l'extrémité du transposon formant une épingle à cheveux d'ADN. Pour l'intégration, PB ouvre l'épingle à cheveux à chaque extrémité lors d'une étape d'hydrolyse laissant l'extrémité de 4 nt, TTAA disponible. Ensuite, une seconde étape de transestérification va relier chaque extrémité à l'ADN cible. Sur le site donneur vide, les brins d'ADN complémentaires permettent une réparation propre. LE : extrémité gauche, RE : extrémité droite, TIR : répétition terminale inversée, TSD : Duplication du site cible [Source : (Chen *et al.*, 2020)].

Le domaine C-terminal riche en cystéine (CRD pour *Cysteine-Rich Domain*) forme un motif en doigt de zinc et participe à la dimérisation de la transposase et à sa liaison aux extrémités LE et RE du transposon (Morellet *et al.*, 2018). Sept cystéines sont présentes dans la transposase IFP2 avec un espacement bien conservé.

3. Mécanisme de transposition : excision et insertion

La transposition de *piggyBac* est un mécanisme de type « couper coller ». Elle est catalysée par une série d'étapes d'hydrolyses et de transestérifications qui génèrent un intermédiaire excisé avec une formation en *hairpin* (épingle à cheveux) aux extrémités (**Figure 6**). Comme le détaillent (Mitra, Fain-Thornton and Craig, 2008), la transposition de *piggyBac* se déroule en quatre étapes (**Figure 6**). Lors de la première étape, une coupure à la jonction du transposon 3' et de l'ADN donneur expose un 3' — OH sur l'extrémité du transposon. Étape 2 : le 3' — OH agit comme un nucléophile et attaque les 4 nucléotides dans l'ADN donneur adjacent sur le brin complémentaire, générant une épingle à cheveux sur l'extrémité du transposon et libérant le transposon du « backbone » du donneur. L'ADN flanquant du donneur contient les extensions 5' TTAA qui peuvent se rejoindre pour combler l'écart entre les donateurs et donner une excision précise. L'étape 3 : l'épingle à cheveux sur le transposon est reconnue par la transposase, laissant un excédent de 4 nt à l'extrémité 5' du transposon linéaire excisé et exposant de nouveau le 3'OH à l'extrémité 3'. Étape 4 : Le 3' — OH à l'extrémité du transposon se joint alors de manière covalente à l'ADN cible permettant l'intégration du transposon dans un nouveau locus.

Il a été montré que seules les 35 paires de bases en 5' et 63 paires de bases en 3' sont nécessaires comme domaine terminal pour permettre la transposition d'un vecteur *piggyBac* (Solodushko, Bitko and Fouty, 2014). Bien que les domaines terminaux complets soient importants pour une transposition efficace, ils pourraient être retirés sans perdre l'efficacité de l'intégration tant que ces domaines sont présents dans un environnement proche. En effet, une étude a mis en évidence la nécessité de la présence de séquences TIR courtes flanquant un gène d'intérêt pour l'intégration du transgène. La présence de séquences internes du transposon près de la transposase est nécessaire. Ces séquences internes ne doivent pas être présentes près du transgène (Solodushko, Bitko and Fouty, 2014).

4. Devenir du site chromosomique donneur

Au niveau des sites génomiques d'excision, des extrémités 5' sortantes du tétranucléotide TTAA sont produits aux deux extrémités. Celles-ci simplement s'apparient et l'ADN double brin est reconstitué sur les deux brins grâce aux enzymes de réparation de l'ADN. Ce processus permet de restaurer la séquence à son état initial, sans empreinte du passage de *piggyBac*. Toutefois, il arrive que la réparation du site du donneur échoue, ce qui entraine de petites insertions et/ou délétions, à une fréquence d'environ 1 % des événements d'excision lors de la transposition chromosomique dans les cellules souches embryonnaires (ES) de souris (Yusa *et al.*, 2011).

La transposition *piggyBac* présente donc deux caractéristiques uniques. Premièrement, elle ne nécessite pas de synthèse d'ADN. P*iggyBac* produit le tétranucléotide TTAA 5' attaché aux intermédiaires de transposition et l'utilise pour se coupler à l'extrémité 5' TTAA de l'ADN simple brin pendant l'intégration. Deuxièmement, les extrémités des transposons excisées forment une structure en épingle à cheveux. Ce mécanisme a été le premier à être observé dans des transposons eucaryotes, bien qu'il ait été observé dans les transposons bactériens Tn5 et Tn10 qui appartiennent à la famille IS4.

5. Sites d'intégrations

Le motif TTAA qu'utilise *Ifp2* est assez fréquent (18 713 270 sites) dans le génome humain, mais le statut de la chromatine et donc l'accessibilité de ces cibles peut affecter la préférence du transposon *Ifp2* pour s'insérer. Il est donc important de comprendre comment le transposon *Ifp2* reconnaît son site d'intégration. Ce point est particulièrement important lorsque le transposon *Ifp2* est utilisé comme mutagène ou comme vecteur de transgenèse. Li *et al* ont identifié plus de 30 000 sites potentiels d'intégration de *Ifp2* dans des cellules souches embryonnaires de souris (Li *et al.*, 2011). Lorsqu'il est mobilisé à partir de sites chromosomiques donneurs, le transposon *Ifp2* présente des sauts de « courte distance », ce qui est fréquemment observé chez les transposons ADN. Environ 10 à 15 % des intégrations ont été trouvées à moins de 5 Mb des sites donneurs (Yusa, 2014).

Concernant le statut épigénétique des sites d'intégration utilisés par *Ifp2*, ceux-ci sont clairement associés à de la chromatine ouverte (euchromatine). Cette observation est en fait triviale puisque qu'en toute logique, un transposon ne va pouvoir s'insérer dans l'ADN chromosomique que si la chromatine lui permet d'accéder à l'ADN. En toute logique, les sites utilisés sont des sites hypersensibles à la Dnase I, enrichis en histones H3 avec une lysine 4

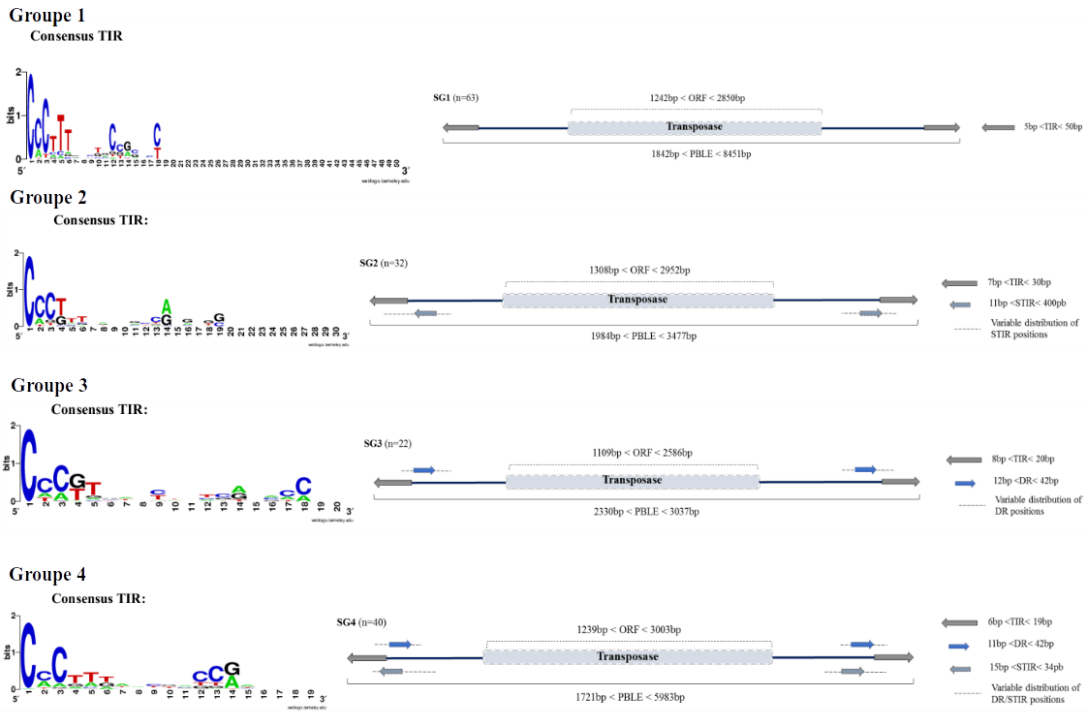


Figure 7 - Groupes de *pble* définis par Bouallègue et al, 2017 en fonction de l'organisation de leurs extrémités.

Le groupe 1 concerne les *pble* possédant des TIR, le groupe 2 est pour les *pble* ayant des TIR et des STIR, les *pble* du groupe 3 ont des TIR et des DR, le groupe 4 concerne les *pble* ayant des TIR, des DR et des STIR [Source : (Bouallègue *et al.*, 2017)].

triméthylée (H3K4me3). Comme ce type de profil de site est plus fréquemment rencontré au niveau des promoteurs, il n'est pas rare d'y trouver aussi un site de liaison de la polymérase pol II. En revanche, la présence de domaines associés à la lamine (LAD) observée dans l'hétérochromatine révèle des régions de chromatine condensée dans lesquels l'ADN chromosomique n'est pas accessible et donc des régions dans lesquelles l'intégration du transposon *Ifp2* n'a pas lieu. *Ifp2* s'intègre préférentiellement dans les régions génomiques contenant des gènes exprimés comme les régions d'initiation de la transcription (Wilson, Coates and George, 2007; Wang *et al.*, 2008; Yoshida *et al.*, 2017). Ces intégrations préférentielles dans la chromatine ouverte ont également été observées avec le transposon *piggyBat* (élément apparenté à *piggyBac* isolé à partir du génome de *Myotis lucifugus*) lorsqu'il est utilisé dans des tests de transposition effectués avec des cellules humaines et les fibroblastes de chauve-souris (Mitra *et al.*, 2013).

6. Diversité des éléments de la famille piggyBac

Pendant longtemps, l'élément *Ifp2* est resté le seul membre de l'actuelle superfamille *piggyBac*. Cependant en 2000, une première équipe montre la présence d'éléments ressemblant fortement à *Ifp2* (95 % d'identité nucléotidique) chez la mouche orientale des fruits, *Bactrocera dorsalis* (Handler and McCombs, 2000). La publication des génomes de multiples espèces animales dans les années suivantes a permis l'identification de nouveaux éléments de la famille *piggyBac*. La superfamille des transposons *piggyBac* est très ubiquiste dans les espèces eucaryotes. Sa distribution taxonomique qui dans un premier temps était restreinte aux différents ordres des insectes s'est vue diversifiée au sein des Eucaryotes, à l'exception des plantes. On retrouve des éléments de cette famille, qu'on appelle *pble* pour « *piggyBac* like elements », à la fois chez les insectes, les champignons, les crustacés et plusieurs espèces de vertébrés (Penton, Sullender and Crease, 2002; Sarkar *et al.*, 2003; Xu *et al.*, 2006; Wang *et al.*, 2006, 2010; Hikosaka *et al.*, 2007; Wu *et al.*, 2008; Sun *et al.*, 2008; Carpes *et al.*, 2009; Daimon *et al.*, 2010; Luo *et al.*, 2014; Wu and Wang, 2014).

Au moins cent-sept séquences différentes de *pble* possédant une transposase d'une taille supérieure à 300 acides aminés peuvent être recensées dans banques de données. Ces séquences de *pble* sont retrouvées dans les génomes allant des insectes aux chordés en passant par les hémichordés et les échinodermes (Bouallègue *et al.*, 2017). Quatre groupes de *pble* ont été mis en évidence sur la base de l'organisation de leurs séquences [(Bouallègue *et al.*, 2017); **Figure 7**], chaque groupe ayant un consensus propre de séquence de TIR. Habituellement chaque TIR

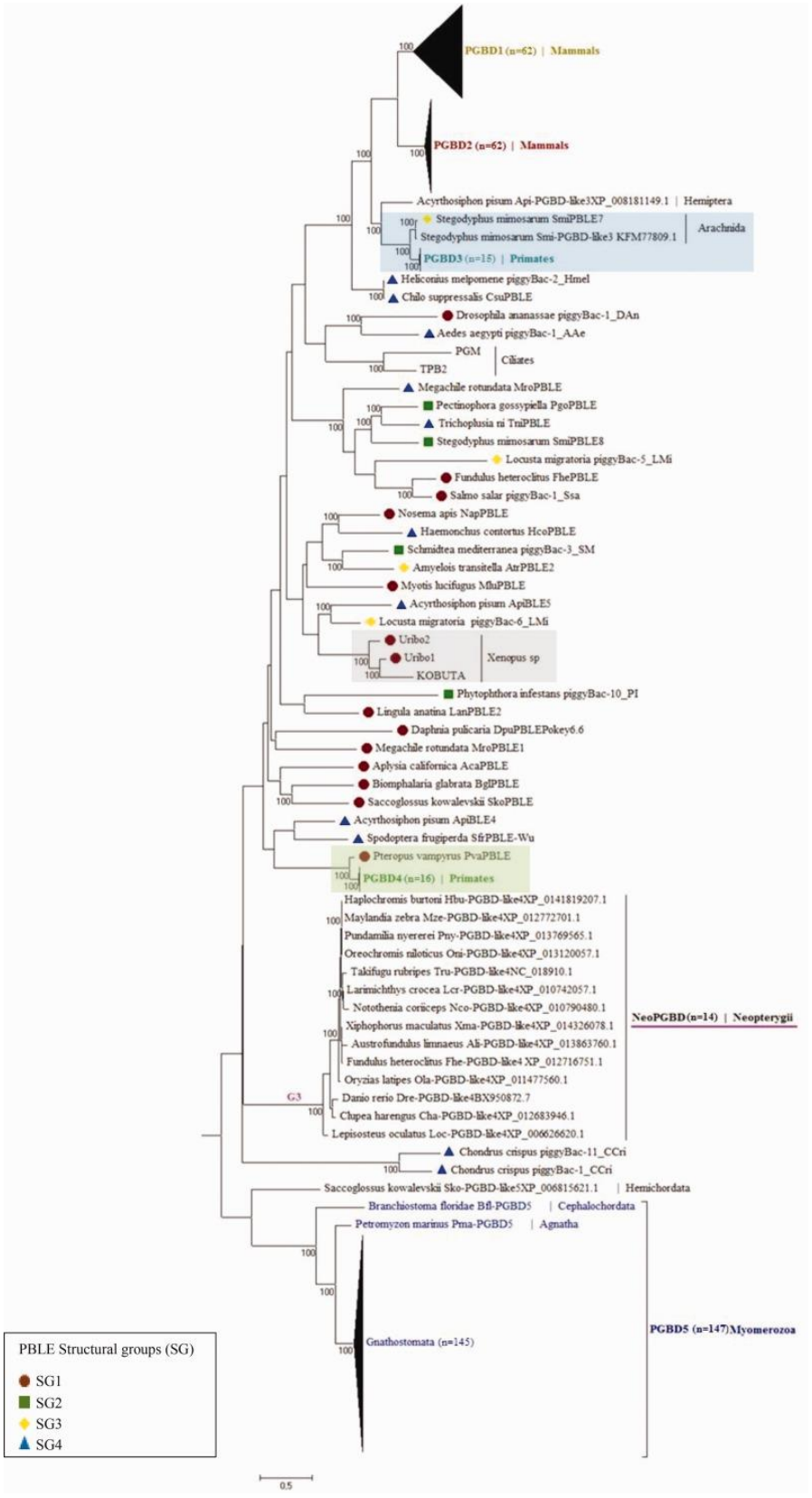


Figure 8 - Arbre phylogénétique de la superfamille piggyBac.

La phylogénie est basée sur l'alignement de séquences d'amino acides de 170 résidus [Source : (Bouallègue *et al.*, 2017)].

est flanqué par un site cible de duplication (ou TSD pour *target site duplication*). Le 1er groupe est constitué de 63 *pble* ayant seulement des TIR courts. Le 2nd groupe rassemble 32 *pble* qui ont des TIR et des inversions sub-terminales répétées (ou STIR pour *subterminal inverted repeats*). Le 3ème groupe est composé de 22 *pble* présentant des TIR et des répétitions directes (DR pour *direct repeats*) et le 4ème groupe est constitué de 40 *pble* qui présentent à la fois des TIR, des DR et des STIR. Comme le soulignent les auteurs (Bouallègue *et al.*, 2017), cette classification est basée sur l'organisation des extrémités des *pble* et non pas sur l'origine de leurs transposases. D'après l'apparentement phylogénétique des transposases de ces *pble* (**Figure 8**), ces éléments se sont différenties en au moins 5 lignées (Bouallègue *et al.*, 2017).

Chez l'homme, on peut recenser 4 *pble* : MER75, MER85, UCON29 et Looper. On suppose que les gènes *pgbd3* et *pgbd4* ont respectivement une origine commune avec MER85 et MER75 (A Smit, Riggs and Olson, 1996; Sarkar *et al.*, 2003; Sinzelle, Izsvák and Ivics, 2009; Bailey *et al.*, 2012). On peut supposer que *pgbd1* et *pgbd2* soient apparentés à UCON29 et Looper.

III. Domestication de gènes contenus dans des éléments transposables.

Les ET sont des séquences d'ADN qui peuvent se déplacer d'un endroit à un autre dans les génomes. Ils sont présents chez presque toutes les espèces et leurs mouvements ont façonné l'évolution de leur génome, générant de la variation génétique, des transferts horizontaux de gènes, le remodelage du génome et l'émergence de réseaux de régulation distincts [Bourques et al, 2018 ; Feschotte and Pritham, 2007]. Les ET peuvent aussi être à l'origine de nouveaux gènes dérivés des ORF qu'ils contiennent dans leurs séquences. Ces nouveaux gènes sont porteurs de nouvelles fonctions dans le génome hôte. Ce processus est appelé « domestication moléculaire » ou exaptation. Tandis que la séquence nucléotidique des copies originelles d'ET diverge et dégénère en accumulant des mutations au cours du temps, les éléments domestiqués sont conservés grâce aux effets de la sélection liés à leur nouvelle fonction. Ils sont alors présents en une seule copie et localisés dans des locus orthologues depuis leur apparition dans les lignées d'espèces apparentées.

Chez les eucaryotes, les gènes dérivés d'ORF d'éléments de type rétrotransposons à LTR (Kokošar and Kordiš, 2013) et les transposons à TIR (Sinzelle, Izsvák and Ivics, 2009) sont les mieux décrits et étudiés dans la littérature scientifique.

Les gènes dérivés de rétrotransposons à LTR proviennent d'ORF codant l'intégrase, la protéine gag ou la protéine d'enveloppe. Le séquençage du génome humain et murin a révélé que plus de 50 gènes codaient pour des protéines issues du gène *gag* des rétrotransposon à LTR Gypsy. De nombreux gènes eucaryotes anciens codant pour des protéines présentent une similitude de séquence avec les intégrases de type DDE de rétroéléments. De tels gènes ont été identifiés chez l'humain, les poissons et les vers (Lloréns and Marín, 2001). La différence entre les gènes domestiqués dérivés de l'intégrase et l'intégrase ancestrale des rétrotransposons est l'absence de motif catalytique DDE dans les gènes dérivés de l'intégrase (Lloréns and Marín, 2001; Ono *et al.*, 2001; Kapitonov and Jurka, 2004; Brandt *et al.*, 2005; Campillos *et al.*, 2006).

Il existe aussi des gènes dérivants de rétrovirus. Le criblage du génome humain pour les gènes rétroviraux codants a révélé que de nombreuses copies complètes de gènes d'enveloppe virale (*env*) dérivaient des gènes *env* des rétrovirus endogènes (ERV) (de Parseval and Heidmann, 2005). Dans le génome humain, 59 polyprotéines virales intactes ont été détectées dans des positions dispersées, et 29 d'entre elles provenaient de gènes *env* gammarétroviraux (Villesen *et al.*, 2004). Certains de ces gènes *env* jouent un rôle fonctionnel dans la formation du placenta, interface nutritionnelle et protectrice entre la mère et le fœtus en développement.

L’activation transcriptionnelle des ERV dans le placenta contribue à l’alternance de l’expression des gènes voisins (Prudhomme, Bonnaud and Mallet, 2005; Alzohairy *et al.*, 2013).

Les gènes dérivés de transposons à TIR proviennent principalement d’ORF codant des transposases. Ils ont principalement conduit à l’émergence de deux types de protéines : des facteurs de transcription (Sinzelle, Izsvák and Ivics, 2009) et des recombinases impliquées dans des phénomènes de réarrangements programmés des génomes. Au cours de l’évolution des eucaryotes et des procaryotes, des domestications de gènes dérivés d’éléments transposables ont eu lieu sur plusieurs grandes fonctions : système immunitaire, déterminisme du sexe, différentiation germen-soma, virulence.

1. DDE transposon as Public Goods

Author Proof

Chapter 14
DDE Transposon as Public Goods

Check for updates

Louis Tsakou, Célia Vicari, Laura Helou, Vivek Keshri, Sabyasachi Das, Yves Bigot, and Pierre Pontarotti

Abstract DDE Transposons have been recruited at least four times as site-specific recombination activating gene allowing programmed DNA elimination in eukaryotes. The described cases are RAG in jawed vertebrates, Kat 1 and Alpha 3 in the *Kluyveromyces lactis* yeast and Piggymac/TPB1 TPB2 and TPB6 in ciliates. The domesticated RAG is the most known case. It constitutes the enzymatic core of the Jawed vertebrates V(D)J recombination machinery. It directs random assembly and joining of gene segments during the development of B and T cells helping in the generation of the enormous gene diversity encoding antibodies or T cell receptors. It was shown in the case of RAG that the shift from DDE transposon to site-specific recombination activating gene is an evolutionary phenomenon that did not require dramatic changes. This explains why the co-option of DDE transposon as site-specific recombination activating gene can occur in a convergent manner. As numerous genes coding for DDE transposases are widespread through numerous members of the life tree, it is expected that several of them might correspond to domesticated transposons involved in programmed DNA elimination and maybe in the generation of receptor diversity. The domestication of DDE transposon could have been and still be of an extreme importance for organisms' evolution.

L. Tsakou · C. Vicari · V. Keshri · P. Pontarotti ( )
Aix Marseille Univ IRD, APHM, MEPHI, IHU Méditerranée Infection, Marseille, France
e-mail: Pierre.pontarotti@univ-amu.fr

L. Helou · Y. Bigot
UMR INRAE 0085, CNRS 7247, PRC, Centre INRAE Val de Loire, 37380 Nouzilly, France

S. Das
Emory Vaccine Center and Department of Pathology and Laboratory Medicine, Emory University, 1462 Clifton Road North-East, Atlanta, GA 30322, USA

P. Pontarotti
SNC5039 CNRS, Paris, France

© Springer Nature Switzerland AG 2020

P. Pontarotti (ed.), *Evolutionary Biology — A Transdisciplinary Approach*,
https://doi.org/10.1007/978-3-030-57246-4_14

UNCORRECTED PROOF

491382_1_En_14_Chapter ☑ TYPESET ☐ DISK ☐ LE ☑ CP Disp.:13/8/2020 Pages: 22 Layout: T1-Standard

Author Proof

14.1 Preamble

Eukaryotes and prokaryotes do not have the same DNA content in their genome over stages of their cell differentiation. Indeed genomes are modified during development or cell differentiation. This phenomenon is called developmentally regulated genome rearrangement (DRGR) (Zufall et al. 2005). DRGRs start by a DNA breaks either due to a specific endonuclease activity or specific chromatin structure that is related to a mechanisù inducing a DNA recombination event.

14.2 Developmentally Regulated Genome Rearrangement via a Specific Endonuclease

All the reported cases so far correspond to the domestication of selfish DNA that displayed at the minimum an endonuclease activity derived from a DDE transposase, a homing endonuclease, or a prophage recombinase.

The homothallic switching endonuclease (HO) is a domesticated homing endonuclease (Fig. 14.1). The HO endonucleases are mobile genetic elements that correspond to DNA fragments encoding these proteins over all their length, from their first 5' to the last 3' nucleotide. These DNA fragments use the homing endonuclease they encode to mediate their mobility (Keeling and Roger 1995; Koufopanou and Burt 2005). Homing endonucleases specifically recognize and cleave a specific oligonucleotide motif that is generally located within a very conserved motif of some house-keeping protein-coding sequence. After DNA cleavage, a DNA fragment coding an HO endonucleases specifically inserts in frame with the parasitized gene in the middle of its own recognition sequences (Fig. 14.1). In a diploid cell that is heterozygous for a homing endonuclease, the gene lacking the parasitic element becomes cleaved at the recognition site, and the broken chromosome is invaded by the parasitic DNA fragment by homologous recombination using the homing endonuclease-containing gene as a template. The mating type in budding yeasts is derived from the domestication of one of these parasitic DNA fragments. The mating types in budding yeasts are encoded by the mating-type (MAT) loci *MAT*a and *MAT*α. Some yeasts have the ability to change their mating type (mating-type switching) without going through mating or meiosis." In *Saccharomyces cerevisiae,* switching is initiated when the homothallic switching (HO) endonuclease induces a DNA double-strand break (DSB) in the *MAT* locus. Next, the replacement is completed through a gene conversion, in which transcriptionally silent copies of *MAT* genes, known as "hidden MAT left, *HML*α" and "hidden MAT right, *HMR*a," are copied into the expressed *MAT* locus.

Prophage excision involved in the specific rearrangement. (Feiner et al. 2015, Fig. 14.2)

During the differentiation process in procaryote, prophages and derived prophages are excised from the genomic region using a prophage recombinase in order to

UNCORRECTED PROOF

Author Proof

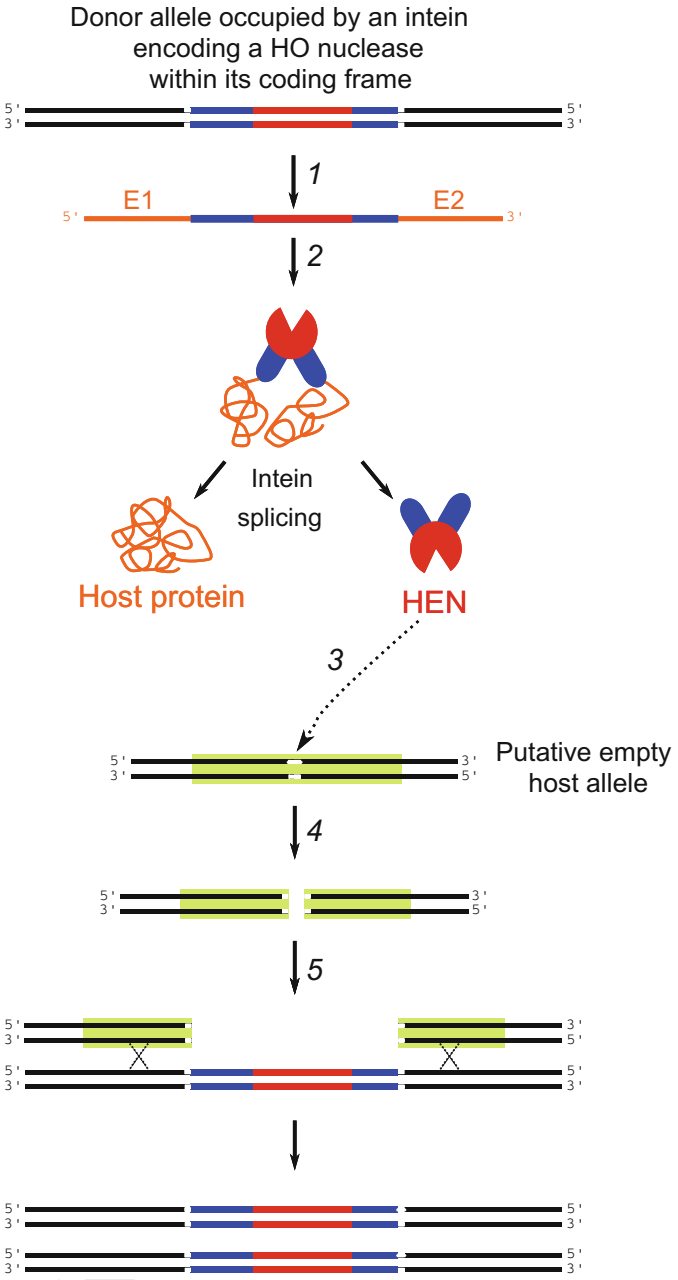


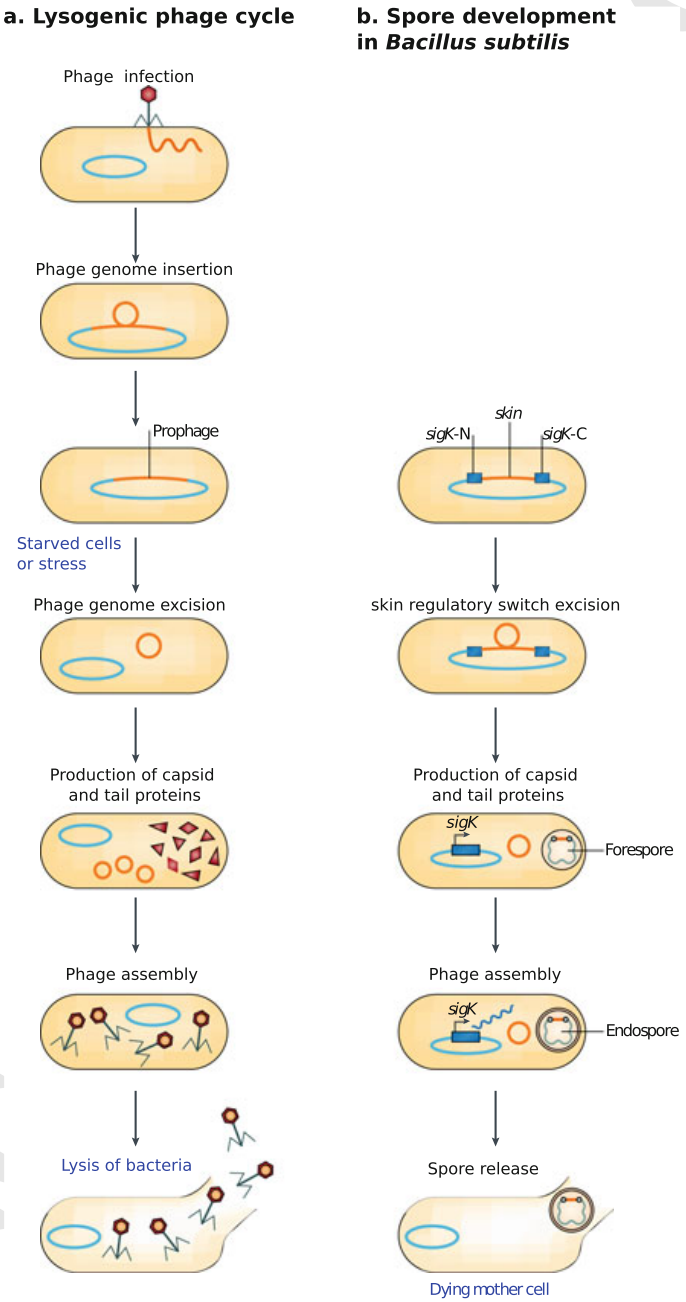
Fig. 14.1 Organization, expression, splicing, and mobility from one allele occupied by an intein to an unoccupied allele. The main steps of the intein mobility and homing are numbered. Step 1 is the RNA transcription. Black lines indicate DNA strands of the occupied and unoccupied alleles by an intein-coding DNA fragment, and cyan blue and light green bars indicate the coding-frame of the DNA region transcribed for each allele kind. Blue and red lines indicated regions coding for the intein (blue) and the homing endonuclease (HEN) moieties (red) in the DNA and the corresponding RNA transcripts. Step 2 is translation into protein. "pacmans" HEN moieties are shown in red and blue ellipses are N- and C-terminal intein moieties. Step 3 is site-specific recognition of an unoccupied allele by HEN. Blank spaces indicated the insertion site specifically cleaved by the HEN in both alleles. Step 4 is specific cleavage by the HEN. Finally, steps 5 and 6 depicted the invasion of a DNA fragment encoding intein from an occupied allele toward an unoccupied one by DNA strand invasion at DNA replication. This figure was constructed from derived information and graphic elements from Pièguet al. (2015)

491382_1_En_14_Chapter ☑ TYPESET ☐ DISK ☐ LE ☑ CP Disp.:13/8/2020 Pages: 22 Layout: T1-Standard

Author Proof

combine ORFs in frame. This process is involved in at least three phenomenons: sporulation (*Bacillus* and *Clostridium*), heterocyst differentiation (Cyanobacteria), and monocytogenes phagosomal escape (*Listeria*).

In the case of the "well analyzed" *Bacillus subtilis*, gene rearrangement occurs in the sigK gene that is disrupted into two segments by the insertion of SKIN (sigK intervening element) which is a remnant of the ancestral prophage (Abe et al. 2017a). During sporulation, SKIN is excised from the chromosome to combine the ORFs in



UNCORRECTED PROOF

Author Proof

◀**Fig. 14.2** Parallel between the lysogenic cycle of a phage (**a**) and the bacterial process involving a phage-derived mechanism activating sporulation (**b**). **a** Temperate phages enter a cycle in which their phage genome is first going to be integrated into the bacterial chromosome to become a prophage that is going to persist in what is considered a phage latent or dormant state in which bacterial cells will be viable and will not produce of phage particles. Prophages are replicated together with the bacterial host chromosome during host-cell replication and switch into lytic production upon exposure to stress. Chronologically, the entry in the lytic phase starts with the excision of the phage genome, its gene transcription then the production of phagie capsid and tail proteins that are thereafter assembled in phage particles. The last step is the lysis of the bacteria that releases phages and kills the bacteria. **b** Regulation of mother cell-specific genes during sporulation in *Bacillus subtilis*. A phage regulatory switch (phageRS), named *skin*, is inserted within the *sigK* gene. *SigK* encodes a protein, σK, that regulates the expression of late-stage sporulation genes in the mother cell. *skin* excises itself at the initiation of the sporulation process, leaving an intact sigK gene able to encode a functional σK protein. Post excision, σK expression activates the mother cell's late-stage sporulation genes. Following excision, the excised skin element is eventually lost in the mother cell, which dies late during sporulation. By contrast, the forespore, which did not undergo element excision, gives rise to an endospore that still encodes the skin element within its sigK gene. This figure was constructed from derived information and graphic elements from

frame. In addition to sigK, many other examples of sporulation-specific gene rearrangement occur, suggesting that this phenomenon is widespread and common in spore-forming bacteria (the intervening sequence can correspond to only a recombinase of prophage origin. The recombinase has two functions: it catalyzes the DNA cleavage at the recombination site and join the DNA molecule ends of the restored protein-coding genes (Abe et al. 2017b) (Fig. 14.2)

Heterocysts in cyanobacteria are specialized cells with a role in nitrogen fixation. They provide nitrogen to vegetative cells. Differentiation in a heterocyst is due the excision of fragments that interrupt three different loci: nitD, fdxN, and hupL that encodes a dinitrogenase alpha subunit, a ferredoxin and an uptake hydrogenase large subunit, respectively. The excised fragments have prophage origins and include at least the recombinase gene (Hilton et al. 2016).

Finally, to promote phagosomal escape, *Listeria* monocytogenes need to excise a temperate prophage integrated into the comK locus during bacterial phagocytosis in order to activate the comk expression (Feiner et al. 2015)

14.3 Programmed Break Specified by Chromatin Signals

14.3.1 Mating-Type Switching in the Fission Yeast S. pombe

As in the case of *S. cerevisiae,* the yeast *Schizosaccharomyces pombe* is able of mating-type switching. This mechanism occurred in an independent manner, the two species that belong to two highly divergent clades of yeast. The *S. pombe* genome contains one active (mat1) and two silenced (mat2 and mat3) mating-type loci that share similarities with the *S. cerevisiae* MAT proteins. However, the mechanism

UNCORRECTED PROOF

Author Proof

of switching is different from that involved in *S. cerevisiae*. Instead of cleavage mediated by an HO endonuclease (Fig. 14.2), a fragile chromosomal site consisting of an epigenetic mark at the mat1 locus in switching-competent cells leads to a dsDNA break during replication (Klar et al. 2014).

14.3.2 Immunoglobulin Switching

Immunoglobulin isotype class switching is a mechanism that changes a B cell's production of immunoglobulin from one type to another. It involves an intrachromosomal deletional rearrangement that focuses on the region of 1–10 kb of repetitive switch (S) located upstream of each heavy chain isotype gene. Class switch recombination is induced by a break done by cytidine deaminase AID (AID-APOBEC family) on accessible chromatin which is expressed during the B cell development (Yu and Lieber 2019).

The variable lymphocyte receptors (VLRs) system (Fig. 14.3). In agnathes (cyclostomes), the VLR system involved in adaptive immunity includes a germline VLR that do not code for functional protein but instead encodes the only portion of the amino and carboxyl termini of the mature VLRs. The sequences encoding those portions are separated by non-coding intervening regions. In lymphocytes, the germline VLRs are assembled by somatic DNA rearrangement into a mature VLR that encodes the functional receptor via the insertion of LRR cassette that flank the germline VLR. The germline VLR is broken by the AID-APOBEC enzyme at the intervening sequences between the C-terminal and N-terminal portions of the VLR genes where the chromatine is accessible. Then gene conversion starts thanks to sequence identity between the intervening sequence and sequences surrounding the LRR cassettes (Boehm et al. 2018).

14.3.3 Immunoglobulin Diversity Driven by Gene Conversion in Birds and Some Mammals

In birds and some mammals, only one pair of functional V and J segments is found for both the Ig light and heavy chain loci. Therefore, the diversity generated by V(D)J recombination is limited. However, several pseudo-V coding segments are found upstream the functional V segment in genes coding light and heavy chains. These pseudo-V segments are used as a template for gene conversion to diversify the single functional V segment. To launch the diversification process, a break done by the Activation-induced deaminase (AID) within the single functional V segment is required. (see for example Arakawa et al. 2002).

Other DRGRs have been described in the literature (Wang and Davis 2014), and they likely correspond to an infinitesimal number of cases that occur in nature. At the

UNCORRECTED PROOF

Author Proof

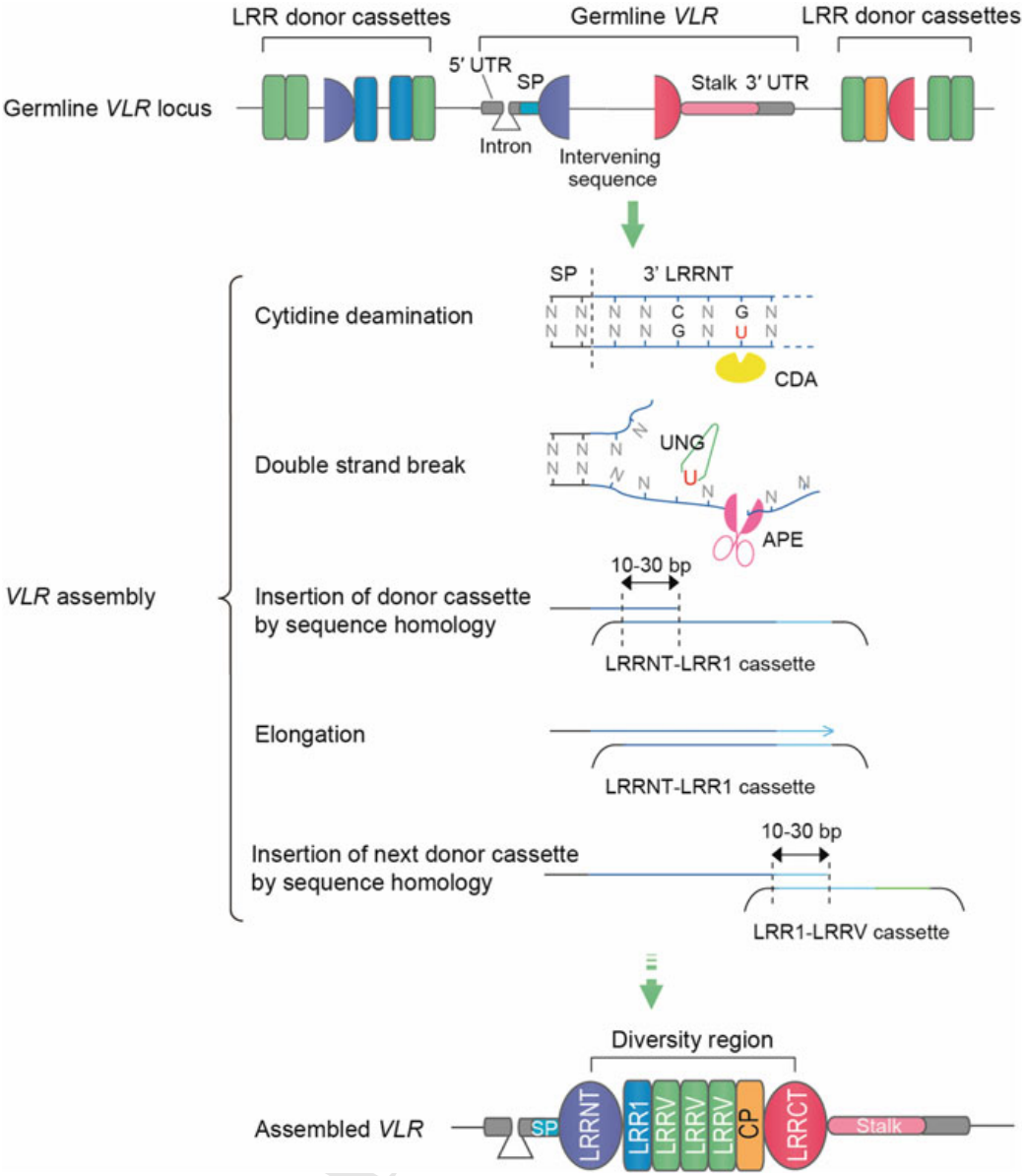


Fig. 14.3 Variable lymphocyte receptors (VLRs) system. Schematic diagram of a germline *VLR* gene and the postulated gene conversion-like (copy choice) mechanism for *VLR* assembly. The germline *VLR* gene is incomplete and contains invariant regions encoding for 5'-end of the N-terminal LRR (5′LRRNT) and 3′-end of the C-terminal LRR (3′LRRCT) and stalk. Hundreds of different LRR cassettes are located upstream and downstream of a pre-assembled germline *VLR* gene. The non-coding intervening sequence between 5′LRRNT and 3′LRRCT is replaced by the donor LRR cassettes that are sequentially copied either from 5′ to 3′ or 3′ to 5′ direction. At the beginning of the gene assembly process, a cytosine deaminase (CDA) converts cytosine (C) to uracil (U) in the germline *VLR* gene. The uracil is then removed by uracil-DNA glycosylase (UNG), leaving an apurinic (AP) site. The AP site activates nicking activity of AP endonuclease (APE) which leads to a DNA double-strand break. To repair this break, homologous recombination starts based on the sequence homology of 10-30 bp between the donor and acceptor LRR cassettes. This process is repeated along with deletion of the intervening sequence until the completion of a mature *VLR* gene. The diversity region of a mature *VLR* is composed of a 3'LRRNT, LRR1, multiple LRRVs, connecting peptide (CP), and a 5'LRRCT

491382_1_En_14_Chapter ☑ TYPESET ☐ DISK ☐ LE ☑ CP Disp.:13/8/2020 Pages: 22 Layout: T1-Standard

Author Proof

molecular level, as will be seen in the next section beside the prophage and among other domesticated selfish elements, domesticated DDE transposons seem a solution often retained by evolution to achieve programmed DNA elimination.

14.4 DDE Transposon and Domesticated DDE Transposon

DNA transposition is the process by which a discrete segment of DNA is either moved or copied into a new genomic location. Several distinct types of enzymes catalyze DNA transposition, one of the most abundant kinds are the DDE transposases thus named for conserved essential acidic residues located at the active catalytic site. The DDE transposase coding gene in the transposon is flanked by two Terminal Inverted Repeats (TIRs). To achieve transposition, the transposase recognizes these TIRs to perform the excision of the transposon which is after this or in a concerted manner with excision is re-inserted into a new genomic location. Upon insertion, the target site DNA is duplicated, resulting in Target Site Duplications (TSDs).

Several DDE transposons have been domesticated by their hosts retaining their DNA binding capacity (see for review (Sinzelle et al. 2009; Jangam et al. 2017)) and in some cases, the transposase and their related TIR as site-specific recombination activating gene. The few well-studied examples will be presented in the next sections.

14.4.1 Shift from a DDE Transposon to a Mechanism of Programmed DNA Elimination

DDE transposons have been recruited as site-specific recombination activating gene at least four times as programmed DNA elimination system, mediating biological differentiation processes: RAG (RAG1-RAG2) in jawed vertebrates, *Kat 1* and *Alpha 3* in *K. lactis* yeast and *PiggyBac* (TPB1 TPB2 and TPB6) in the ciliate *Tetrahymena*. The first described and best known example is the V(D)J RAG.

14.4.2 The Recombination Activating Gene (RAG) Paradigm

Jawed vertebrates has a specific adaptive immune system based on lymphocytes that express highly diverse, clonally distributed antigen receptors encoded by genes that are non-functional in the germline and assembled by recombination during lymphocyte development (Teng and Schatz 2015). This assembly reaction, known as V(D)J recombination, operates on arrays of V, D, J polypeptide-coding segments of immunoglobulin and T-cell receptor loci. V(D)J recombination is initiated early in lymphocyte development by a site-specific endonuclease: RAG1 and 2. The RAGs

UNCORRECTED PROOF

Author Proof

cleave at a conserved recombination signal (RSS) that flanks each V D and J segments. The finding that the RAGs have transposase activity, supports a model in which co-option of the components of a transposon played a critical role in the evolution of the jawed vertebrate adaptive immune system (see for review Flajnik 2016) (Fig. 14.4).

The hypothesis was that the RAG proteins derive from transposase genes of RAG transposons while the split antigen receptor genes derive from the insertion of the terminal inverted repeats (TIRs) of this transposon into a Ig-like receptor gene exons with the inserted TIRs becoming the RSSs. The presence of a RAG transposase core and TIRs in non-vertebrates (Kapitonov and Jurka 2005) was consistent with this model. The discovery of a complete RAG transposon in amphioxus (*Branchiostoma belcheri*) (Huang et al. 2016) strengthens this hypothesis. The discovery of an active

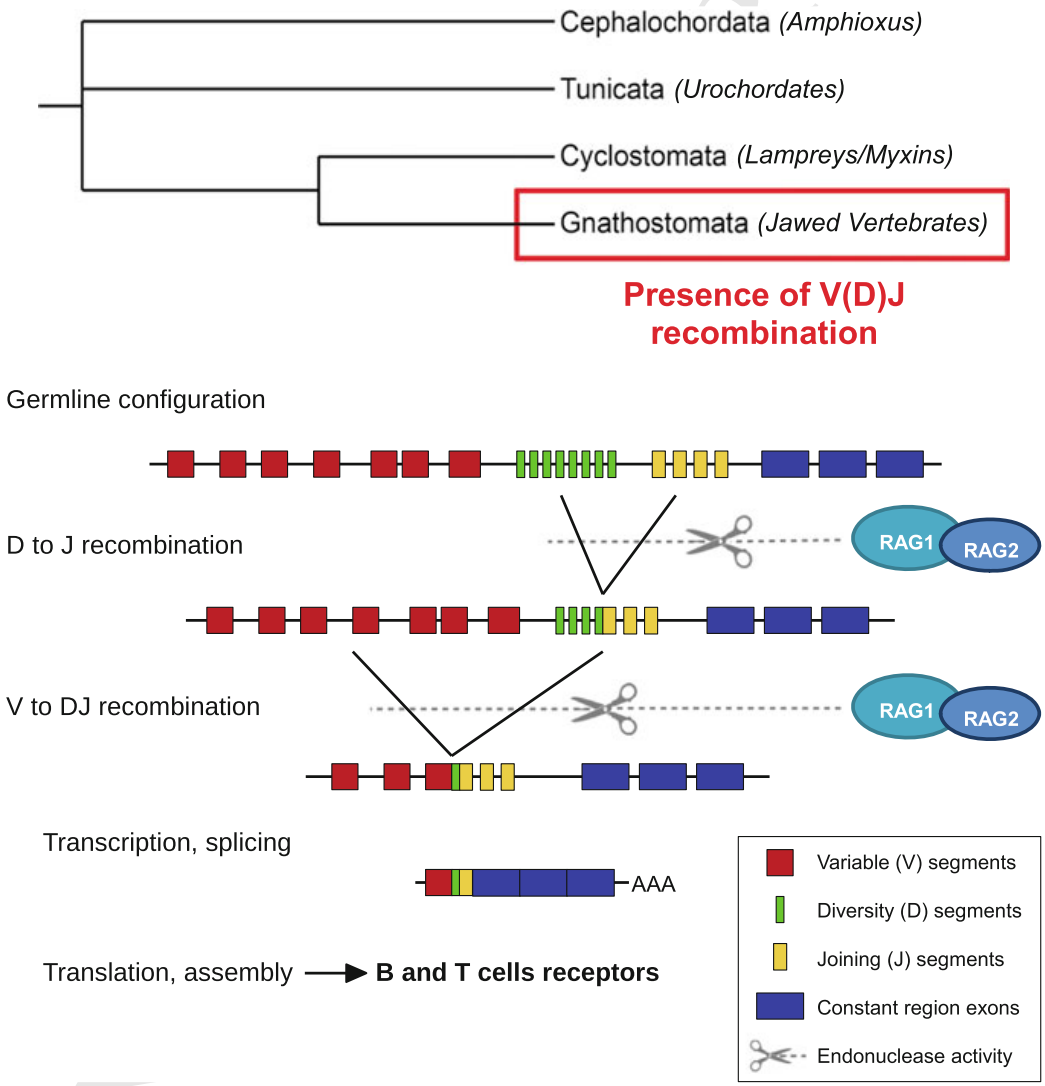


Fig. 14.4 RAG implication in the V(D)J recombination mechanism in jawed vertebrates. The chordates consensus tree shows the position of jawed vertebrates among chordates. The V(D)J mechanism was adapted from Janeway et al., 2001

Author Proof

RAG transposon in the hemichordata *Ptychodera flava* and several fossilized transposons in several Deuterostomia (Morales Poole et al. 2017) indicate that the RAG transposon was present in the deuterostomia ancestor and remained active to date in several lineages of this clade, while it was co-opted as part of V(D)J recombinase in jawed vertebrates about five hundred millions years ago. The RAG transposon includes TIRs that are related in sequence to the RSS heptamer and the transcribed open reading frame encodes RAG1-like and RAG-2 like proteins with a biochemical activity similar to those of the RAGs, including DNA cleavage via a nick-hairpin mechanism (Huang et al. 2016) (Fig. 14.5).

After DNA cleavage by the jawed vertebrate RAG or *Branchiostoma belcheri* RAG transposon, there are two predominant fates for the excised fragment flanked by RSSs or TIRs: joining of the ends to form a signal joint or integration into a new locus. The jawed vertebrate recombinase RAG strongly favors the joining of the ends. Indeed, RAG in the jawed vertebrates actively directs cleaved signals and coding ends into the NHEJ DNA repair pathway for joining coding segments in frame. In contrast, *Branchiostoma belcheri* RAG transposon appears to strongly favor integration after excision, nevertheless, it allows some TIR-TIR joints to form (Huang et al. 2016; Zhang et al. 2019). It is possible that the transposase partially prevents the interaction between the TIR and the NHEJ repair pathway and that the jawed vertebrate RAG lost this property. Zhang et al. (2019) started to uncover the mechanism beyond the domestication and evidenced important amino acid positions involved in transposition/or suppressing transposition (Fig. 14.7).

The mechanism of co-option can be described as follows: (1) insertion of an active transposon into a given genomic locus, (2) the translocation of the transposase gene; in some cases, the native TIRs remain within the genomic locus. (3) The TIR like sequence is recognized by the domesticated transposase (Known under the name of RSSs in the case of RAG) and the TIRs together with the internal sequence are excised (4) the excised sequence lost the ability to insert another genomic region and the two flanking ends are joined by the non-homologous end-joining (NHEJ) DNA repair pathway, to form a coding joint (CJ) (Fig. 14.6).

In conclusion, biochemical functions of the DDE transposon and the sequence-specific recombination activating system are very similar, hence the biochemical shift from a transposon to a sequence-specific recombination activating system corresponds an easy evolutionary step requiring a very limited number of events to pass form a "wild" state to a domesticated one (Fig. 14.3). This is also supported by the fact that many other DDE transposons have been co-opted as recombination activating site-specific endonuclease as described in the next paragraphs.

Piggymac/TPB2/TPB1/TPB6 in ciliates (Baudry et al. 2009; Cheng et al. 2016; Nowacki et al 2009) (Fig. 14.7).

Ciliates are unicellular organisms able to perform DNA rearrangements during development in order to differentiate a somatic macronucleus that is metabolically active from their transcriptionally silent germinal micronucleus. After duplication of the germinal genome, a large proportion of their germinal micronuclear genome is eliminated to differentiate a somatic macronuclear genome through the loss and in particular through the excisions of thousands of internal eliminated sequences. This

UNCORRECTED PROOF

Author Proof

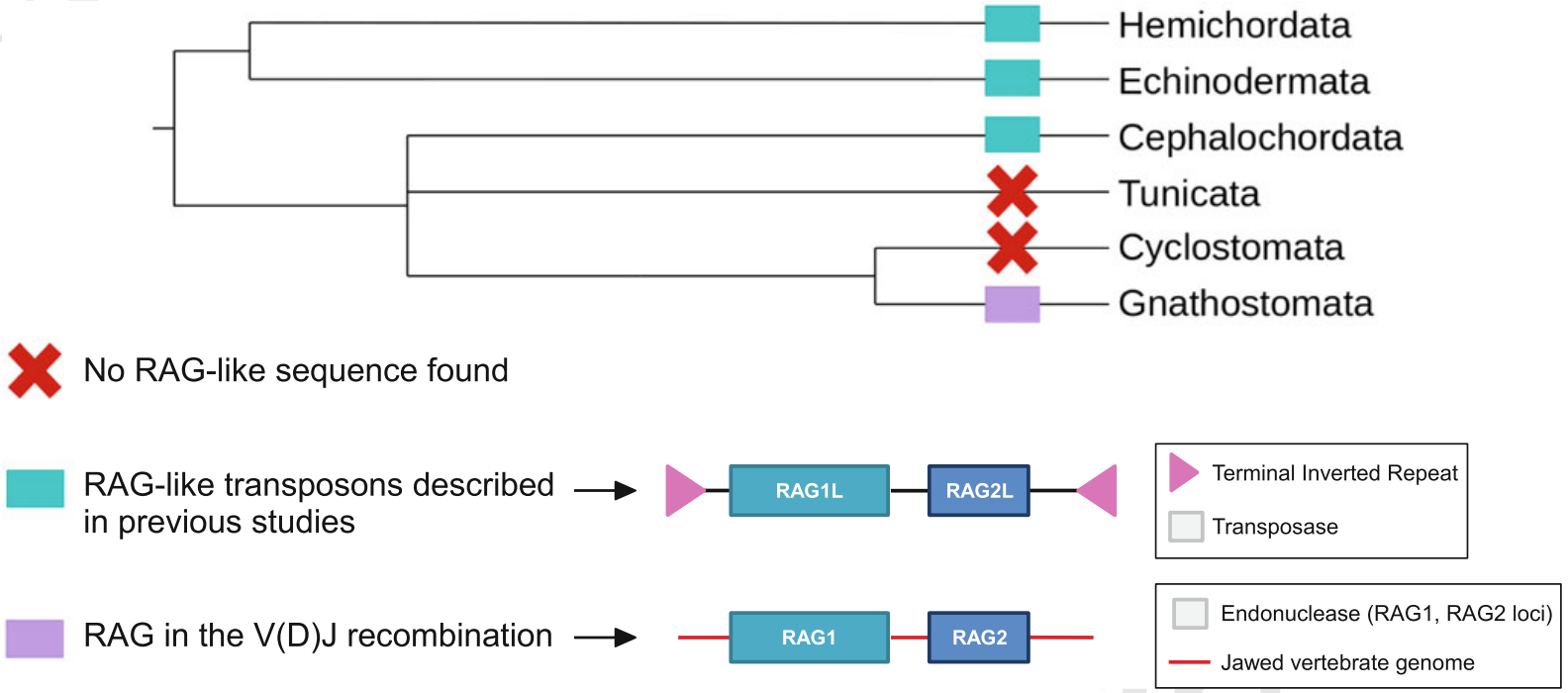


Fig. 14.5 Taxonomic distribution of the RAG and RAG-like sequences known in Deuterostomes. This is a deuterostomes consensus tree with the clades presenting RAG in the V(D)J recombination in the adaptive immune system, the clades presenting RAG-like sequences and the clades presenting no RAG-like sequences until today. described in Morales Poole et al. (2017)

491382_1_En_14_Chapter ☑ TYPESET ☐ DISK ☐ LE ☑ CP Disp.:13/8/2020 Pages: 22 Layout: T1-Standard

Author Proof

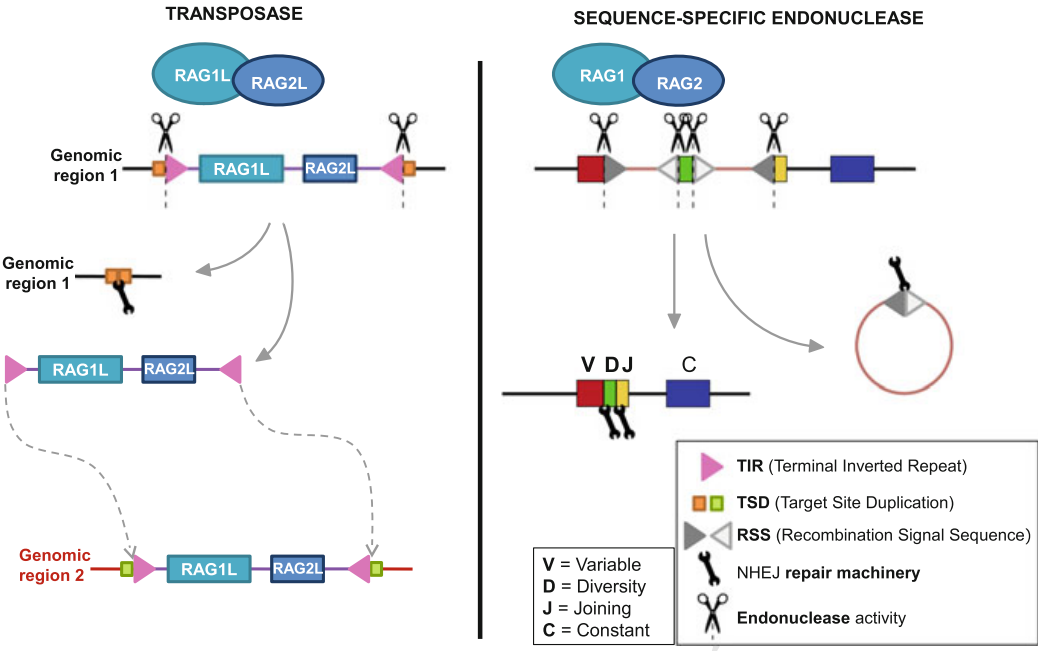


Fig. 14.6 Simple biochemical switch between the RAG-like transposase activity and the RAG endonuclease/recombinase activity in the V(D)J recombination of jawed vertebrates. Adapted from Huang et al. (2016)

is done by domesticated DDE transposon. Most of the eliminated sequences in the paramecium and tetrahymena are named, respectively, by Piggymac and TPB2 and are derived from transposons belonging to the piggybac family (Fig. 14.7). In contrast to the domesticated *piggyBac* transposase genes in the macronuclear genomes of oligohymenophorea ciliates, *Oxytricha* bears thousands of active transposase genes within the *Tc1/mariner* superfamily.

In the case of the Tetrahymena, the domesticated transposons TPB1 and TPB6 excise 12 internal eliminated sequences that disrupt exons. TPB1 and TPB6 recognize the TIR like sequences. TPB2 and likely a Piggymac in paramecium seem to recognize RNA intermediates. In the case of TPB2, the TIR direct interaction with the TIR has been lost and the domesticated transposase performs its excision via scnRNA-directed heterochromatin involved specialization; thus, TPB2 to recognize heterochromatin rather than TIRs. Thus Tetrahymena TPB2 seemingly represents another level of transposase domestication. The situation is less clear in the case of Piggymac where the domesticated transposase seems to recognize both IES ends and the intermediate RNA. It is likely that in the case of TPB2 and Piggymac, the transposition mechanism was gradually grafted into a heterochromatin formation pathway. In all these cases, the internal eliminated sequence form a circle that prevents reintegration in the genome as this is also the case for jawed vertebrates with the chromosomal excision product resulting from the RAG activity.

The phylogenetic analysis (not shown) indicated that possibly a PiggyBac transposon entered the Oligohymenophorea (that include the paramecium and tetrahymena phyla) ancestor and became later domesticated. The ancestor

UNCORRECTED

Author Proof

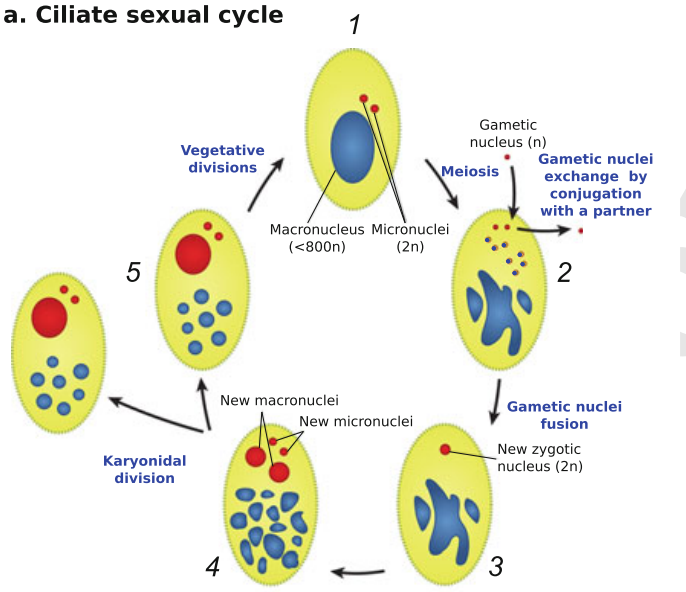


Fig. 14.7 Brief overview of biological and genomic features of the ciliate system. **a** Main steps of the sexual cycle in ciliate species belonging to the *Paramecium* genus. At the entry of the cycle in vegetative cells, the macronucleus starts disorganizing while the 2 micronuclei (2n) trigger meiosis in vegetative cells (*1*) resulting in eight gametic nuclei (n). Seven or six of these gametic micronuclei, respectively, degenerate after the acquisition of one gametic micronuclei or the exchange of one of them with another *Paramecium* cell. The fusion of both remaining gametic micronuclei (*3*) leads to a new zygotic nucleus (2n) that thereafter differentiate in new macronuclei and microneclei (*4*), before karyonical division and differentiation of a new vegetative cell. **b** Organization of genome sequences in the micronucleus and the macronucleus. Here, the micronucleus genome is reduced to one chromosome containing the genes (blue arrows) that are for some of them split by insertion element sequence (IES; orange rectangle). DNA transposons (orange double arrows) can also be interspersed in chromosomes. During macronucleus differentiation, IES are precisely excised what restores split gene ORFs (the faint blue line means that the 2 ORF are fusioned in one ORF), transposons are eliminated what fragments chromosomes before the remaining chromosomal fragments are amplified. This figure was constructed from derived information and graphic elements in Nowacki et al. 2011

UNCORRECTED PROOF

Author Proof

Piggymac/TPB2/TPB1/TPB6 domesticated transposon evolved in the two lineages: tetrahymena and paramecium Piggymac and TPB2/TPB1/TPB6. In the case of tetrahymena, a duplication gave rise to TPB2 in one part and TPB1/TPB6 ancestor in the other. TPB2 lost the possibility to recognize its TIR but gained the possibility to interact with the scan RNA. TPB1/TPB6 represent maybe the ancestral state and TPB2 could represent another state in the domestication process.

Kat 1 and MAT alpha3 (Barsoum et al. 2010; Rajaei et al. 2014)

As we said above, some yeasts have the ability to change their mating type. In *Saccharomyces cerevisiae,* switching is initiated by homothallic switching (HO) endonuclease. In the related yeast *Kluyveromyces lactis* HO has been replaced by two domesticated DDE transposons: MATalpha 3 involved in the switch d *MAT*α to *MATa.Kat1* involved in the switch *de MATa to MAT*α.

*MAT*α*3* which was domesticated from transposable elements belonging to the Mutator Like Element Family MULEs. Regulated excision of this element results in a double stranded DNA break, (DSB) that stimulates recombination from the genome and initiates mating-type switching from *MAT*α to *MATa.* It has to be noted that in the case of this domesticated transposon the TIR-like sequences have been replaced by other sequences which are different on the left and right sides. The left side is a low complexity sequence with a long stretch of T and A. The right side contains a conserved motif that is conserved in sibling species-species of the *Kluyveromyces* genus. The other domesticated transposon named Kat1 evolved from hAT (*hobo/Activator/Tam3*) transposases and is involved in the switching from *MAT*a to *MAT*α. Kat1 cleaves the *MAT*a locus at two different positions, resulting in DSBs that stimulate recombination. Kat1 recognizes a TIR like sequence. In both cases, the intervening DNA is joined into a circle.

In conclusion, DDE transposons evolved as site-specific recombination activating genes many times during evolution and therefore this is a case of isoconvergent evolution of site-specific recombination activating genes.

14.4.3 Simple Evolutionary Shift Can Explain Convergent Evolution

Isoconvergent evolution, the independent evolution of similar features from the same ancestral state in different evolutionary lineages (Pontarotti and Hue 2016), could be explained in part by natural selection where the new feature gave an advantage to the individual. Isoconvergent evolution could be due to the limited number of evolutionary pathways resulting from developmental and functional constraints on the evolutionary process (Losos 2011)—Functional constraints imposing a finite number of accurate adaptations, a finite number of mechanisms can be used to answer functional problems. Finally, isoconvergent evolution can be explained in part by the ease of transition from ancestral state to a derived state. This last aspect is not really discussed in the literature.

UNCORRECTED PROOF

Author Proof

Losos (2011) pointed out that the wings powering flight in vertebrates have been built in different ways in birds, pterosaurs and bats. In all these cases, the wings represent modified forelimbs. The combination of wings and forelimbs, in theory, would be not very useful in real life. He concluded that this was due to lack of constraints; however, we underline here that besides the constraints, it was easier to modify forelimbs to get wings than to start from nothing. The same reasoning could be applied for the DDE transposon co-option as sequence-specific recombination activating systems and the biochemical shift from a transposase to a sequence-specific recombination activating endonuclease is an easy evolutionary step.

14.4.4 Transposases Form the Largest Family in the Diverse Genomes of Life

Transposase-encoding genes are greatly over-represented in sequenced genomes and metagenomes relative to other coding sequences (Aziz et al. 2010). Some of the transposase coding genes could correspond to domesticated transposons, active transposons and maybe to fossilized ones. From this and the chapter developed above, we hypothesize that many DDE transposons have been recruited as recombination activating site-specific endonuclease systems. Many domesticated candidates have already been described in the literature, some of them lost the endonuclease domain butand conserved the DNA binding domain and they could be involved in novel chromatin-modifying complexes (Feschotte 2008) while others are involved in centromere binding, chromosome segregation, meiotic recombination (Sinzelle et al. 2009). Some articles also described large-scale systematic analysis to search for domesticated transposons including DDE transposons (Hoen and Bureau 2015; Bouallègue et al. 2017). In the human genome 26 putative DDE transposases has been described (Arnaoty et al. 2012), among these putative domesticated transposases as far as we know other than the RAGs, only the function of one of them, PBGD5 was really investigated, and its nuclease activity has been shown, however we do not know whether it recognizes specific sequences and if it is really involved in recombination (Henssen et al. 2015, 2016, 2017). Another candidate has been tested for its transposase activity (Majumdar et al. 2013) where the authors show that THAP9 gene encodes an active DDE DNA transposase.

We need to test further the hypothesis that many transposons have been recruited as recombination activating site-specific endonuclease. If our hypothesis is true, many domesticated DDE transposons acting as specific DNA endonucleases should be found throughout the life diversity.

UNCORRECTED PROOF

Author Proof

14.4.5 Search for Domesticated Transposons Involved in Programmed Recombination. Test of the Hypothesis

This paragraph is a small guideline. The strategy to search for domesticated transposons involved in programmed recombination can be performed by two complementary approaches: Look for the candidate domesticated transposase and look for a somatically rearranged genomic region.

14.4.6 Look for the Candidate Transposase Co-opted as Site-Specific Endonuclease Involved in Genomic Recombination

Different strategies have been developed in particular by Hoen and Bureau (2015) and Bouallègue et al. (2017). The strategy described here was adapted from these publications. We first have to look for the sequence that codes for complete transposase; these can be done by profile search using an alignment with a known transposase (Eddy 2011).

The ORF found needs to be checked for a bona fide catalytic site and the conservation of the DDE motif. Then in order to be sure that the transposase is likely to be active it is necessary to show that the protein evolved under constraint. This can be done by calculating the ratio of non-synonymous to synonymous substitutions (dN/dS) which is a useful measure of the strength and mode of natural selection acting on protein-coding genes (Jeffares et al. 2015).

The next step is to differentiate between transposase belonging to an active transposon from site-specific recombination activating gene (domesticated transposase). The following criteria must be present in the case of site-specific endonuclease. The domesticated transposase must be in single copy without pseudogene like sequences that could correspond to recent transposition events. The next criteria is based on the fact that the domesticated transposase should be unable to transpose thus the domesticated transposase should remain in the same genomic region in different species—conserved synteny—(Rascol et al. 2009). The higher the number of divergent species display conserved synteny, the higher is the probability that the transposase has been domesticated. For example, PGBD5 is present in all chordate genomes and belongs to a conserved synteny (Pavelitz et al. 2013) indicating that it is likely to have been domesticated.

The following step is to search the domesticated sequence recognition signal: find RSS/TIR like sequence. One way to search for the sequence recognition signal is to perform Chip-Seq experiment (Park 2009), and the other, that could be complementary to the Chip-Seq experiment is to be guided by sequence data. This could be done by looking for the active transposon which is the most similar to the domesticated transposon and to determine the TIR sequence of the transposon. The TIR sequence should be similar to the site-specific DNA sequence recognized by the domesticated

UNCORRECTED PROOF

Author Proof

transposase, as this is the case for the RAG transposon and the domesticated RAG (see Kapitonov and Jurka 2005; Huang et al. 2016; Morales Poole et al. 2017). The TIR like sequences can then be searched in the genome(s) coming from the cells where the "domesticated" transposase is transcribed. It should be noted that the endonuclease could recognize RNA intermediates as this is the case for TPB2 in-ciliates where the TIR recognition has been lost and instead a chromatin structure is recognized.

14.4.7 Look for Somatically Rearranged Genomic Regions

The next step is to look for rearranged genomic regions, and this can be done in several ways. The first way would be to sequence the genome from the cell or the tissue where the domesticated transposase is expressed (sequence 1) compare this genome with a reference genome (sequence 2) from the same species (same individual) in order to look for the rearranged sequence in sequence 1. Another way is to use high throughput genome translocation sequencing methodology based on the ability of a double-strand break to translocate to a fixed 'bait' double-strand break generated by a nuclease (Hu et al. 2016).

In both cases, the genomic region, where the rearrangement occurred, should be flanked by terminal repeats; these terminal repeats corresponding to the RSS should be compared to the TIR of the active transposon (the most similar one).

Candidate regions using the second methodology have been described by Wei et al. (2016). They described 27 recurrent double-strand breaks that occurred during neural stem/progenitor cell differentiation and they named it as recurrent double-strand break cluster. The authors hypothesized that the break region will be joined to the distal break region thus potentially leading to new genetic information via for example a novel exon combination that can be generated by recombination between intronic regions at the DNA level via "exon shuffling." (Alt and Schwer 2018) In the case where the double-strand break followed by a recombination occurs via the action of domesticated DDE transposon a possible candidate could be PGBD5. Indeed PGBD5 seems to be expressed in the brain and fetal brain and therefore possibly in the neural stem/progenitor cells (Pavelitz et al. 2013) and PGBD5 is able to induce double double-strand breaks in non-physiologic condition (Henssen et al. 2015, 2017).

14.4.8 Are There Other Domesticated Transposases that Are Able to Generate Diversity?

One important question concerns the possibility that some domesticated transposase and their TIRs could create diversity through the rearrangement of distinct tandem

UNCORRECTED PROOF

Author Proof

repeat paralogous gene segments (such as in V(D)J recombination). These processes correspond to an excision between at least 2 sets of paralogous fragments with a combinational joining. In these cases, the paralogous fragments should be flanked by the TIR like sequences located 3′ from each of the first paralogous fragments family and 5′ to each of the second fragment paralogue family. Such genomic rearrangement can be searched. The corresponding transposon can be searched as described above.

14.5 Conclusion

We discussed in this chapter the possibility that DDE transposons have been recruited several times as systems involved in programmed DNA elimination and perhaps in the generation of receptor diversity. This is due to the fact that the functional shift is an easy one and that the huge number of transposase domains are present and across the tree of life. Therefore, DDE transposons along with other so-called selfish element encoding nucleases (homing nuclease and prophage) could have been major players in the evolution of biodiversity. The co-option of DDE transposons as the regulatory element is also very important. This has been largely discussed, Justin Goodrich this Issue) (Sinzelle et al. 2009; Jangam et al. 2017). Another important role of DDE transposons is the one in the horizontal gene transfer (HGT). Horizontal transfer of DDE transposons (HTT) has been widely reported in eukaryotes (El Baidouri and Panaud 2015). However, in most cases, the literature discusses HGT and HTT in a non-integrated manner. However, the role of DDE transposons as the carrier of additional antibiotic resistance genes is well known in bacteria (Babakhani and Oloomi 2018), and at least the role of DDE transposon as carrier one case in yeast is also known. (McDonald et al. 2019). It is likely that most of the HGT are driven by transposons, but as most of the studied cases HGT corresponds to ancient events the transposase and the corresponding TIR has been lost so an effort in the analysis of recent HGTs should be done. Therefore, DDE transposons could have been major players in organismal evolution because it helps in genetic exchange between species.

The DDE transposon evolutionary trajectory should be included in the public goods hypothesis for the evolution of life on Earth (McInerney et al. 2011). According to this hypothesis, nucleotide sequences are simply seen as goods, passed from one organism to another through both vertical and horizontal transfer. The interesting things about DDE transposons are that they evolved in vertical and horizontal manner and they should be seen as goods since they increase the possibility of transfer and can be co-opted for example as recombination activating site-specific endonucleases involved in programmed DNA elimination.

Acknowledgements This work was supported by the French Government under the «Investissements d'avenir» (Investments for the Future) program managed by the Agence Nationale de la Recherche (ANR, fr: National Agency for Research), (reference: Méditerranée Infection 10-IAHU-03).

UNCORRECTED PROOF

Author Proof

References

Abe K, Shimizu SY, Tsuda S, Sato T (2017a) A novel non prophage(-like) gene-intervening element within gerE that is reconstituted during sporulation in Bacillus cereus ATCC10987. Sci Rep 7:11426. https://doi.org/10.1038/s41598-017-11796-8

Abe K, Takamatsu T, Sato T (2017b) Mechanism of bacterial gene rearrangement: SprA-catalyzed precise DNA recombination and its directionality control by SprB ensure the gene rearrangement and stable expression of spsM during sporulation in Bacillus subtilis. Nucleic Acids Res 45:6669–6683. https://doi.org/10.1093/nar/gkx466

Alt FW, Schwer B (2018) DNA double-strand breaks as drivers of neural genomic change, function, and disease. DNA Repair (Amst) 71:158–163

Arakawa H, HauschiLd J, Buerstedde JM (2002) Requirement of the activation-induced deaminase (AID) gene for immunoglobulin gene conversion. Science (80-) 295:1301–1306. https://doi.org/10.1126/science.1067308

Arnaoty A, Pitard B, Bateau B, Bigot Y, Lecomte T (2012) Novel approach for the development of mew antibodies directed against transposase-derived proteins encoded by human neogenes. Methods Mol Biol 859:293–305. https://doi.org/10.1007/978-1-61779-603-6_17

Aziz RK, Breitbart M, Edwards RA (2010) Transposases are the most abundant, most ubiquitous genes in nature. Nucleic Acids Res 38:4207–4217. https://doi.org/10.1093/nar/gkq140

Babakhani S, Oloomi M (2018) Transposons: the agents of antibiotic resistance in bacteria. J Basic Microbiol 58:905–917

Barsoum E, Martinez P, Åström SU (2010) α3, a transposable element that promotes host sexual reproduction. Genes Dev 24:33–44. https://doi.org/10.1101/gad.557310

Baudry C, Malinsky S, Restituito M, Kapusta A, Rosa S, Meyer E, Bétermier M (2009) PiggyMac, a domesticated piggyBac transposase involved in programmed genome rearrangements in the ciliate Paramecium tetraurelia. Genes Dev 23:2478–2483. https://doi.org/10.1101/gad.547309

Boehm T, Hirano M, Holland SJ, Das S, Schorpp M, Cooper MD (2018) Evolution of alternative adaptive immune systems in vertebrates. Annu Rev Immunol 36:19–42. https://doi.org/10.1146/annurev-immunol-042617-053028

Bouallègue M, Rouault JD, Hua-Van A, Makni M, Capy P (2017) Molecular evolution of piggyBac superfamily: from selfishness to domestication. Genome Biol Evol 9:323–339. https://doi.org/10.1093/gbe/evw292

Cheng CY, Young JM, Lin CYG, Chao JL, Malik HS, Yao MC (2016) The piggyBac transposon-derived genes TPB1 and TPB6 mediate essential transposon-like excision during the developmental rearrangement of key genes in Tetrahymena thermophila. Genes Dev 30:2724–2736. https://doi.org/10.1101/gad.290460.116

Eddy SR (2011) Accelerated profile HMM searches. PLoS Comput Biol. https://doi.org/10.1371/journal.pcbi.1002195

El Baidouri M, Panaud O (2015) Horizontal transfers and the new model of TE-driven genome evolution in Eukaryotes. In: Evolutionary biology: biodiversification from genotype to phenotype. Springer International Publishing, pp 77–92

Feiner R, Argov T, Rabinovich L, Sigal N, Borovok I, Herskovits AA (2015) A new perspective on lysogeny: prophages as active regulatory switches of bacteria. Nat Rev Microbiol 13:641–650

Feschotte C (2008) Transposable elements and the evolution of regulatory networks. Nat Rev Genet 9:397–405

Flajnik MF (2016) Evidence of G.O.D'.s miracle: unearthing a RAG transposon. Cell 166:11–12

Henssen AG, Henaff E, Jiang E, Eisenberg AR, Carson JR, Villasante CM, Ray M, Still E, Burns M, Gandara J, Feschotte C, Mason CE, Kentsis A (2015) Genomic DNA transposition induced by human PGBD5. Elife. https://doi.org/10.7554/elife.10565

Henssen AG, Jiang E, Zhuang J, Pinello L, Socci ND, Koche R, Gonen M, Villasante CM, Armstrong SA, Bauer DE, Weng Z, Kentsis A (2016) Forward genetic screen of human transposase genomic rearrangements. BMC Genom 17:548. https://doi.org/10.1186/s12864-016-2877-x

UNCORRECTED PROOF

Author Proof

Henssen AG, Koche R, Zhuang J, Jiang E, Reed C, Eisenberg A, Still E, Macarthur IC, Rodríguez-Fos E, Gonzalez S, Puiggròs M, Blackford AN, Mason CE, De Stanchina E, Gönen M, Emde AK, Shah M, Arora K, Reeves C, Socci ND, Perlman E, Antonescu CR, Roberts CWM, Steen H, Mullen E, Jackson SP, Torrents D, Weng Z, Armstrong SA, Kentsis A (2017) PGBD5 promotes site-specific oncogenic mutations in human tumors. Nat Genet 49:1005–1014. https://doi.org/10.1038/ng.3866

Hilton JA, Meeks JC, Zehr JP (2016) Surveying DNA elements within functional genes of heterocyst-forming cyanobacteria. PLoS One. https://doi.org/10.1371/journal.pone.0156034

Hoen DR, Bureau TE (2015) Discovery of novel genes derived from transposable elements using integrative genomic analysis. Mol Biol Evol 32:1487–1506. https://doi.org/10.1093/molbev/msv042

Hu J, Meyers RM, Dong J, Panchakshari RA, Alt FW, Frock RL (2016) Detecting DNA double-stranded breaks in mammalian genomes by linear amplification-mediated high-throughput genome-wide translocation sequencing. Nat Protoc 11:853–871. https://doi.org/10.1038/nprot.2016.043

Huang S, Tao X, Yuan S, Zhang Y, Li P, Beilinson HA, Zhang Y, Yu W, Pontarotti P, Escriva H, Le Petillon Y, Liu X, Chen S, Schatz DG, Xu A (2016) Discovery of an active RAG transposon illuminates the origins of V(D)J recombination. Cell 166:102–114. https://doi.org/10.1016/j.cell.2016.05.032

Jangam D, Feschotte C, Betrán E (2017) Transposable element domestication as an adaptation to evolutionary conflicts. Trends Genet 33:817–831

Jeffares DC, Tomiczek B, Sojo V, dos Reis M (2015) A beginners guide to estimating the non-synonymous to synonymous rate ratio of all protein-coding genes in a genome. Methods Mol Biol 1201:65–90. https://doi.org/10.1007/978-1-4939-1438-8_4

Kapitonov VV, Jurka J (2005) RAG1 core and V(D)J recombination signal sequences were derived from Transib transposons. PLoS Biol 3:0998–1011. https://doi.org/10.1371/journal.pbio.0030181

Keeling PJ, Roger AJ (1995) The selfish pursuit of sex. Nature 375:283

Klar AJS, Ishikawa K, Moore S (2014) A unique DNA recombination mechanism of the mating/cell-type switching of fission yeasts: a review. Microbiol Spectr. https://doi.org/10.1128/microbiolspec.mdna3-0003-2014

Koufopanou V, Burt A (2005) Degeneration and domestication of a selfish gene in yeast: molecular evolution versus site-directed mutagenesis. Mol Biol Evol 22:1535–1538. https://doi.org/10.1093/molbev/msi149

Losos JB (2011) Convergence, adaptation, and constraint. Evolution (N Y) 65:1827–1840. https://doi.org/10.1111/j.1558-5646.2011.01289.x

Majumdar S, Singh A, Rio DC (2013) The human THAP9 gene encodes an active P-element DNA transposase. Science (80-) 339:446–448. https://doi.org/10.1126/science.1231789

McDonald MC, Taranto AP, Hill E, Schwessinger B, Liu Z, Simpfendorfer S, Milgate A, Solomon PS (2019) Transposon-mediated horizontal transfer of the host-specific virulence protein ToxA between three fungal wheat pathogens. MBio. https://doi.org/10.1128/mbio.01515-19

McInerney JO, Pisani D, Bapteste E, O'Connell MJ (2011) The public goods hypothesis for the evolution of life on Earth. Biol Direct 6:41. https://doi.org/10.1186/1745-6150-6-41

Morales Poole JR, Huang SF, Xu A, Bayet J, Pontarotti P (2017) The RAG transposon is active through the deuterostome evolution and domesticated in jawed vertebrates. Immunogenetics 69:391–400. https://doi.org/10.1007/s00251-017-0979-5

Nowacki M, Higgins BP, Maquilan GM, Swart EC, Doak TG, Landweber LF (2009) A functional role for transposases in a large eukaryotic genome. Science 324:935–938

Park PJ (2009) ChIP-seq: advantages and challenges of a maturing technology. Nat Rev Genet 10:669–680

Pavelitz T, Gray LT, Padilla SL, Bailey AD, Weiner AM (2013) PGBD5: a neural-specific intron-containing piggyBac transposase domesticated over 500 million years ago and conserved from cephalochordates to humans. Mob DNA 4:23. https://doi.org/10.1186/1759-8753-4-23

UNCORRECTED PROOF

Author Proof

526 Pontarotti P, Hue I (2016) Road map to study convergent evolution: A proposition for evolutionary
527 systems biology approaches. Evolutionary biology: convergent evolution, evolution of complex
528 traits, concepts and methods. Springer International Publishing, Heidelberg, pp 3–21
529 Rajaei N, Chiruvella KK, Lin F, Åström SU (2014) Domesticated transposase Kat1 and its fossil
530 imprints induce sexual differentiation in yeast. Proc Natl Acad Sci U S A 111:15491–15496.
531 https://doi.org/10.1073/pnas.1406027111
532 Rascol VL, Levasseur A, Chabrol O, Grusea S, Gouret P, Danchin EGJ, Pontarotti P (2009)
533 CASSIOPE: an expert system for conserved regions searches. BMC Bioinf 10:284. https://doi.
534 org/10.1186/1471-2105-10-284
535 Sinzelle L, Izsvák Z, Ivics Z (2009) Molecular domestication of transposable elements: from
536 detrimental parasites to useful host genes. Cell Mol Life Sci 66:1073–1093
537 Teng G, Schatz DG (2015) Regulation and evolution of the RAG recombinase. In: Advances in
538 immunology. Academic Press Inc, pp 1–39
539 Wang J, Davis RE (2014) Programmed DNA elimination in multicellular organisms. Curr Opin
540 Genet Dev 27:26–34
541 Wei PC, Chang AN, Kao J, Du Z, Meyers RM, Alt FW, Schwer B (2016) Long neural genes harbor
542 recurrent DNA break clusters in neural stem/progenitor cells. Cell 164:644–655. https://doi.org/
543 10.1016/j.cell.2015.12.039
544 Yu K, Lieber MR (2019) Current insights into the mechanism of mammalian immunoglobulin class
545 switch recombination. Crit Rev Biochem Mol Biol 54:333–351
546 Zhang Y, Cheng TC, Huang G, Lu Q, Surleac MD, Mandell JD, Pontarotti P, Petrescu AJ, Xu A,
547 Xiong Y, Schatz DG (2019) Transposon molecular domestication and the evolution of the RAG
548 recombinase. Nature 569:79–84. https://doi.org/10.1038/s41586-019-1093-7
549 Zufall RA, Robinson T, Katz LA (2005) Evolution of developmentally regulated genome
550 rearrangements in eukaryotes. J Exp Zool Part B Mol Dev Evol 304:448–455

UNCORRECTED PROOF

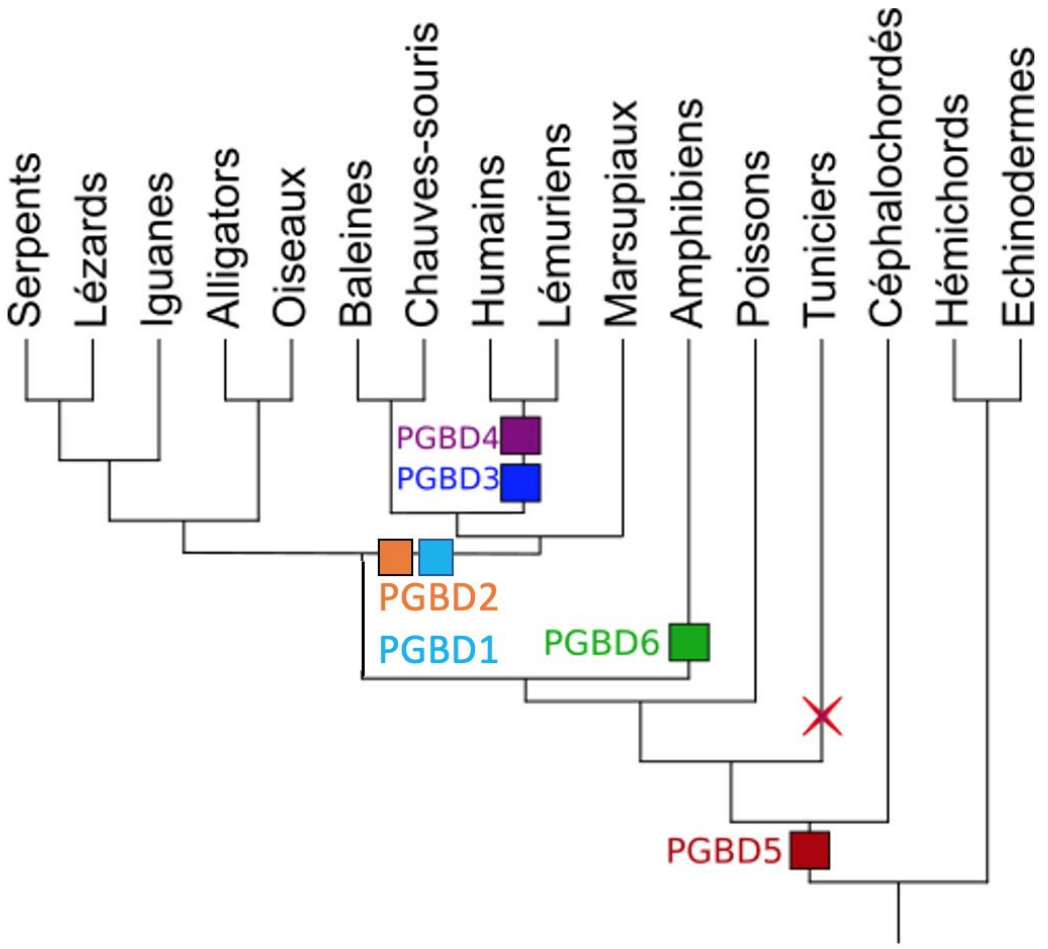


Figure 9 - Répartition des *pgbd* chez les vertébrés.

Chaque carré coloré représente un *pgbd*. Les *pgbd1* et *pgbd2* se trouvent chez les mammifères, les *pgbd3* et *pgbd4* se trouvent chez les primates, le *pgbd5* est présent chez tous les vertébrés et céphalochordés mais pas chez les tuniciers. Le *pgbd6* (anciennement Kobuta) est spécifique aux anoures.

2. Domestication de gènes dérivés de transposons de la famille de piggyBac

Chez les ciliés des transposases *piggyBac* domestiquées jouent un rôle essentiel dans la différentiation au cours du cycle cellulaire. PiggyMac chez Paramecium tetraurelia et TPB2 chez Tetrahymena thermophila (Baudry *et al.*, 2009; Cheng *et al.*, 2010, 2016) interviennent lors du développement du macronoyau à partir du micronoyau en éliminant de courtes séquences internes (Internal Eliminated Sequences, IES) ce qui permet de reconstruire des gènes fonctionnels.

D'après la littérature, il existe six cas de gènes codant une transposase *piggyBac* domestiquées (PGBD) et ayant des histoires évolutives différentes dans les génomes de vertébrés, PGBD1 à 6. Parmi ces transposases domestiquées *piggyBac*, 5 sont présentes dans le génome de l'homme, *pgbd*1 à 5 (pour *piggyBac transposable element derived*) (**Figure 9**).

L'organisation des gènes codants ces transposases domestiquées diffère notamment par leur nombre d'exons. Dans certains cas, la séquence codante dérivée de la transposase est fusionnée avec un autre domaine.

2.1. *Pgbd1*

Pgbd1 est présent chez tous les mammifères, il est composé de 7 exons. *Pgbd1* est un gène sous pression de sélection, comme l'indique son ratio dN/dS (aussi appelé ka/ks) qui est évalué à 0,432. Ce ratio permet d'évaluer la contrainte sélective sur un gène. Il résulte du rapport du nombre de substitutions non synonymes dN sur le nombre de substitutions synonymes dS. Sa valeur permet d'estimer quel type de pression de sélection s'exerce au cours du temps. La sélection est purifiante si dN/dS < 1, neutre si dN/dS = 1 ou positive/diversifiante si dN/dS > 1 (Hurst, 2002). Pour *pgbd1*, la valeur est proche de 0 indiquant une sélection purifiante en faveur d'une domestication fonctionnelle. L'histoire évolutive d'un gène domestiqué et son origine permet une meilleure compréhension et identification du rôle de ce gène. À ce jour, aucun *pble* proche du *pble* ancêtre à l'origine de *pgbd1* n'a été identifié.

Chez les mammifères, la protéine PGBD1 varie de 312 à 826 acides aminés selon l'annotation disponible dans les banques de données (Bouallègue *et al.*, 2017). Chez l'humain, PGBD1 est une protéine de 809 aa. Elle est composée de deux domaines. Le domaine N-terminal de 290 aa correspond à un domaine Zn-Scan (**Figure 10**). Le domaine C-terminale de 519 aa correspondant à une transposase de *piggyBac.* Le domaine catalytique de PGBD1 est

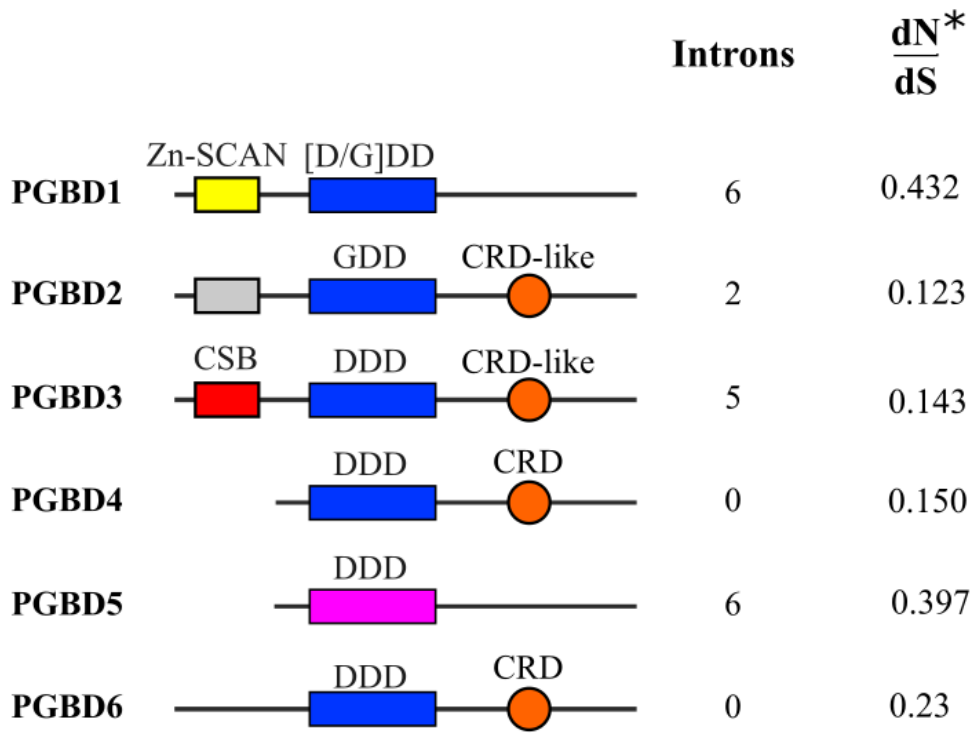


Figure 10 - Caractéristiques des PGBD chez les vertébrés.

Pour chaque séquence domestiquée, la case colorée représente des domaines. Le jaune représente le domaine Zn-SCAN, le gris un domaine inconnu, le rouge le domaine CSB, le bleu clair le domaine SCAN et le vert le domaine zF-FCS. Le cercle orange représente les domaines CRD ou similaires. Le bleu est pour la triade catalytique avec la composition des résidus. La triade catalytique est en rose pour le PGBD5 car elle n'est pas alignée avec les autres séquences domestiquées et avec IFP2. * Le dN/dS est reporté de la publication de (Bouallègue *et al.*, 2017).

différent de celui de IFP2. En effet, selon les espèces, le premier acide aminé n'est pas toujours conservé et la triade peut être de la forme [G/D]DND. Contrairement à IFP2, PGBD1 ne possède pas de CRD (Sarkar *et al.*, 2003; Bouallègue *et al.*, 2017). La similitude moyenne calculée entre les protéines PGBD1 (62 espèces) est de 85 % (Bouallègue *et al.*, 2017).

La fonction de PGBD1 est inconnue. Dans certaines banques de données comme GenBank (Accession No. D88259), PGBD1 est aussi connue sous le nom de HUCEP-4, une protéine cérébrale nommée HUCEP-4 (Sarkar *et al.*, 2003; Sinzelle, Izsvák and Ivics, 2009).

2.2. *Pgbd2*

Pgbd2 est présent chez tous les mammifères. Le gène est composé d'un seul exon. Le ratio dN/dS a été estimé à 0,123 (Bouallègue *et al.*, 2017). Aucun *pble* proche de l'ancêtre de *pgbd2* n'a pu être mis en évidence.

Chez les mammifères, la séquence de PGBD2 varie de 586 à 759 acides aminés. Chez l'humain, PGBD2 est une protéine de 592 acides aminés. Le domaine de catalytique de PGBD2 n'est pas conservé (GDGD) suggérant que cette protéine n'est pas une recombinase. Elle a probablement été domestiquées sur la base de ses autres activités, liaison à l'ADN et capacité à interagir avec d'autres protéines et a probablement un rôle autre (**Figure 10**). PGBD2 possède un CRD (Sarkar *et al.*, 2003; Bouallègue *et al.*, 2017). La similitude moyenne calculée entre ces protéines (62 espèces) est de 88 % (Bouallègue *et al.*, 2017). Sa fonction reste inconnue.

D'après la littérature, PGBD1 et PGBD2 semblent évolutivement proches de PGBD3 qui est présent uniquement chez les primates (Bouallègue *et al.*, 2017).

2.3. *Pgbd3*

Pgbd3 est présent uniquement chez les primates. Ce gène aussi appelé *ERCC6* (*Excision Repair 6, Chromatin Remodeling Factor*) est composé de 6 exons. L'exon dérivant de la transposase piggyBac est le sixième exon. Les exons 1 à 5 correspondent au gène responsable du syndrome du groupe B de Cockayne (CSB). Le ratio dN/dS est estimé à 0,143. L'élément *pble* ancestral le plus proche de *pgbd3*, Smi-pble7, a été identifié chez l'araignée *Stegodyphus mimosarum* (Bouallègue *et al.*, 2017).

L'épissage alternatif de ce gène permet la formation d'une protéine de fusion CSB-PGBD3 de 1061 acides aminés incluant 468 résidus de CSB. Chez l'humain, PGBD3 est composée de 593 aa. Le domaine catalytique de PGBD3 comme PGBD1 & 2 n'est pas conservé. En effet, les résidus s'alignant avec IFP2 sont de la forme DN[D/G]D. PGBD3 possède un domaine CRD (**Figure 10**). La similitude moyenne calculée entre les protéines (15 espèces) est de 98 % (Sarkar *et al.*, 2003; Newman *et al.*, 2008; Bouallègue *et al.*, 2017).

CSB-PGBD3 est capable de se lier à l'ADN, y compris aux extrémités de transposons endogènes MER85 de types dans le génome humain. Cependant, aucune activité catalytique n'a été détectée. Des éléments de preuves biochimiques et génétiques suggèrent qu'elle pourrait participer dans les mécanismes de réparation de l'ADN (Bailey *et al.*, 2012; Gray *et al.*, 2012; Henssen *et al.*, 2015). Il a aussi été proposé que la transposase PGBD3 ait été initialement domestiquée pour réprimer la transposition de *piggyBac* et de l'élément non autonome associé MER85 (Newman *et al.*, 2008; Sinzelle, Izsvák and Ivics, 2009). D'autres auteurs ont émis l'hypothèse que la protéine de fusion CSB-PGBD3 avec les transposons *piggyBac* non autonomes MER85 peuvent fournir un réseau de régulation génique similaire à celui proposé pour la protéine SETMAR et ses sites de liaison génomique putatifs dérivés des TIR *Hsmar1* ou encore comme le gène de fusion *Krabiner* qui se lie à ses transposons apparentés et contrôle un vaste réseau de gènes et d'éléments cis-régulateurs (Cordaux *et al.*, 2006; Sinzelle, Izsvák and Ivics, 2009; Cosby, Chang and Feschotte, 2019; Tellier and Chalmers, 2019).

2.4. *Pgbd4*

Pgbd4 est présent chez tous les primates. Le gène est composé d'un seul exon. Le ratio *dN/dS* est estimé à 0,150 (**Figure 10**). Un *pble* a été défini comme étant proche de l'ancêtre de *pgbd4*, il s'agit de *Pva-pble*, un élément présent chez la roussette de Malaisie, *Pteropus vampyrus* (**Figure 8**) (Bouallègue *et al.*, 2017).

Chez les primates, la séquence protéique de PGBD4 varie de 585 à 603 acides aminés. Chez l'humain, la protéine fait 585 aa. La triade catalytique de la transposase de piggybac est conservée chez PGBD4. La similitude moyenne calculée entre les orthologues (16 espèces) est de 98 % (Bouallègue *et al.*, 2017). PGBD4 possède un domaine CRD (**Figure 10**). Bien que PGBD4 possède un domaine catalytique conservé, aucune donnée n'a à ce jour été publiée pour montrer qu'il s'agit d'une transposase active. La fonction cellulaire de PGBD4 est pour l'instant non élucidée.

2.5. *Pgbd5*

Le gène *pgbd5* a été domestiqué il y a environ 500 millions d'années (Sarkar *et al.*, 2003; Pavelitz *et al.*, 2013). Il résulte d'un événement de domestication chez l'ancêtre commun des céphalochordés et des vertébrés et il aurait été perdu chez l'ancêtre commun des tuniciers (Delsuc *et al.*, 2006; Putnam *et al.*, 2008). Le gène est composé de plusieurs introns (six chez les céphalochordés, sept chez les agnathes et six chez les autres vertébrés). Le ratio *dN/dS* pour *pgbd5* est estimé à 0,397 (**Figure 10**). Aucun élément ancestral à *pgbd5* n'a pu être identifié à ce jour.

Chez l'humain, la protéine PGBD5 existe sous plusieurs isoformes. 4 isoformes ont été identifiées à ce jour chez l'humain (Hs 554 ; Hs 524 ; Hs 455 et Hs 409). Chez la souris, on retrouve deux isoformes : Mm409 et Mm523 qui correspondent respectivement aux isoformes humaines Hs 409 et Hs 524. La séquence de la protéine PGBD5 varie de 343 acides aminés (*Equus caballus*) à 732 acides aminés (*Elephantulus edwardii*) (Sarkar *et al.*, 2003; Pavelitz *et al.*, 2013; Bouallègue *et al.*, 2017). Cette différence de séquence est due à une faible conservation de la région N-terminale. En effet, si on regarde la conservation de ces séquences en intégrant la région N-terminal elle est de 76 %. Si on retire la région N-terminale, la conservation est de 87 % (147 espèces) (Bouallègue *et al.*, 2017). PGBD5 ne possède pas de CRD et bien que la présence d'une triade catalytique ait été démontrée, cette dernière ne s'aligne pas avec la triade de la transposase de l'élément de référence *Ifp2*.

L'expression de PGBD5 chez l'humain et la souris semble largement limitée à l'embryon précoce et à certaines zones du cerveau embryonnaire et adulte. La protéine PGBD5 est détectée dans les neurones pyramidaux et granulaires de l'hippocampe, de l'amygdale et du bulbe olfactif qui sont respectivement impliqués dans les mémoires spatiale, contextuelle, émotionnelle et olfactive (Sarkar *et al.*, 2003; Pavelitz *et al.*, 2013). PGBD5 s'est révélée être aussi un déclencheur de cancer mais son rôle dans les tissus sains reste encore inconnu (Henssen *et al.*, 2015). Les travaux de Henssen et al. (2015) révèle que PGBD5 est une transposase totalement fonctionnelle. De façon assez étonnante, PGBD5 est capable de transposer l'élément *Ifp2*, le *pble* le plus éloigné de PGBD5 (Bouallègue *et al.*, 2017).

2.6. *Pgbd6*

Dans mon manuscrit, *pgbd6* désigne le gène de la transposase domestiquée *Kobuta* présent chez Xenopus. Le gène de *pgbd6* ne possède qu'un seul exon. Les éléments les plus proches du gène domestiqué qui ont été mis en évidence sont *uribo1* et *uribo2* dans le genre Xenopus. Le gène domestiqué (*pgbd6*) et les *pble* (*uribo1* et *2*) coexistent dans les mêmes génomes.

La protéine PGBD6 est constituée de 610 *aa* (**Figure 10**). Elle possède un domaine catalytique de type [D/N] DND ainsi qu'un CRD (Hikosaka *et al.*, 2007). La fonction de PGBD6 n'est pas encore tout à fait élucidée (Hikosaka *et al.*, 2007). PGBD6 semble être impliquée dans une activité de liaison ou de recombinaison à l'ADN pouvant inactiver les éléments autonomes par hétérodimérisation. Une hypothèse serait qu'elle pourrait inactiver les transposases URIBO par hétérodimérisation de sorte que les transposases actives codées par les *pble uribo* pourraient être sous contrôle et ne pas être nocifs chez l'hôte (Hikosaka *et al.*, 2007).

2.7. *NeoPgbd*

Chez les poissons téléostéens (*Actinopterygii*), de potentiels *pgbd* ont été mis en évidence, appelés *NeoPgbd* (Bouallègue *et al.*, 2017). *NeoPgbd* ne possède qu'un seul exon. Ils sont décrits comme possédant un ORF putatif présentant une forte homologie avec le domaine de la transposase de l'élément *ifp2*. Ils ne présentent pas de promoteurs identifiables, de signaux d'épissage ou sites de liaison de facteur de transcription. La valeur de dN/dS estimée est de 0,094 indiquant une forte sélection purifiante. Aucun *pble* proche de ce gène n'a été identifié jusqu'à présent.

Les transposases NeoPGBD putatives aurait une séquence de 649 à 682 acides aminés. Elles ont 86 % de similitude entre elles. Aucun domaine protéique associé n'a été détecté. Le rôle de ces transposases NeoPGBD n'est pas encore connu.

Objectifs de la thèse

Les éléments génétiques mobiles (EGM) peuvent être impliqués dans l'expression génique lors de reprogrammations chromatiniennes. Certains EGM sont activés lors de processus de différenciation cellulaire comme la gamétogénèse, la mémorisation ou à la suite de stress ou de traitement pharmacologiques. Certains de ces EGM expriment des transposases responsables de leurs propriétés de transposition. Parmi l'ensemble des transposases certaines sont impliquées dans des réarrangements programmés de l'ADN chromosomique nécessaire à la physiologie de l'hôte.

La protéine PGBD5 aurait cette propriété et serait capable d'induire des réarrangements chromosomiques programmés au sein du génome des neurones pyramidaux et granulaires de l'hippocampe, de l'amygdale et du bulbe olfactif. Lors de ces réarrangements, elle pourrait « bouger » des éléments de régulation transcriptionnel et épigénétique (Bire *et al.*, 2016) contenus dans les *pble* et/ou des fragments chromosomiques se comportant comme des *pble*. Par ailleurs, le gène *pgbd5* et la protéine PGBD5 qu'il code sont très conservés chez tous les vertébrés et céphalochordés, le mécanisme d'action qui leur est lié pourrait aussi être très conservé et physiologiquement très important chez les vertébrés.

L'expression inappropriée de PGBD5 pourrait induire l'apparition et le développement de tumeurs rhabdoïdes chez l'enfant (Henssen *et al.*, 2016, 2017).

Une analyse de la banque de données protéomiques PRIDE révèle que PGBD5 est présente dans 100 % des 80 cas étudiés de cancers colorectaux. Dans ce type de cancer, le niveau de transcription de *pgbd5* est corrélé avec le pronostic vital du patient.

L'objectif principal de mes travaux de thèse est de mettre en évidence les cibles de liaison de PGBD5 sur l'ADN des chromosomes. *Pgbd5* étant très conservé depuis l'origine des vertébrés on peut supposer que son mécanisme d'action et ses cibles le sont aussi. Ainsi mes travaux de thèse ont contribué à répondre aux questions suivantes :

i) Est-ce que *pgbd5* partage ou non un ancêtre commun avec les autres *pgbd*, et donc par extensions est capable d'utiliser les sites de liaison à l'ADN des autres *pgbd* ?

ii) Est-ce que PGBD5 est capable de reconnaître des TIR de *pble* proches de son élément ancestral ?

PGBD5 et IFP2 sont aussi eux deux transposases phylogénétiquement éloignées dans la famille des *pble.* Elles présentent en plus une différence d'un domaine à leur extrémité C-terminale

(PGBD5 n'ayant pas de domaine C-terminal correspondant à un CRD contrairement à IFP2). Cela soulève la question :

iii) Le CRD dans IFP2 est-il nécessaire à la catalyse de la transposition ?

iv) S'il existe plusieurs types de *pble* dans un génome, PGBD5 pourrait-elle être capable de lier leurs extrémités et éventuellement les mobiliser pour la transposition ou à minima pour l'excision ?

Trois grands axes de travaux ont été développés pendant ma thèse afin de répondre à ces questions :

Le premier axe de recherche concerne l'efficacité de transposition d'une transposase IFP2/PB sans le domaine CRD, sachant qu'une transposase *piggyBac* domestiquée comme PGBD5 a perdu ce domaine au cours de la domestication. L'objectif de mes travaux a été **de vérifier si le domaine C-terminal est indispensable à l'activité de transposase-recombinase ou s'il n'est qu'un domaine accessoire de régulation de l'activité.**

Le second concerne l'histoire évolutive de *pgbd5*. Aucun *pble* proche de *pgbd5* n'avait été mis en évidence jusqu'à aujourd'hui et l'existence de certaines ambiguïtés sur l'origine de certains *pgbd* nous a poussés à mettre à jour et répertorier les informations concernant ces gènes chez les vertébrés. Ensuite, nous avons procédé à la recherche des éléments les plus proches à chaque *pgbd*. L'objectif de cette première étude a été d'**identifier le *pble* phylogénétiquement le plus proche de *pgbd5*, puis de vérifier s'il était un meilleur substrat qu'*Ifp2* pour la transposition catalysée par PGBD5.**

Le troisième axe de recherche est porté sur les capacités de PGBD5 à exciser puis réintégrer dans de nouveaux locus des *pble* humains. La séquence de ces *pble* humains n'a comme point commun que la présence de TIR et de structures secondaires sachant que la séquence primaire de ces éléments diverge. L'objectif de cette troisième étude est **d'évaluer la capacité de PGBD5 à mobiliser des éléments *pble* présents dans le génome humain**. Cela nous permettra d'identifier plus précisément les cibles de la transposase.

Au cours de ce travail et afin de répondre aux objectifs de la thèse, nous avons développé des outils et des pipelines informatiques. Ils sont présentés dans un quatrième chapitre.

Résultats

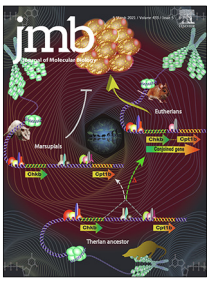
Chapitre I. Le domaine C-terminal de la transposase piggyBac n'est pas nécessaire pour la transposition de l'ADN

1. Contexte et objectifs de l'article

J'ai évalué si le domaine C-terminal des transposases *piggyBac* est nécessaire pour l'activité de transposition. En effet, contrairement à la transposase de *ifp2*, PGBD5 ne possède pas de domaine CRD. Nous avons décidé de tester l'activité de la transposase *piggyBac* sans le domaine CRD. Des études précédentes ont proposé des rôles non exclusifs pour le CRD. Dans la première étude, les auteurs ont conclu que le CRD était essentiel pour la transposition d'*ifp2*, car l'utilisation de PB déficient en CRD (PB1-558) dans un essai d'intégration réalisé sur des cellules de mammifères n'a pas entrainé d'augmentation de l'activité d'intégration par rapport aux contrôles effectués en l'absence de transposase. La deuxième étude a proposé que la dimérisation de PB puisse servir à empêcher une transposition excessive d'*ifp2*, car la suppression du CRD conduit les monomères de PB monomérique à être plus actif en termes de clivage de l'ADN. Cependant, chez *ifp2*, le domaine CRD contient aussi le signal de localisation nucléaire (NLS) qui est requis pour son transport vers le noyau de la cellule. Dans les deux études, aucune n'a considéré que l'absence du CRD impliquait une rétention cytosolique car le NLS était absent. Nous avons donc cherché à vérifier si PB1-558 fusionné à un NLS de virus simien 40 (SV40) était capable d'assurer la transposition de *ifp2* dans des chromosomes de cellules humaines.

2. Résumé des principaux résultats

Nous avons montré que PB1-558 et PB.NLS-1-558 avaient tous les deux un effet négatif sur l'obtention de clones dans les essais d'intégration. Cela explique probablement pourquoi aucune activité d'intégration élevée n'avait été détectée auparavant. Deuxièmement, nous avons découvert que PB.NLS-1-558 est capable d'effectuer la transposition des éléments *ifp2*. Troisièmement, nous avons comparé les propriétés de PB.NLS-1-558 avec celles d'une autre protéine *piggyBac* déficiente en CRD, la protéine murine domestiquée PGBD5. Quatrièmement, nous avons évalué la qualité des extrémités *ifp2* néo-intégrées dans les chromosomes par PB et les deux protéines déficientes en CRD, PB.NLS-1-558 et PGBD5. Enfin, nous avons vérifié si les trois protéines utilisaient des pools de sites d'insertion chromosomiques similaires ou différents, quel que soit l'état des extrémités de transposons *ifp2*. Ma contribution à cette publication s'est essentiellement portée sur l'analyse des sites d'insertion au niveau chromosomique et de l'élément.



The C-terminal Domain of *piggyBac* Transposase Is Not Required for DNA Transposition

Laura Helou [1], Linda Beauclair [1], Hugues Dardente [1], Peter Arensburger [2], Nicolas Buisine [3], Yan Jaszczyszyn [4], Florian Guillou [1], Thierry Lecomte [5], Alex Kentsis [6,7,8] and Yves Bigot [1]*

*1 - **PRC,** UMR INRAE 0085, CNRS 7247, Centre INRAE Val de Loire, 37380 Nouzilly, France*
*2 - **Biological Sciences Department,** California State Polytechnic University, Pomona, CA 91768, USA*
*3 - **UMR CNRS 7221,** Muséum National d'Histoire Naturelle, 75005 Paris, France*
*4 - **Université Paris-Saclay,** CEA, CNRS, Institute for Integrative Biology of the Cell (I2BC), 91198 Gif-sur-Yvette, France*
*5 - **EA GICC 7501,** CHRU de Tours, 37044 Tours Cedex 09, France*
*6 - **Molecular Pharmacology Program,** Sloan Kettering Institute, Memorial Sloan Kettering Cancer Center, New York, NY, USA*
*7 - **Weill Cornell Medical College,** Cornell University, New York, NY, USA*
*8 - **Department of Pediatrics,** Memorial Sloan Kettering Cancer Center, New York, NY, USA*

***Correspondence to Yves Bigot:** yves.bigot@inrae.fr (Y. Bigot)*
https://doi.org/10.1016/j.jmb.2020.166805
Edited by M.F. Summers

Abstract

PiggyBac(PB)-like elements (*pble*) are members of a eukaryotic DNA transposon family. This family is of interest to evolutionary genomics because *pble* transposases have been domesticated at least 9 times in vertebrates. The amino acid sequence of *pble* transposases can be split into three regions: an acidic N-terminal domain (~100 aa), a central domain (~400 aa) containing a DD[D/E] catalytic triad, and a cysteine-rich domain (CRD; ~90 aa). Two recent reports suggested that a functional CRD is required for *pble* transposase activity. Here we found that two CRD-deficient *pble* transposases, a PB variant and an isoform encoded by the domesticated PB-derived vertebrate transposase gene 5 (*pgbd5*) trigger transposition of the *Ifp2 pble*. When overexpressed in HeLa cells, these CRD-deficient transposases can insert *Ifp2* elements with proper and improper transposon ends, associated with deleterious effects on cells. Finally, we found that mouse CRD-deficient transposase *Pgbd5*, as well as PB, do not insert pbles at random into chromosomes. Transposition events occurred more often in genic regions, in the neighbourhood of the transcription start sites and were often found in genes predominantly expressed in the human central nervous system.

© 2021 Elsevier Ltd. All rights reserved.

Introduction

Transposable elements (TEs) gather diverse discrete DNA sequences of prokaryotic and eukaryotic origins that use a wide range of mobility mechanisms to transpose within the genome of their hosts.[1–4] *PiggyBac-like* elements (*pble*) consist of a family of DNA transposons that have so far only been found in animal genomes and with copy numbers that vary widely between host species.[5] *Pbles* are able to jump from one chromosomal locus to another using cut-and-paste transposition which is enzymatically catalysed by the transposase they encode. The first *pble* to be identified was *Ifp2* (a.k.a. *piggyBac*) from the cabbage looper moth *Trichoplusia ni* (Lepidoptera).[6] It is the reference element in the *piggyBac* family for academic research purposes.[7]

The *Ifp2* DNA sequence is 2476 bp in length and contains an open reading frame (ORF) coding a 594 amino acid transposase named PB. The *Ifp2* DNA sequence is flanked by 13 bp long terminal inverted repeats (TIR) and by 19 bp long subterminal inverted repeats (STIR) located internally at 3 and 31 nucleotides of distance of 5′ and 3′ TIR inner ends, respectively. PB excises the transposon from its donor chromosomal locus and reinserts it into a TTAA motif, which gets duplicated upon insertion.

Sequence analyses revealed that PB contains at least 3 domains. The first domain spans from residues 1 to 116, shows no overt structural features and displays an acidic pI of 4.41. The second domain is a macro domain that extends from residues 117 to 535 and has a basic pI (9.29); it contains several highly conserved residues, including the predicted catalytic residues (D268, D346 and D447) that are required for all transposition steps. On both sides of the catalytic domain (residues 263 to 457), recent structures obtained by cryo-electron-microscopy[8] revealed two DNA binding sub-domains (residues 117 to 263 and 457 to 535, respectively) that bind to *Ifp2* TIRs. Finally, the third domain spans from residues 559 to 594, and contains a cysteine-rich domain (CRD, pI = 9.07) for which the atomic structure was first solved by nuclear magnetic resonance.[9] This CRD was shown to bind to a 5′-TGCGT-3′/3′-ACGCA-5′ motif that is contained within the 19 bp subterminal inverted repeat between positions 178 and 199 of the *Ifp2* sequence.[9] The CRD was also found to be vital for the dimerization of PB;[10] the removal of the last seven residues being sufficient to yield a monomeric protein that still binds to *Ifp2* ends *in vitro*. Finally, this CRD contains a nuclear localisation signal (NLS) that is required to mediate PB nuclear localisation.[11]

Two studies[9,10] proposed non-exclusive roles for the CRD and the ability of PB to mediate transposition. In the first study, the authors concluded that the CRD was essential for *Ifp2* transposition because the use of a CRD-deficient PB (PB.1-558) in an integration assay performed in mammalian cells did not lead to an increase in integration activity compared to controls done in the absence of the transposase.[9] The second study proposed that PB dimerisation might serve to prevent excessive transposition of *Ifp2* since removal of the CRD led the monomeric PB to be more active in transposition excision.[10] It is also possible that the absence of the CRD may cause cytosolic retention since it contains a nuclear localization sequence (NLS).

Here we aimed to determine whether PB.1-558 fused to a simian virus 40 (SV40) NLS was able to mediate the transposition of *Ifp2* into human cell chromosomes. First, we showed that both PB.1-558 and PB.NLS-1-558 had a negative effect on obtaining clones in integration assays. Second, we found that PB.NLS-1-558 was able to carry out the transposition of *Ifp2* elements. Third, we compared the properties of PB.NLS-1-558 with those of another CRD-deficient *piggyBac* protein, the domesticated murine and human PGBD5 protein. Fourth, we evaluated the quality of *Ifp2* ends neo-integrated into chromosomes by PB and the two CRD-deficient proteins, PB.NLS-1-558 and PGBD5. Finally, we determined whether the three proteins used similar or different pools of chromosomal insertion sites, whatever the state of *Ifp2* transposon ends.

Results

PB.1-558 needs an NLS to locate into nuclei

We made a CRD-deficient PB mutant (PB.1-558) and a construct in which its N-terminal end was fused with a SV40 NLS (PB.NLS-1-558). Such a position for the NLS does not modify the transposition activity of PB and preserves the activity of the added localization motif or protein domain.[12,13] To assess cellular localization of both proteins we made two more constructs in which the green fluorescent protein (GFP) was C-terminally fused to PB.1-558 and PB.NLS-1-558. An expression vector encoding GFP was used as a diffusion control within the cytoplasm and nucleus (Figure 1(a)–(c)) and a vector encoding a PB-GFP fusion was used as a control for active import into the nuclei (Figure 1(d)–(f)).[11] Our data revealed that PB.1-558 is not enriched in the nucleus (Figure 1(g)–(i)), which contrasts with its PB.NLS-1-558 counterpart that is almost completely nuclear (Figure 1(j)–(l)).

PB variants display cytotoxicity

Prior to assaying the ability of the full-length PB, PB.1-558, and PB.NLS-1-558 to trigger transposition of an *Ifp2* source, we checked whether these proteins impacted random integration rates into HeLa cell chromosomes. We used a DNA plasmid containing a gene cassette coding for a neomycin resistance (NeoR) without the *Ifp2* sequence, the pBSK-NeoR plasmid. This was performed using a classic integration assay (see material and methods section). We observed that the integration rate of pBSK-NeoR into chromosomes was significantly lower in the presence of each of the three variants compared to a control GFP sequence (Figure 2(a); ~1.25, 1.31, 2.67 folds (1/fold change) for PB, PB.1-558, and PB.NLS-1-558, respectively). This indicated that PB and its variants have a deleterious effect on cells, including those that display random and stable integrations of NeoR into their chromosomes. The cytotoxicity of PB.NLS-1-558 is higher than that of PB and PB.1-558.

The results of integration assays (Figure 2(b)) performed with the *Ifp2*-NeoR transposon donor plasmid confirmed that PB.1-558 has a negative

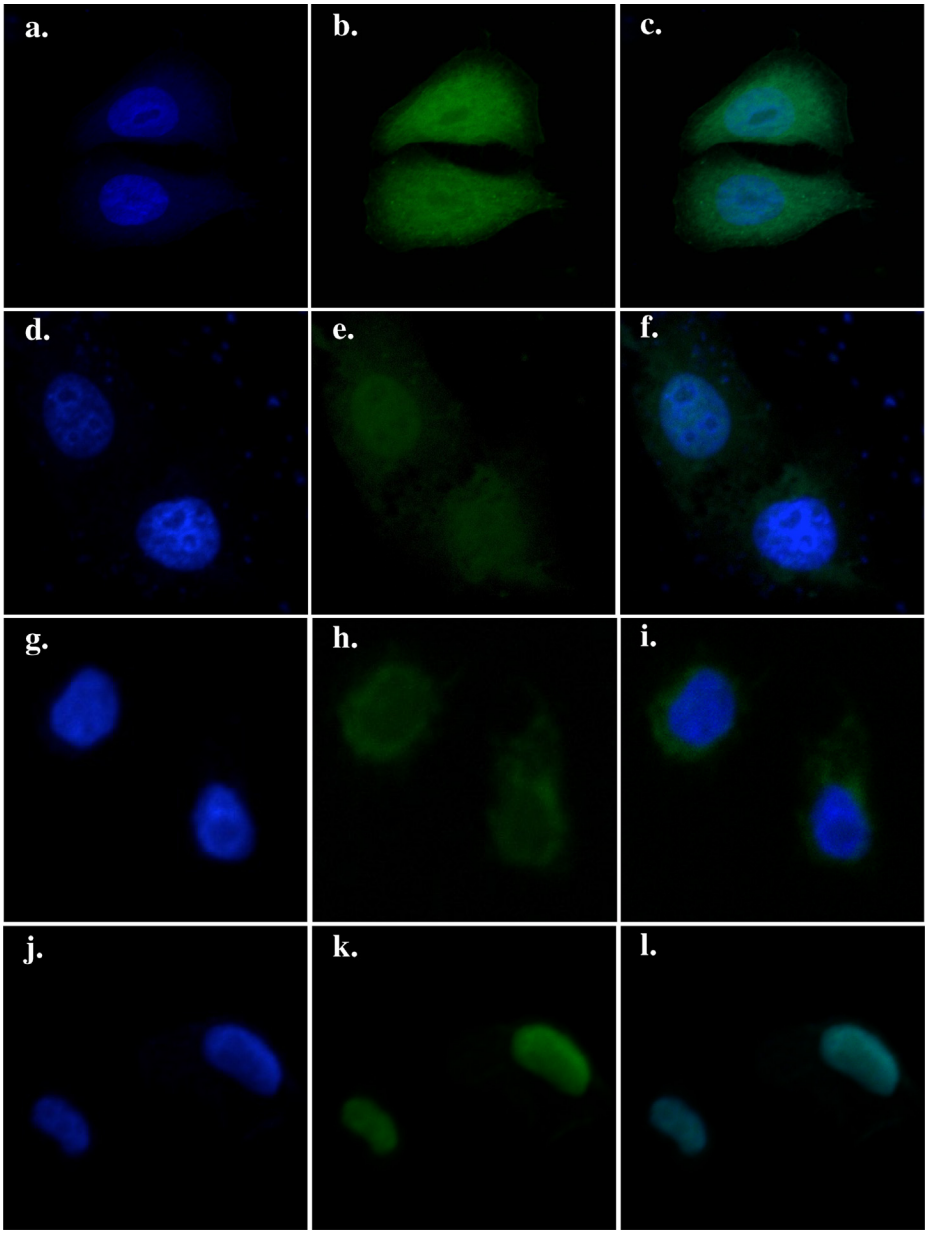


Figure 1. Cellular localization of GFP fusions in HeLa cells transiently transfected with a vector expressing GFP (a, b, c), PB-GFP (d, e, f), PB.1-558-GFP (g, h, i) and PB.NLS-1-558-GFP (j, k, l). The left panels (a, d, g, j) show GFP fluorescence, the middle panels (b, e, h, k)) show the nuclear genomic DNA staining by Hoechst 33342, the right panels (c, f, I, l) correspond to merge pictures.

effect on obtaining NeoR clones. The number of clones was 1.5 times lower than that obtained with the GFP control and 60 times lower than that obtained with PB. This indicated that integration assays performed with PB.1-558 are affected by both: (1) the cytotoxicity of PB.1-558, and (2) the integration of the NeoR cassette into chromosomes.

The number of NeoR clones obtained with PB.NLS-1-558 was 2.5 times higher than that with the GFP control. Due to the cytotoxicity of PB.NLS-1-558, this number was likely underestimated. After correcting for PB.1-558 toxicity rate (Figure 2(c)) we estimate the number of integration events to be ~7 times higher what is found with the GFP control. Overall, this means that PB.NLS-1-558 is roughly 8 times less efficient than PB for obtaining NeoR clones in an integration assay done in HeLa cells. Because cytotoxicity hampered our ability to directly evaluate integration rates and was likely dose-dependent as observed with PGBD5,[14] we focused our investigation on the ability of PB.NLS-1-558 to trigger *Ifp2* transposition by characterizing integration events into chromosomes by NeoR clones.

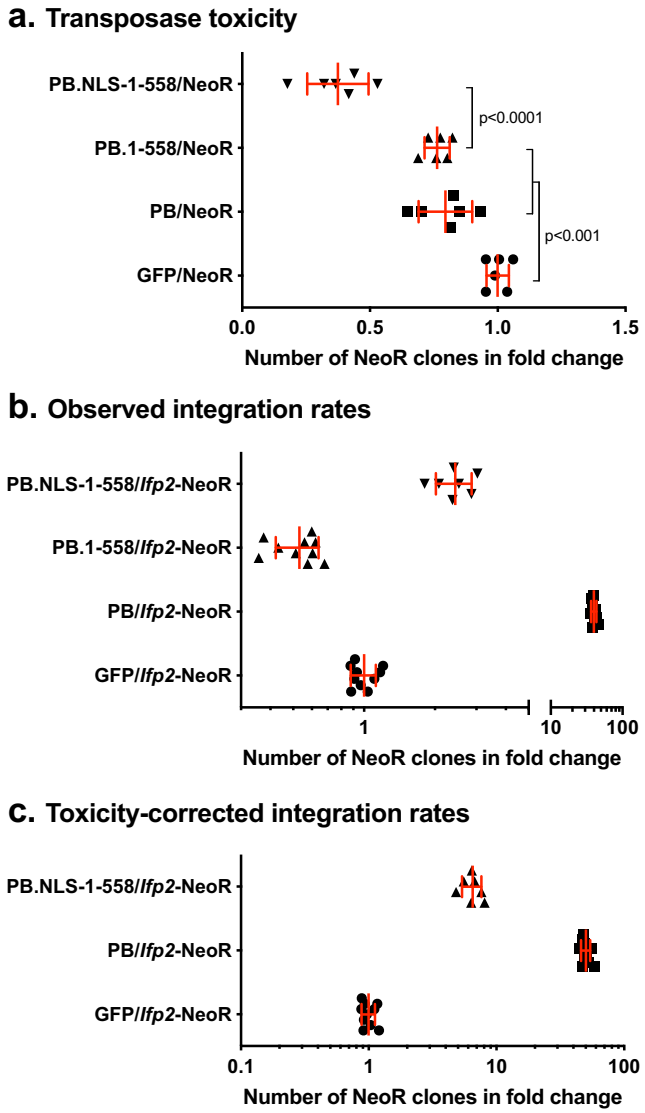


Figure 2. Box plot representations of integration assay results. (a) impact of the two PB variants (PB.1-558 and PB.NLS-1-558), PB and GFP on the rate of random integration of a NeoR cassette. (b) rates of NeoR clones resulting from the integration of *Ifp2*-NeoR when recombination was mediated by the two PB variants, PB and GFP. (c) rates of NeoR clones resulting from the integration of *Ifp2*-NeoR when recombination was mediated by PB.NLS-1-558, PB and GFP and corrected by the rate of toxicity of each protein calculated in (a). In (b) and (c), integration rates were expressed in rate NeoR clones that were normalized using controls done with GFP. In each plot, the red lines represented the median and the standard deviation, respectively.

Features of sites targeted by PB and PB.NLS-1-558 when integrating Ifp2 into chromosomes

To verify the presence of transposition events and to determine their sequence features, we produced fragment populations corresponding to *Ifp2*-chromosome junctions. These were made by LAM-PCR using genomic DNA (gDNA) of NeoR clone populations that were sequenced using Illumina Miseq technology. To prepare gDNA samples we used ~60,000 clones from integration assays done with *Ifp2*-NeoR and PB, and ~1000 clones from integration assays done with *Ifp2*-NeoR and PB.NLS-1-558. Previous results of integration assays performed in HEK293 cells found that the rate of *Ifp2* integration into chromosomes by proper transposition (i.e. with a perfect duplication of the "TTAA" TSD and conservation of the TIR sequence) was about 96–98% when PB was used as a transposase source.[15,16]

Using DNA sequence alignments, we characterized 7623 *Ifp2*/chromosome junctions resulting from integration events mediated by PB and 516 junctions mediated by PB.NLS-1-558, with *Ifp2*-NeoR as a transposon source (supplementary Tables 1(a) and 2(a)). Sequenced

junctions at the 5′ and the 3′ of *Ifp2* ends were not equally represented in the sequence data, likely because of efficiency differences at certain steps of DNA fragment amplification during the LAM-PCR. Sequence junctions were further examined taking into account the conservation of TSD and TIR sequences, two features that were required to keep the capacity of neo-inserted elements to be efficiently remobilized during excision and insertion, i.e. to remain "active in transposition".[15,16] Four kinds of junctions were observed: those displaying (i) a full TIR sequence and a TTAA TSD (red bars from positions 101 to 104 and 2222 to 2225 in Figure 3), (ii) a region containing an intact TIR and a TTAA TSD juxtaposed to a little piece of plasmid backbone (black bars from positions 1 to 100 and 2226 to 2301 in Figure 3(a) and (b)), (iii) no TTAA TSD but a full TIR sequence (blue bars from positions 102 to 105 and 2218 to 2221 in Figure 3), and (iv) no TIR sequence lacking one or several nucleotides at its outer end (black bars from positions 107 to 178 and 2147 to 2217 in Figure 3). The summary of results in Table 1 indicates that the rate of proper events when PB was used as a transposase was similar to that previously observed in other cell types, but 3.3% of junctions nevertheless displayed improper TSD, TIRs or both. Interestingly, 19.0% of junctions mediated by PB.NLS-1-558 were found to be proper, thus demonstrating that this variant is able to trigger canonical transposition events even though less efficiently than PB. Our results also indicated that PB.NLS-1-558 integrated *Ifp2* into non-canonical TSDs approximately 25 times more often than PB. Furthermore, TIRs were damaged or accompanied by a piece of backbone sequence of variable length juxtaposing the transposon in the transposon donor plasmid in about 65% of *Ifp2*/chromosome junctions (while they represented only 2.25% of junctions among integration events triggered by PB). These observations suggested that the observed junctions resulted from both proper transposition events and improper integration events that could be mediated by both PB variants, but the rates of each kind of integration events were dramatically different between the two proteins.

Next, we identified 5′ and 3′ junctions for events that occurred exactly at the same chromosomal

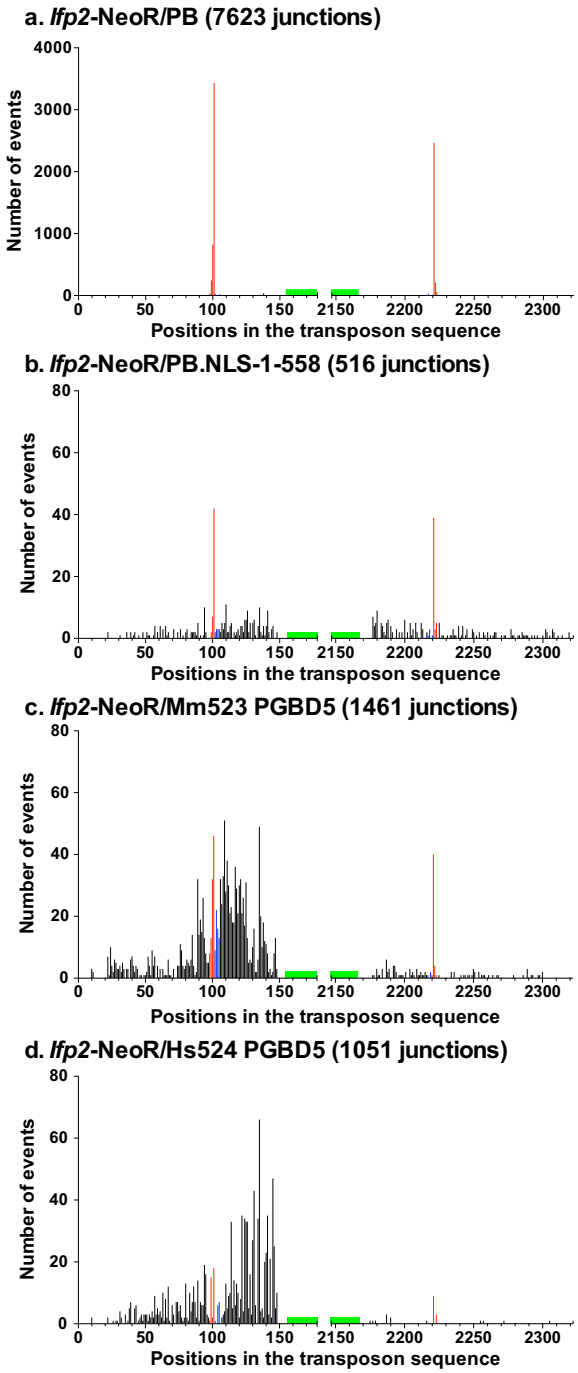


Figure 3. Number and location of transposon breakpoints in *pble* sequences after transposition into chromosomes. Histogram distributions of Ifp2-NeoR extremities transposed by PB (a), PB.NLS-1-558 (b), Mm523 (c) and Hs524 (d) (detailed in supplementary Tables 1 to 4). Red bars indicated insertion events with perfectly conserved TSD and TIR while blue bars located those in which TIR were perfectly conserved but the TSD did not correspond to a canonical TTAA at the outermost extremities of pbles. Black bars represented breakpoints within the transposon sequence and within plasmid backbone sequences juxtaposed to the transposon. Each bar corresponded to the number of junctions found at a single nucleotide position. Green boxes located the position of primers anchored within the transposon sequence and used at the last step of LAM-PCR. These graphics described the relative importance of wounds at transposon ends under our experimental conditions. However, they could not allow calculating wound rates at each of both ends due to the fact that the final LAM-PCR products in each dataset came from the gathering of several LAM-PCR reactions.

Table 1 Presence of canonic TTAA TSD and-or TIR among transposon/chromosome junctions resulting from LAM-PCR products and originating from events mediated by PB variants.

Transposase source	Transposon source	Proper TSD and TIR	Proper TSD improper TIR	Improper TSD/proper TIR	Improper TSD and TIR
PB	*Ifp2*-NeoR	96.7% (7370)	1.05% (78)	1.05% (80)	1.2% (95)
PB-NLS-1-558	*Ifp2*-NeoR	19.0% (98)	2.5% (13)	3.1% (16)	75.4% (388)

insertion sites in both datasets. We observed that 7446 chromosomal sites were used and found 177 unambiguous insertion sites in our Lumpy raw file that were occupied several times. In these 177 sites, integration events occurred in both *Ifp2* orientations when mobility was mediated by PB (supplementary Table 1(b)). A careful examination of the resulting bam file with IGV[17] revealed 11 cases of putative single integration events (supplementary Table 1(b), case highlighted in cyan blue). Among them, three corresponded to *Ifp2* transposons displaying at least one TIR damaged at its outer end, and one TIR was inserted into a duplicated TSD corresponding to a duplicated CATG motif. As previously described[18,19] we also found four sites in which both integration events occurred into non-canonical TSD (CATG, TATC, ACAT, TTCC; supplementary Table 1(b)) and 16 sites where two events occurred with the insertion of transposons with at least one improper end. These results suggest that virtually any type of non-canonical integration can be found at a very low frequency when *Ifp2* was transposed by PB. In data resulting from the transposition of *Ifp2* by PB.NLS-1-558, we found that 516 chromosomal sites were used and identified 32 unambiguous insertion sites for which integration events occurred in both orientations of the transposon. Four of these putatively corresponded to single integration events (supplementary Table 2(b), case highlighted in cyan blue, two would correspond to canonical integrations by transposition and two with non-canonical TIR or TSD). The main difference with PB is that improper events were dramatically more frequent when transposition was mediated by PB.NLS-1-558.

PGBD5 a natural domesticated CRD-deficient pble transposase

We compared the transposition features of the PB.NLS-1-558 variant to those of murine and human orthologues of the oldest domesticated *piggyBac* transposase since the origin of vertebrates, PGBD5 (Mm523 and Hs524).[20] Alignment of three protein sequences (Figure 4) revealed that both CRD-deficient proteins displayed an acidic N-terminal domain and a second domain with a basic pI (~9.2) containing an apparent catalytic triad composed of 3 acidic amino acid residues that were essential for transposition activity.[20] Another shared feature was their ability to trigger *Ifp2* transposition.[14,21] This transposition ability was rather unexpected for PGBD5 compared to PB.NLS-1-558 because the PGDB5 catalytic triad was not located at the same positions as in *pble* transposases (Figure 3, bold residues highlighted in yellow). PGBD5 acquired a new putative NLS that is centrally located in the sequences of Mm523 and Hs524 (Figure 3, RKRKKRK motif typed in green and underlined). In agreement with the literature[14] we observed that the ectopic expression of murine PGBD5 isoform of 523 amino acids (Mm523) reduced the apparent efficiency of obtaining NeoR clones (Figure 5(a)) that is close to that of PB.NLS-1-558 in HeLa cells (Figure 2(a)). In integration assays done with the *Ifp2*-NeoR transposon donor plasmid under experimental conditions similar to those used above for PB.NLS-1-558, the rate of NeoR clones obtained with Mm523 (Figure 5(b)) was similar to that obtained with the GFP control. In order to verify whether this was due to PGBD5 cytotoxicity, we used a second cellular system developed in human rhabdoid tumor G401 cells and in which the endogenous expression of PGBD5 (Hs524) was found to have little impact on cell viability.[14] Under these experimental conditions, we found that the rate of NeoR clones was sevenfold higher than that of the GFP control (Figure 5(b)). Together, this indicates that the expression rate of CRD-deficient *pble* transposases strongly impact the outcome of integration assays.

The sequence features of integration events were studied through *Ifp2*-chromosome junctions obtained with Mm523 and Hs524. We prepared gDNA samples from ~1800 and 1600 NeoR clones obtained from integration assays done with *Ifp2*-NeoR and, Mm523 or Hs524, respectively. Using the Mm523 gDNA sample, we obtained 1461 transposon/chromosome junctions that were analyzed as described above (supplementary Table 3(a)). The profiles of transposon/chromosome junctions were found to be similar between integration events mediated by PB.NLS-1-558 and Mm523 (Figure 3(c), and Table 2 versus last raw in Table 1). This was also verified by examining chromosomal sites where we found integration events in both orientations within the 1461 chromosomal sites used (supplementary Table 3(a) and (b)). When the junctions were categorized and analyzed in terms of percentages at each *Ifp2* end for PB, PB.NLS-1-558 and Mm523 we observed that: (i) proper junctions occurred more often at the 3′ end than at the 5′ end (Figure 6, red bars), and (ii) among improper

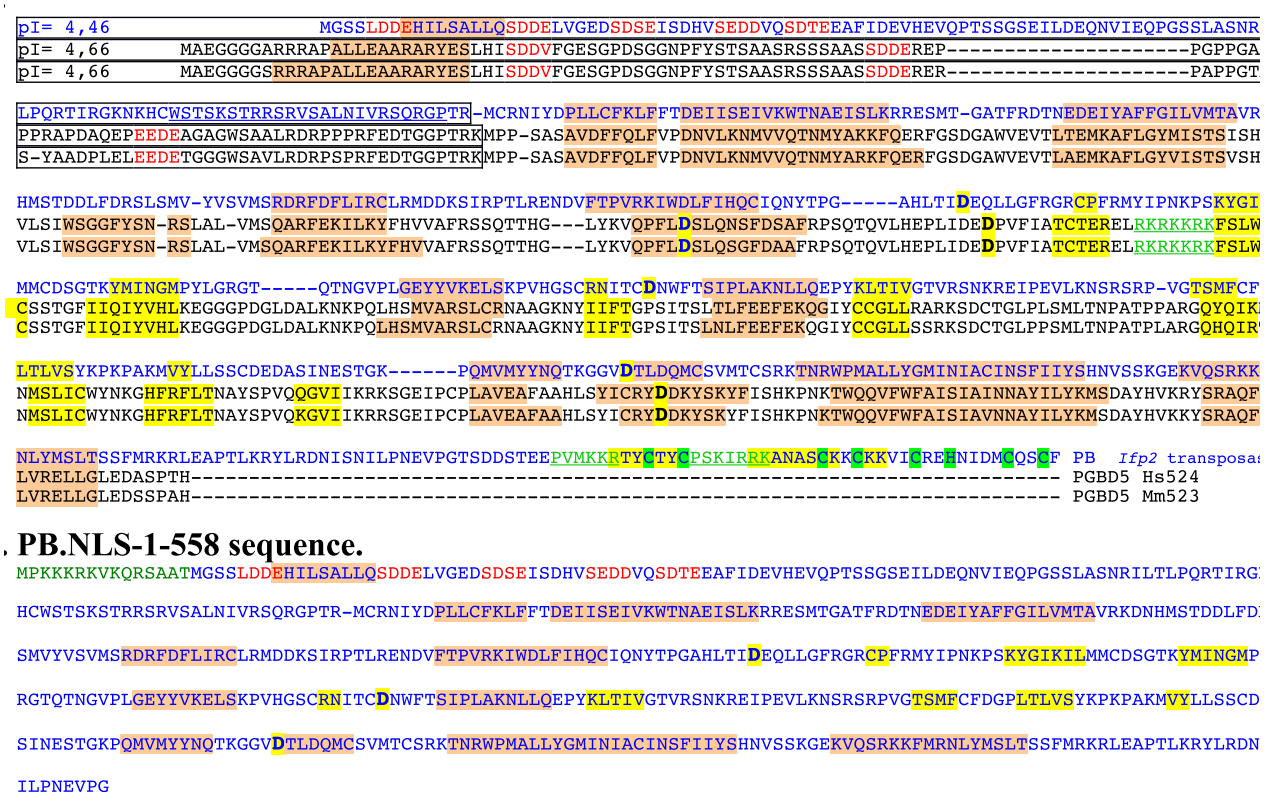


Figure 4. Sequence features of the *Ifp2* transposase (PB) variants and two Mm523-like PGBD5 isoforms. (a) Protein sequence alignment of Ifp2 transposase (PB) with two murine and human domesticated PGBD5 proteins corresponding to the orthologous Hs524 and Mm523 isoforms. (b) Sequence features of PB.NLS-1-558. Secondary structure predictions calculated with psipred (http://bioinf.cs.ucl.ac.uk/psipred/) and Jpred4 (http://www.compbio.dundee.ac.uk/jpred/) were highlighted in pink for α-helices and in orange for β-strands. The three proteins share two domains: a N-terminal domain that was few structured, with an acid pI (boxed regions) and repeated acid motifs (in red letters), a domain of ~400 amino acid residues that display a basic pI. PB contained a third C-terminal domain, the CRD, that contains cysteins (highlighted in green) able to assemble zinc finger folds. Aspartic residues inactivating the recombinase catalytic activity were bolded and highlighted in yellow.[17,21] The PB NLS and the putative NLS in PGBD5 isoforms were underlined and typed in green.

junctions, those without a canonical TSD and those located within TIR and juxtaposed with transposon sequences (Figure 6, blue bars and internal black bars, i.e. wounds at transposition ends as exemplified in 22 occurred more often than those located within the plasmid backbone sequences juxtaposed near the TSD and TIR of the donor plasmid (Figure 3, flanking black bars). Using the Hs524 gDNA sample, we obtained 1051 transposon/chromosome junctions (supplementary Table 4(a)). The junction profile was overall similar to those of both CRD-deficient proteins (Figures 3(c) and 6 (d)), but it displayed a marked difference in that there were fourfold and twofold less proper insertion events by transposition than in those obtained with PB.NLS-1-558 and Mm523, respectively. Unexpectedly, this suggests that PGBD5 is more prone to trigger improper integration in rhabdoid tumor G401 cells, consistent with the proposal that PGBD5 exhibits aberrant activities in human rhabdoid tumors.[14] Since we observed that PB.NLS-1-558, Mm523 and Hs524 display similar junction profile, we wondered if their insertion site preferences might be similar.

Features of insertion sites targeted by PB variants and PGBD5 when integrating Ifp2 into chromosomes

PB and PGBD5 have been shown to integrate *Ifp2* into intragenic regions more frequently than expected by chance, specifically within transcription start site (TSS) regions flanking (±5 kbp) protein-coding genes.[23–26] Using our junction data, we observed that PB, PB.NLS-1-558, Mm523, Hs524 did not distributed *Ifp2* integrations at random into intergenic and intragenic regions (Figure 7(a); Chi2, $p = 2.08 \times 10^{-95}$, 2.21×10^{-10}, 0.0026, and 3.72×10^{-81} respectively), but with a significant enrichment for intragenic regions (hypergeometric test, $p \ll 0.01$ for the four proteins). Similar investigations were also

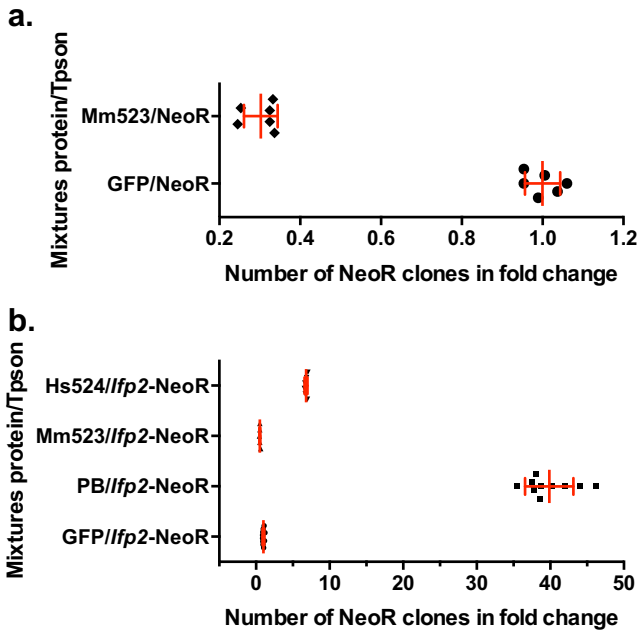


Figure 5. Graphic representations of integration assay results. (a) Impact of Mm523 on the rate of random integration of a NeoR cassette. (b) Rates of integration of an *Ifp2*-NeoR when recombination was mediated by PB and Mm523 in HeLa cells, and Hs524 in G401 cells. Integration rates were expressed in rate NeoR clones that were normalized using controls done with GFP (green). In each plot, the red lines represented the median and the standard deviation, respectively.

done within regions flanking TSSs of five types of genes coding for: (i) proteins, (ii) non-coding RNA (ncRNA), (iii) micro RNA (miRNA), or being annotated in hg38 as (iv) pseudogenes or (v) uncharacterized genes. We also found that *Ifp2* was integrated more frequently than expected by chance within regions flanking TSSs in the 5 types of genes, except for the two CRD-deficient proteins into uncharacterized genes (Figure 7(b); Chi2, $p = 2.58 \times 10^{-9}$, 2.16×10^{-8}, 8.27×10^{-12}, and 3.43×10^{-9} respectively), and with a significant enrichment in each type of genes (hypergeometric test, $p \ll 0.01$ for the four proteins), except for the miRNA and ncRNA genes in the PB.NLS-1-558 and Mm523 datasets, respectively (hypergeometric test, $p = 0.043$ and 0.051). In addition to these global distribution features, a striking feature was that our Lumpy raw files, after manual investigation using IGV, contained 166 (i.e. 177-11), 24 (44-20), 23 (26-3) and 11 (13-2) chromosomal sites, each displaying a fragment containing the *Ifp2* element inserted in both orientations, i.e. inserted at least twice into these sites when integration events were mediated by PB, PB.NLS-1-558, Mm523 and Hs524, respectively. We also found common insertions sites among the PB, PB.NLS-1-558, Mm523 and Hs524 datasets (Table 3, lines 1, 3 and 5). These insertion events occurred at the same nucleotide position site but this was not due to sample contamination since they resulted from *Ifp2* integration events in different orientations and in some cases from properly and improperly integrated *Ifp2* transposons. The number of these observations was increased when using a 1000 bp window on both sides at each chromosomal insertion site (Table 3, lines 2, 4 and 6; regions called below insertion sites-containing regions (ISCR)) (Figure 8).

The choice of an insertion site by any *pble* transposase does not fully occur at random since a TTAA motif is used. In the human genome model hg38, there are 18,713,270 TTAA motifs. Public data about DNAse I hypersensitivity mapping revealed that 98% of them are located in open chromatin in HeLa cells (but also in HEK cells), i.e. accessible to DNA binding proteins such as transposases. This means that the probability of integrating an *Ifp2* transposon twice into a single target site lies about 1.8×10^{-7} in hg38. Taking into account the size of datasets used herein, to find several insertions by chance into the same site is therefore unexpected under our experimental conditions.

Given the putative impact of some specific genomic features of HeLa cells such as their aneuploidy,[27–30] we further investigated ISCR features in other cell lines taking advantage of public datasets. We used three of them that were produced from integration assays performed with *Ifp2* and PB in HEK293,[23] in HCT116[15] and in CD4+[26] cells (21,967, 172,866 and 8954 chromosomal sites, respectively (Table S4(a)–(c)).

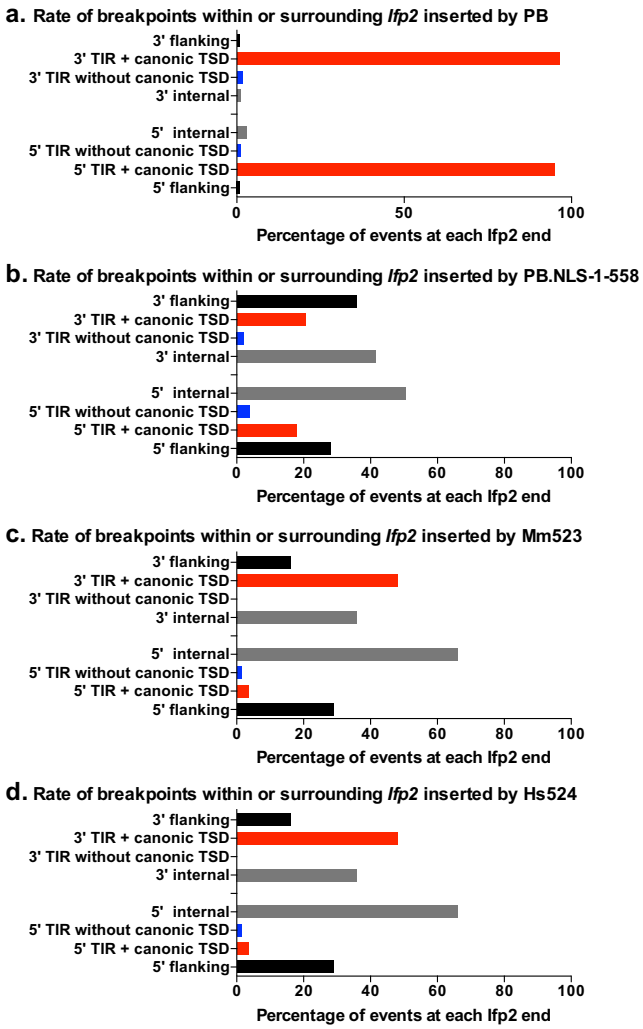


Figure 6. Features of transposon breakpoints in lfp2 transposed by PB (a), PB.NLS-1-558 (b), Mm523 (c) and Hs524 (d). Black bars corresponded to percentages of breakpoints located within the plasmid backbone flanking the transposon or those located within inner transposon regions (from the position 2 in TIR to the primer used for the LAM-PCR). Red bars corresponded to those within transposons that displayed intact canonical TTAA TSD and TIRs. Blue bars corresponded to *pbles* displaying noncanonical TSD but intact TIRs.

First, we confirmed that the rate of insertions mediated by PB into intragenic regions in each of the four cell lines (HeLa, HEK293, HCT116 and CD4+; Table 4(a), column 4) is 7 to 18% higher than expected by chance (51.6%; Table 4(b), column 1). This preference could not be explained by the numbers of TTAA motifs in intragenic regions (53.1%; Table 3(c), column 2), which is close to that expected by chance. In spite of variations of aneuploidy and chromatin profiles between the four cell lines, the insertion preference into intragenic regions does not appear to correlate to these features.

In order to verify the statistical consistency of insertion sites shared between datasets, all pairs of datasets were compared taking into account the variation of TTAA motif distribution between intra and intergenic regions (Table 4(c)). P-values indicated that the number of commonly used chromosomal insertion sites was significantly more elevated than expected by chance (Table 3; raws 1–16) whatever the window used around the insertion sites (0 or 1000 bp). We also observed that 18 ISCR were shared by the four datasets obtained with *lfp2* and PB in the four cell lines.

We also examined insertion datasets obtained with the transposon *sleeping beauty* in HEK293 and CD4+ cells (28,490 and 8290 insertion sites, respectively.[15,26] Taking into account the distribution of its TA targets in hg38 (Table 4(c)), results in Table 3 revealed that this transposon does not display a significant preference between available putative TA target sites. It displayed higher rates of insertion into intragenic regions (~62.5%, Table 4

Table 2 Presence of canonic TTAA TSD and-or TIR among transposon/chromosome junctions resulting from LAM-PCR products and originating from events mediated by Mm523 and Hs524.

Transposase source	Transposon source	Proper TSD and TIR	Proper TSD improper TIR	Improper TSD/proper TIR	Improper TSD and TIR
Mm523 PGBD5	*Ifp2*-NeoR	10.0% (147)	5.4% (80)	4.3% (64)	80.3% (1188)
Hs524 PGBD5	*Ifp2*-NeoR	4.6% (48)	2.1% (22)	1.3% (14)	92.0% (967)

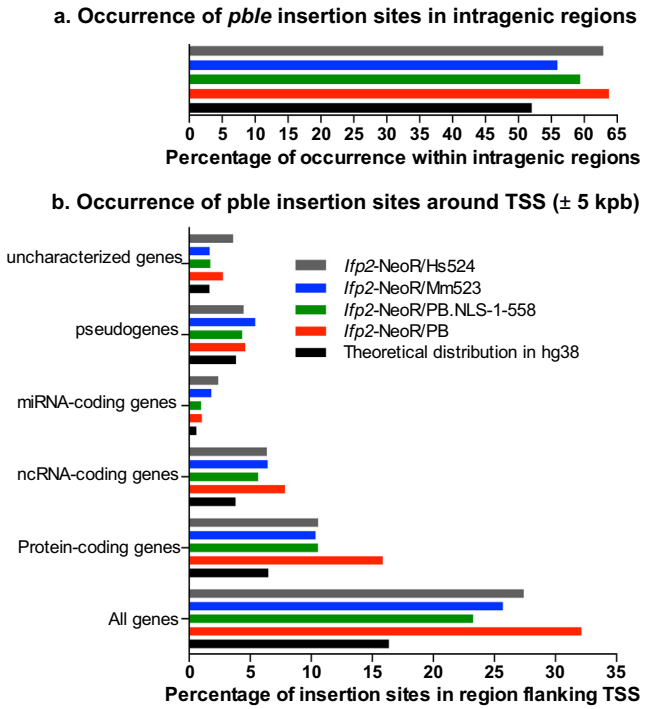


Figure 7. Proportions of Ifp2 insertions mediated by PB, PB.NLS-1-558**, Mm523 and Hs524 in intragenic regions (a) and regions containing TSS (b)** taking into account the five gene categories: protein-coding genes, ncRNA-coding genes, miRNA-coding genes, pseudogenes and uncharacterized genes. Black bars indicated the expect percentage in a random distribution.

(a)) than predicted by chance (51.6%, Table 4(b)). In contrast to PB, this can however be correlated with TA density that is dramatically increased in these regions (60.4%, Table 4(c)) compared to intergenic ones.

In all, our data reveal that PB, PB.1-558, Mm523 and Hs524 insert *Ifp2* preferentially into intragenic regions with some level of site preference between available TTAA target sites.

Features of genes targeted by Ifp2 insertions mediated by PB variants and PGBD5

We wondered whether insertion site preferences of *pble* transposases might also be seen at the level of some intragenic regions. We postulated that experimental conditions of transposition assays are conducive to forced integration of transposons into chromosomes. Therefore, we predicted that insertions should be enriched among ISCR shared by several datasets than among those unique to each dataset.

In the ontology analysis done with the 410 genes overlapped by ISCR and shared by at least two datasets among those obtained with PB, PB.NLS-1-558, Mm523 and Hs524 in HeLa cells (Figure 6 (a)), we found that 35/50 significant terms were directly related to the nervous system (Table 5). This issue was therefore further investigated by verifying whether there was an enrichment of “neuron genes” among ISCR overlapping with genes. For this purpose, we used the 6854 “neuron genes” identified in hg19 based on the expression properties of 29,165 genes (protein-, ncRNA- and miRNA-coding plus some uncharacterized genes and pseudogenes) in 216 distinct human brain structures.[31] We found that neuronally expressed genes were significantly enriched in each of the PB, PB.NLS-1-558, Mm523 and Hs524 datasets obtained in HeLa cells (Table 4(a), columns 5, 6, 7). They were again

Table 3 Number of chromosomal sites in common in each dataset pair and their statistical consitency in permutation tests.

Feature of dataset 1			Feature of dataset 2					Random permutations features			
Cell line 1	Transposon 1	Transposase 1	Cell line 2	Transposon 2	Transposase 2	Window around insertion sites (±) in both datasets	Number of sites observed	Average number of sites expected per chance	Standard deviation	Z score	Probability = H0 was no differences between obs. and exp.
HeLa	*Ifp2*	PB	HeLa	*Ifp2*	PB.NLS-1-558	0	1	N.A.	N.A.	N.A.	$p < 0.001$
HeLa	*Ifp2*	PB	HeLa	*Ifp2*	PB.NLS-1-558	1000	37	N.A.	N.A.	N.A.	$p < 0.001$
HeLa	*Ifp2*	PB	HeLa	*Ifp2*	Mm523-PGBD5	0	83	N.A.	N.A.	N.A.	$p < 0.001$
HeLa	*Ifp2*	PB	HeLa	*Ifp2*	Mm523-PGBD5	1000	175	N.A.	N.A.	N.A.	$p < 0.001$
HeLa	*Ifp2*	PB	G401	*Ifp2*	Hs524-PGBD5	0	24	N.A.	N.A.	N.A.	$p < 0.001$
HeLa	*Ifp2*	PB	G401	*Ifp2*	Hs524-PGBD5	1000	42	N.A.	N.A.	N.A.	$p < 0.001$
HeLa	*Ifp2*	PB	HEK293	*Ifp2*	PB	0	21	N.A.	N.A.	N.A.	$p < 0.001$
HeLa	*Ifp2*	PB	HEK293	*Ifp2*	PB	1000	471	260.162	17.0215	12.3865	0
HeLa	*Ifp2*	PB	HCT116	*Ifp2*	PB	0	238	67.01	8.4329	20.2763	0
HeLa	*Ifp2*	PB	HCT116	*Ifp2*	PB	1000	2387	N.A.	N.A.	N.A.	$p < 0.006$
HeLa	*Ifp2*	PB	CD4+	*Ifp2*	PB	0	24	N.A.	N.A.	N.A.	$p < 0.001$
HeLa	*Ifp2*	PB	CD4-	*Ifp2*	PB	1000	213	97.684	10.2987	11.197	0
HEK293	*Ifp2*	PB	HCT116	*Ifp2*	PB	0	391	N.A.	N.A.	N.A.	$p < 0.001$
HEK293	*Ifp2*	PB	HCT116	*Ifp2*	PB	1000	6600	5419.484	93.3269	12.0060	0
HEK293	*Ifp2*	PB	CD4+	*Ifp2*	PB	0	362	N.A.	N.A.	N.A.	$p < 0.001$
HEK293	*Ifp2*	PB	CD4+	*Ifp2*	PB	1000	752	N.A.	N.A.	N.A.	$p < 0.001$
HCT116	*Ifp2*	PB	CD4+	*Ifp2*	PB	0	16	81.825	9.0919	17.5073	0
HCT116	*Ifp2*	PB	CD4+	*Ifp2*	PB	1000	5235	2213.501	50.9891	59.2576	0
HEK293	*Sleeping Beauty*	SB	CD4+	*Sleeping Beauty*	SB	0	3	N.A.	N.A.	N.A.	$p > 0.083$
HEK293	*Sleeping Beauty*	SB	CD4+	*Sleeping Beauty*	SB	1000	385	349.967	19.09	1.8350	$p = 0.034$

N.A., not appropriated to use a Z-test as the distribution of the 1000 permutations results did not fulfil normality in a Shapiro-Wilk test. In the most right column is typed in red probabilities supporting that there was no difference at a threshold of 0.01.

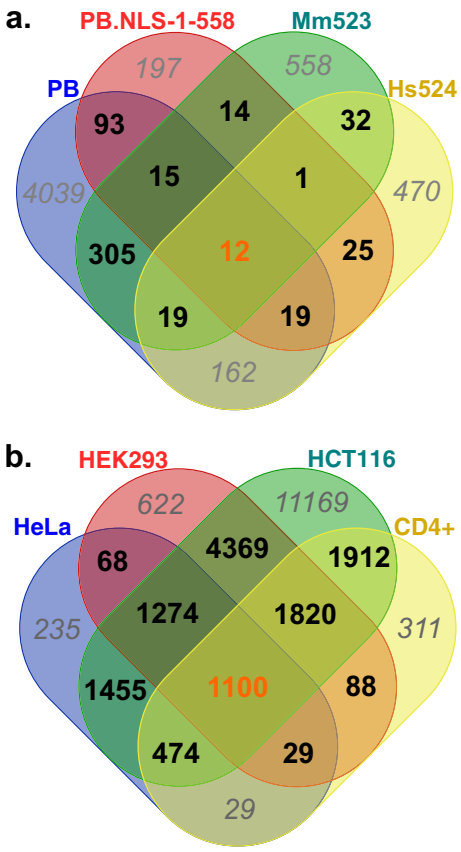


Figure 8. Venn diagram representations of intragenic regions overlapped by insertion site-containing regions (±1000 bp) between transposition assays done with PB, PB.NLS-1-558, Mm523 and Hs524 (a) and with PB in HeLa, HEK293, HCT116 and CD4+ cells (b). The numbers of intragenic regions specific of datasets were italicized and typed in grey. Those shared by all datasets were typed in orange.

enriched among the 697 genes overlapped by ISCR that were shared by at least two of four datasets. This enrichment in neuron genes was 31.7–38.8% in each of the four datasets and 37.2% (259/697) among the 697 shared genes. These results therefore support the notion that neuron genes are preferred regions for *pble* transposases to insert *Ifp2*.

These observations were confirmed using datasets obtained in HEK293, HCT116 and CD4+ cells. First, we found that neuronally expressed genes were significantly enriched in each of the four datasets (Table 4(a), columns 5, 6, 7). This did not appear to be related to the target density since the percentage of ISCR in those genes was found to be 3–8% more elevated than the rate of TTAA motifs in those genes (Table 4(a), column 4 versus Table 4(c), column 3). We found that genes overlapped by ISCR and shared by at least two of four datasets did not display a significant enrichment in neuron genes (3884/12618; 30.8%; Figure 6(b)). However, a very strong enrichment was found when only ISCR shared by the four datasets were kept for the analysis (705/1100, 64% neuron genes) and a strong depletion in neuron genes was found among genes occurring in only one dataset (2273/12186; 18.7%).

Finally, we evaluated whether the insertion preferences into neuronally expressed genes were specific to PB by analysing the same features in datasets obtained with two unrelated transposons.[15,26] Data obtained with *sleeping beauty* in HCT116 and CD4 cells and with a *TcBuster* in HCT116 cells indicated that both transposons also inserted more frequently into neuronally expressed genes than in other genes (Table 4(a), column 4, 5, 6). Their insertion preferences were also about 6–8% above the density in their respective target motif (Table 4(c), column 3). For *sleeping beauty*, we also found that there was an enrichment in neuronally expressed genes among genes that overlapped by ISCR and shared by both datasets (1143 neuron genes for 2813 genes; 40.63%) and a depletion in those which were only found in one dataset (2625/9791; 26.8%).

Altogether, these last results reveal that *pble* transposases insert *Ifp2* more often than expected by chance into neuronally expressed genes. However, this apparent preference is also displayed by *sleeping beauty* and very likely by *TcBuster,* indicating that it is not specific of *pble* transposases. This is not related to the gene size and the number of TTAA and TA because the densities in target motifs are very close in neuron and non-neuron genes (TTAA: 5.318 ± 0.029 and 5.114 ± 0.015 motifs/kbp, respectively; TA: 44.68 ± 0.156 and 44.07 ± 0.083 motifs/kbp, respectively). Therefore, this might result from the enhanced accessibility of neuronally expressed genes or their association with DNA repair and chromatin remodeling factors that support DNA transposition, a property that would be shared by multiple cell lines as exemplified here.

Discussion

This study generated two sets of novel insights. First, the two CRD-deficient transposases PB.NLS-1-558 and PGBD5 (Mm523 and Hs524) mediate canonical *Ifp2* transposition, but also non-canonical events that may result from events of improper transposition, transposase-dependent integration by recombination and random integration. We assume that these CRD-deficient transposases operate at a reduced efficiency than the "wild-type" or "full-length" *piggyBac* transposase but their cytotoxicity on host cells and the possibility that they do integration by transposase-dependent recombination indicated that they have nuclease activity, as previously suggested for PGBD5 isoforms.[13] Furthermore, cytotoxicity issues often arise from the balance between the level of expressed protein and the efficiency of mechanisms responsible for the mainte-

Table 4

(a) Rate of regions containing insertion sites in genes and neuron genes in hg38

1, Dataset features: cells, transposon/transposase sources	2, No of different ISCR*	3, No &% of different ISCR overlapping a gene***	4, No of genes overlapped by ISCR**	5, No &% of different ISCR overlapping a neuron gene***	6, No &% of neuron genes among genes overlapped by ISCR**	7, p****
HeLa, *Ifp2*/PB	7,623	5,060 (66.4%)	4,663	2,231 (29.2%)	1,587 (34.0%)	2.1×10^{-82}
HeLa, *Ifp2*/PB.NLS-1-558	516	327 (63.2%)	376	123 (23.8%)	123 (37.6%)	2.5×10^{-6}
HeLa, *Ifp2*/Mm523	1,479	863 (58.4%)	956	321 (21.7%)	303 (31.7%)	3.1×10^{-11}
G401, *Ifp2*/Hs524	1052	665 (63.2%)	740	279 (26.5%)	258 (38.8%)	1.3×10^{-14}
HEK293, *Ifp2*/PB[22]	21,967	15,226 (69.3%)	9,370	7,173 (32.6%)	3,032 (32.4%)	1.9×10^{-150}
HCT116, *Ifp2*/PB[14]	172,866	113,648 (67.7%)	23,578	49,688 (28.8%)	5,945 (25.3%)	1.3×10^{-88}
CD4+, *Ifp2*/PB[25]	8,954	6,940 (77.5%)	5,763	3,134 (35.0%)	1,903 (33.0%)	6.2×10^{-89}
HCT116, *sleeping beauty*/SB[14]	28,490	17,534 (61.54%)	10,460	7,776 (27.3%)	3,229 (30.9%)	2.5×10^{-128}
CD4+, *sleeping beauty*/SB[25]	8,290	5,441 (64,63%)	5,133	2,406 (29.0%)	1,750 (34.1%)	2.9×10^{-93}
HCT116, *TcBuster*/ TCBUSTER[14]	17,227	11,841 (68.7%)	8,522	5,517 (32.0%)	2,820 (33.1%)	4.9×10^{-151}

(b) Numbers and sequence coverages (%) of genes and neuron genes in hg38

1, Genes	2, Neuron genes
30,077 (51.6%)	6854 (22.8%)

(c) Numbers and rate of potential insertion sites for PB, SB and TcBuster in hg38

	1, Number of potential insertion sites in hg38	2, Number & % of potential insertion sites in genes in hg38	3, Number & % of potential insertion sites in neuron genes in hg38
Ifp2 (TTAA)	18,713,270	9,943,117 (53.1%)	4,521,501 (24.2%)
sleeping beauty (TA)	152,412,514	92,065,399 (60.4%)	36,803,438 (24,1%)
TcBuster (NNNTANNN)	130,938	74,733 (57.1%)	33,940 (25.9%)

* ISCR = insertion sites-containing regions, i.e. plus 1000 bp upstream and downstream each insertion site.
** 2 genes can overlap an ISCR.
*** Each genes and neurons genes = exons and introns of their transcriptional unit plus 5000 bp upstream and downstream.
**** Hypergeometric test using H0 = no enrichment in neuron genes among genes overlapped by an ISCR.

Table 5 Significant ontology terms resulting from the analysis of the 817 genes shared by PB, PB.NLS-1-558, Mm523 and Hs524.

GO ID	GO Term	P value*	Nr. Genes
Fused GO terms: kinase activity (10.62% genes)			
GO:0016301	Kinase activity	0.0042217	80
Fused GO terms: establishment of cell polarity (2.26% genes)			
GO:0030010	Establishment of cell polarity	0.0043640	17
Fused GO terms: cranial nerve morphogenesis (2.92% genes)			
GO:0021602	*Cranial nerve morphogenesis*	0.0003926	9
GO:0021953	*Central nervous system neuron differentiation*	0.0302390	19
Fused GO terms: regulation of signal transduction (25.76% genes)			
GO:0023051	Regulation of signaling	0.0080511	161
GO:0048583	Regulation of response to stimulus	0.0132815	183
GO:0010646	Regulation of cell communication	0.0058346	160
GO:0009966	Regulation of signal transduction	0.0016884	145
Fused GO terms: synapse organization (6,51% genes)			
GO:0034330	*Cell junction organization*	0.0013124	49
GO:0034329	*Cell junction assembly*	0.0237032	32
GO:0050808	*Synapse organization*	0.0004306	36
GO:0099173	*Postsynapse organization*	0.0053155	19
GO:0099084	*Postsynaptic specialization organization*	0.0137634	8
Fused GO terms: regulation of small GTPase mediated signal transduction (14.61% genes)			
GO:0044093	Positive regulation of molecular function	0.0181959	92
GO:0007264	Small GTPase mediated signal transduction	0.0043479	40
GO:0008047	Enzyme activator activity	0.0134638	37
GO:0043087	Regulation of GTPase activity	0.0009387	37
GO:0060589	Nucleoside-triphosphatase regulator activity	0.0007998	30
GO:0030695	GTPase regulator activity	0.0007258	28
GO:0043547	Positive regulation of GTPase activity	0.0225952	30
GO:0051056	Regulation of small GTPase mediated signal transduction	0.0001020	31
GO:0005096	GTPase activator activity	0.0008977	26
Fused GO terms:nervous system development (37.32% genes)			
GO:0007275	*Multicellular organism development*	0.0000272	238
GO:0009653	*Anatomical structure morphogenesis*	0.0000158	139
GO:0060322	*Head development*	0.0024237	53
GO:0000902	*Cell morphogenesis*	0.0001483	67
GO:0032989	*Cellular component morphogenesis*	0.0017502	52
GO:0048468	*Cell development*	0.0067125	107
GO:0048731	*System development*	0.0005126	212
GO:0051128	*Regulation of cellular component organization*	0.0203388	114
GO:2000026	*Regulation of multicellular organismal development*	0.0489716	100
GO:0048513	*Animal organ development*	0.0350042	158
GO:0000904	*Cell morphogenesis involved in differentiation*	0.0047322	49
GO:0007399	*Nervous system development*	0.0000000	142
GO:0007417	*Central nervous system development*	0.0000264	69
GO:0022008	*Neurogenesis*	0.0000022	99
GO:0051960	*Regulation of nervous system development*	0.0018437	59
GO:0120036	*Plasma membrane bounded cell projection organization*	0.0000839	88
GO:0007420	*Brain development*	0.0059673	50
GO:0120035	*Regulation of plasma membrane bounded cell projection organization*	0.0233900	44
GO:0048699	*Generation of neurons*	0.0000007	96
GO:0120039	*Plasma membrane bounded cell projection morphogenesis*	0.0006627	48
GO:0030182	*Neuron differentiation*	0.0000003	90
GO:0048666	*Neuron development*	0.0001405	71
GO:0045664	*Regulation of neuron differentiation*	0.0477549	42
GO:0031175	*Neuron projection development*	0.0003100	64
GO:0048667	*Cell morphogenesis involved in neuron differentiation*	0.0016870	43
GO:0016358	*Dendrite development*	0.0431617	22
GO:0061564	*Axon development*	0.0137671	37
GO:0007409	*Axonogenesis*	0.0105043	35

* Term PValue corrected with Bonferroni step down. In italic are indicated GO terms related to neurogenesis and neuron.

nance of genome integrity, which varies widely from one cell line to another. Such effects could also be related to the cell cycle. This might explain why PB.NLS-1-558, which was always located in the nucleus, had a more negative effect than PB.1-558 (Figure 2(a)), which was mainly cytoplasmic during most of the cell cycle and in contact with chromosomes only during the cell division phase.

The second insight from this work concerns the ability of PB and both CRD-deficient *pble* transposases to trigger integration events that did not seem to occur at random into chromosomes. However, we cannot assume strict insertion site specificity of *pble* transposases since only a small part of the observed insertions events in datasets are the same, down to the same nucleotide position. However, our data demonstrated that these transposases displayed real preferences for insertion into regions containing genes and frequently close to their TSS. In addition, our results supported that *pble* transposases frequently targeted their *pble* insertions into genes committed to the central nervous system function. This last point will need further experimental confirmation. Indeed, the lengths of genes involved in nervous system function were, on average, longer than those of other genes (for review see 32. An alternative interpretation might be that the insertion sites we found were preferentially located in neuronally expressed genes due to their size. However, our results take into account the genome coverage and the density in TTAA motif and support that the observed insertion preferences are not related these factors.

Contribution to the understanding of piggyBac transposition

Previous studies suggested two roles for the CRD of PB. The CRD might be an essential component of DNA-binding to the ends of *Ifp2*, mandatory for transposition.[9] This was confirmed in another study, which also indicated that the CRD was essential for the assembly of the transposase dimer, the active oligomer form for transposition.[10] Here we demonstrate that two CRD-deficient *piggyBac* transposases were able to trigger proper *pble* transposition. Therefore, the CRD is not essential for transposition but seems necessary for triggering proper transposition events, probably by driving precise DNA cleavages at *pble* ends and directing a strict choice of TSD.

Evolutionary reasons for domesticating a CRD-deficient pble transposase

These results, as well as previously published data,[14,20,21] have highlighted two properties of PB, and perhaps of other pble transposases, that may have previously been underestimated in the context of transposase-coding gene domestication. First, while PB mostly mediates proper Ifp2 transposition, it is also sometimes responsible for improper transposition that leads to neo-inserted elements that are difficult or impossible to re-mobilise during new rounds of transposition. Second, pble transposases might display strong preferences for insertion and genome rearrangements. Indeed, we noted that PGBD5 is highly expressed in the mammalian nervous system[20] and that current publicly available data (https://www.gtexportal.org/home/gene/PGBD5, https://www.proteinatlas.org/search/PGBD5) widely support this conclusion. However, these data concern mRNA expression and data regarding mRNA translation and protein expression will be required. Nevertheless, if acquiring a mechanism for triggering irreversible DNA rearrangements in the early steps of vertebrate evolution in the nervous system can be considered advantageous, then PGBD5 seems to possess all the properties required to play such a role. The evolutionary history of the RAG1/RAG2 proteins[33] suggests that each time a domesticated transposase has emerged during evolution its domestication was concurrent with the domestication of its transposon targets. In the PGBD5 context, verifying that pbles are domesticated and are used as binding targets by PGBD5 for genome rearrangements will be challenging. Indeed, while PGBD5 is a highly conserved protein in vertebrates, the pble landscape in these genomes varies drastically from one host species to the next. In the human genome three pbles unrelated to PGBD5 are annotated: MER75, MER85 and Looper. In the mouse genome only one pble closely related to human *Looper* is annotated. In the zebrafish seven pbles that were not related to PGBD5 and to those present in human and mouse genomes, have been annotated. In the chicken genome no pble has been found so far.[34,35] It is possible that PGBD5 has been domesticated in order to mobilise multiple *pbles* for recombination. Its protein sequence conservation in chicken and the absence of *pble*s in this species suggests that it binds to other DNA binding targets, which may be related to the PGBD5-specific signal (PSS) sequences observed in human rhabdoid tumors.[14]

Materials and Methods

cDNA cloning of PGBD5 murine isoforms

A single mouse brain (strain C57Bl6) was used for total RNA extraction using Tri-reagent (Sigma-Aldrich, St-Louis, MO, USA). cDNA synthesis was carried out using Omniscript RT kit and oligo dT primers (Qiagen, Valencia, CA, USA). PCR primers with appropriate flanking restriction sites were synthesized by Eurofins Genomics, Ebersberg, Germany. PCR was performed with Phusion High-Fidelity PCR Master Mix (ThermoScientific). Following agarose gel electrophoresis, PCR fragments were extracted (QIAquick gel extraction kit, Qiagen), submitted to enzyme restriction (*Eco*RI/*Xba*I for the long N-

term isoform and *Eco*RI/*Xho*I for the short N-term isoform), purified (QIAquick PCR purification kit, Qiagen) and kept for cloning. Their sequence identity was verified by Sanger sequencing (Eurofins Genomics, Ebersberg, Germany). The primers used to amplify Mm523 (Accession No: XM_006530804.1) isoforms are supplied in supplementary data 1(a).

Integration assay

Plasmid expression for transposases. The plasmids pCS2-PB and pCS2-PB.NLS-1-558 encode the V5 tagged PB transposases. Each cDNA was inserted into the multi-cloning site of the pCS2+ vector (Life Technologies, Paisley, UK) as described.[36] The plasmid pCS2-Mm523 encodes a two myc tagged PGBD5 isoform 524 amino acid residues in size. Mm523 cDNA was inserted into the multi-cloning site of a modified pCS2 vector with an in-frame N-term 5XMyc tag.[37] The plasmid pCS2-GFP plasmid was built by cloning the gene coding for the green fluorescent protein gene into the multi-cloning site of the pCS2+ vector. pCS2-GFP was used as a negative control of transposition (i.e. absence of transposase expression).

Plasmids donor of transposon. The plasmid pBSK-IFP2-TIR5′-NeoR-TIR3′ (supplementary data 1(b)) was built by introducing the IFP2 5′ and 3′ terminal regions (262 and 400 bp, respectively) into the pBluescript SK plasmid (pBSK). A cassette (NeoR) containing a SV40 promoter, the neomycin phosphotransferase ORF and a sv40 terminator was cloned between transposon ends as described.[36] NeoR was cloned in its middle using a BamHI site that was added to its sequence during DNA synthesis. The plasmid pBSK-NeoR was built by cloning the NeoR cassette into the multi-cloning site of a pBSK plasmid as described.[38]

Integration assay in HeLa cells. Assays were monitored as described.[36] Briefly, each sample of 100,000 cells in a well of a 24-well plates of plaque assays was co-transfected with JetPEI (Polyplus-transfection, Illkirch-Graffenstaden) and 400 ng DNA plasmid and with equal amounts of donor of NeoR cassette included or not within a transposon and transposase sources (1:1 ratio). Two days post-transfection, each cell sample was transferred to a cell culture dish (100 mm diameter) and selected with a culture medium containing 800 μg/mL G418 sulfate (Eurobio Scientific, Les Ulis) for 15 days. After two washing with 1X saline phosphate buffer, cell clones were fixed and stained overnight with 70% EtOH-0.5% methylene blue and colonies >0.5 mm in diameter were counted. Experiments were performed at least twice in triplicate.

Integration assay in G401 cells. Assays were monitored as described.[14] Two clonal cell G401 lines were used. The first line was lentivirally transduced to constitutively express specific shRNA suppressing the expression of Hs524 PGBD5.[14] The second line was modified as a control to constitutively express shRNA to target GFP which is not expressed, thereby preserving the endogenous expression of Hs524 PGBD5. Briefly, each sample of 100,000 cells in a well of a 24-well plates of plaque assays was transfected with jetOptimus and 500 ng DNA plasmid pBSK-IFP2-TIR5′-NeoR-TIR3′ as recommended by the supplier (Polyplus-transfection, Illkirch-Graffenstaden). Two days post-transfection, each cell sample was transferred to a cell culture dish (100 mm diameter) and selected with a culture medium containing 2 mg/mL G418 sulfate (Eurobio Scientific, Les Ulis) for 15 days. After two washing with 1X saline phosphate buffer, cell clones were fixed and stained overnight with 70% EtOH-0.5% methylene blue and colonies >0.5 mm in diameter were counted. Experiments were performed at least twice in triplicate.

Cellular localization of green fluorescent protein-fusion proteins

Plasmid expression for transposase-GFP fusions. The plasmids pCS2-PB-GFP, pCS2-PB.1-558, pCS2-PB.NLS-1-558 and pCS2-Mm523-GFP were made as described.[39]

Cell manipulation. HeLa cells were plated at a density of 5×10^4 cells per well in 1 cm^2 Lab-Tek™ chamber slides (Fisher Scientific, Waltham, MA, USA) and grown in DMEM (Gibco/Life Technologies, Paisley, UK) supplemented with 10% heat inactivated fetal bovine serum (FBS, Eurobio, France) at 37 °C in a humidified atmosphere containing 5% CO_2 for 48 h. Cells were transfected with 500 ng plasmid DNA and jetPEI™ (Polyplus Transfection, Illkirch, France) at an N/P ratio of 5 in DMEM 10% FBS following the Manufacturer's instructions. Cells were then incubated with the complexes for 4 h. The transfection medium was then discarded and replaced by fresh DMEM supplemented with 10% FBS before being incubated for 48 hours at 37 °C.

Imaging. Cells on slides were fixed in 1X PBS/2% paraformaldehyde at RT for 15 min, and then permeabilised with PBS/1% (w/v) Triton-X100 for 10 min. The slides were washed three times for 5 min with 1x PBS. Nuclei were stained using Vectashield Vibrance "Antifade Mounting Medium (hardening) + DAPI" (Vector Laboratories, Burlingame CA, USA). All images of fluorescence were collected with an LSM 700 laser scanning microscope and the associated Zen software (Carl Zeiss, Oberkochen, Germany). All images shown

correspond to one focal plane (0.5 μm). Images to be used for figures were pseudocolored by LSM Image browser software (Carl Zeiss, Thornwood, NY) and Photoshop (Adobe Systems, San Jose, CA) was on the resulting tiff files only to adjust for brightness and contrast.

Recovery of integration sites

LAM-PCR and Illumina libraries. Integration assays were done to produce cell populations containing integrated copies of the donor transposon. Fifteen days post-transfection, cell clones were harvested for genomic DNA preparation using the DNeasy kit (Qiagen, Hilden, Germany). Linear amplification-mediated PCR (LAM-PCR) was performed to amplify the vector-genomic DNA junctions of *Ifp2* vectors as described.[40] All PCR were done using the high fidelity Q5 DNA Polymerase (New England Biolabs, Ipswich, MA). For both approaches, 1 μg DNA was used for twice 50 rounds of linear amplification using a biotinylated primer anchored near one end of the NeoR cassette to enrich DNA species containing transposon-chromosomal DNA junctions (for sequences of (B)-NeoR 5′ and 3′ primers, see supplementary data 1(c)). One reaction was done per ends. The single-stranded products were immobilized on streptadivin-coated magnetic beads (Dynabeads M-280 Streptavidin, Invitrogen, Carlsbad, CA). All subsequent steps were performed on the magnetic bead-bound DNA. Two washes with water followed each step. Second strand synthesis was performed with random hexamer primers (Roche, Basel, Switzerland) using Klenow DNA polymerase (New England Biolabs, Ipswich, MA). The double-stranded DNA was split in two batches and subjected to restriction digests with *DpnI* for the first one and *Pci*I, *Nco*I and *Bsp*HI for the second one using restriction enzymes. The DNA fragments with a CG-3′ or a CATG-3′ overhang ends were ligated to linkers displaying appropriate overhang ends and made from annealed oligonucleotides (supplementary data 1(c)).

To increase the specificity of the full process, an initial PCR was done using one biotinylated primer anchored within the 5′ or 3′ region of the transposon donor and one primer anchored within the linker (for sequences of (B)-TIR-UTR 5′ and 3′, and LC1 primers, see supplementary data 1(c)). PCR products were immobilized on streptadivin-coated magnetic beads and purified as described above. Next, the bead-bound DNA was subjected to a nested PCR using nested primers anchored within transposon ends and within linkers (supplementary data 1(c)). Final PCR products were purified, quantified and gathered in equimolar DNA amounts for each transposon vector (4 populations of LAM-PCR products) before being used to make Illumina libraries using NEBNext® Ultra™ II DNA Library Prep Kit for Illumina® and NEBNext Multiplex Oligos for Illumina (New England Biolabs, Ipswich, MA). Fragment size selection, library quality control and Illumina sequencing (MiSeq 250 nucleotides, TruSeq SBS Kit v3) were achieved at the Plateforme de Séquençage Haut Débit I2BC (Gif-sur-Yvette, France). DNA quantities were monitored at various steps in the procedure with the Qubit® dsDNA (Molecular Probes, Eugene, USA).

Computer analysis. Trimmomatic[41] was used to filter Miseq reads using default parameters, except for SLIDINGWINDOW:5:20 and MINLEN:100. The purpose of the following steps was to recover chromosome-inserted DNA fragment junctions taking into account the plasmid backbone regions located 100-bp upstream and downstream the *Ifp2*-NeoR transposon. Filtered reads were first mapped to the sequence of plasmid backbone minus the 100-bp regions flanking on both sides the *Ifp2*-NeoR transposon with bwa-mem using default parameters.[42] Unmapped reads were then extracted reads using SAMtools view with parameters-b-f 4[43] and bamToFastq from the BEDTools suite using default parameters.[44] Recovered unmapped reads were aligned using bwa-mem against a bwa bank gathering the sequences of hg38 chromosomes plus those of the *Ifp2*-NeoR transposon flanked by the 100-bp plasmid backbone regions on both sides (supplementary data 1(d)). Default parameters were used excepted for -w 1 and -r 1. The bam files resulting from each dataset alignment were analysed with Lumpy in order to identify split reads.[45] The parameters were -e -mw 2 -tt 0.0 and back_distance:20,weight:1,id:lumpy_v1,min_mapping_threshold:20. Structural variants (SV) characterized by "BND" for the broken end notations and displaying for each of them an SV with two positions, one genomic and one on the transposon, were extracted using a house python program (https://github.com/Leelouh/lumpy2site). Results were filtered taking into account a difference below 3 between the transposon breakpoint calculated by Lumpy and the maximal spread of read alignments in the transposon donor sequence for each integration event. Each TSD nucleotide motif at insertion site was obtained after extracting 10-bp sequences before and after the breakpoint in the chromosome sequences.

Gene ontology (GO) analyses were focused mostly on protein coding genes and those encoding long non-coding RNA (lncRNA). We used hg38 gene annotations from UCSC. Gene ontology was first investigated using DAVID (https://david.ncifcrf.gov/) and AmiGO2 (http://amigo.geneontology.org/amigo) to assess term enrichment. This was followed up by the Cytoscape plugin ClueGO.[46,47]

Access of publicly available data

Sequences corresponding to *Ifp2*, *Sleeping Beauty* and *TcBuster* insertion sites in HEK 293 cells[23] were downloaded from public databases using accession numbers JS717545 to JS799249. Sequences corresponding to *Ifp2* insertion sites in HCT116 cells[15] were recovered at https://www.genetics.org/content/suppl/2012/01/03/genetics.111.137315.DC1. Sequences corresponding to *Ifp2* and *Sleeping Beauty* insertion sites in CD4+ cells[26] were recovered in the GSE58744 at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1419000 and https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1419001. For the last two sources, the positions of insertion sites were transformed in hg38 using liftover at https://genome.ucsc.edu/cgi-bin/hgLiftOver. All sites mapped in hg38 were supplied in supplementary Table 4.

PWMTrain[48] at https://ccg.epfl.ch/pwmtools/pwmtrain.php was used to calculate the position-specific weight matrix of *TcBuster* insertion sites using available data.[23] The numbers and the positions of putative insertion sites in hg38 for *pbles* (TTAA), *sleeping beauty (*TA) and *TcBuster* were calculated using PWMScan[48] at https://ccg.epfl.ch//pwmtools/pwmscan.php.

The list of 6985 neuron genes in hg19 was recovered in supplemental data of 31 and was updated to hg38. 131 genes were removed. They corresponded to artefactual genes coding nc RNA that were withdrawn in hg 38. DNAse1 map for HeLa and HEK293 cells were recovered at http://hgdownload.soe.ucsc.edu/wgEncodeAwgDnaseUwdukeHelas3UniPk.narrowPeak.gzgoldenPath/hg19/encodeDCC/wgEncodeOpenChromDnase/, files and http://hgdownload.soe.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeAwgDnaseUniform/wgEncodeAwgDnaseUwdukeHelas3UniPk.narrowPeak.gz. They were updated in hg38 using liftover.

Values in graphs were medians, quartiles 1 and 3 and spread of experiments done at least twice in triplicate. Shapiro-Wilk tests were used to confirm the normality of each set of samples, t-test to analyse distribution differences between experimental samples, Chi2 test to analyse differences between an experimental distribution and a theoretical one, and logarithmic distribution test to analyse enrichments using free tools and tutorials available at http://www.anastat.fr/outils.php.

Permutation tests (10,000 per test) were computed using in-house bash programs that accounted the distribution in TA and TTAA motifs in hg38. The normality of each distribution of permuted results was verified using a Shapiro-Wilk test using free tools and tutorials available at http://www.anastats.fr/outils.php. When the distributions were normal, probabilities were calculated from Z score at https://www.fourmilab.ch/rpkp/experiments/analysis/zCalc.html. When they were not normal, the distributions were used to determine the 1 and 0.1% thresholds at both tails and the observed values were positioned in regards to those values.

Data deposition

All raw and processed data are available through the European Nucleotide Archive under accession number PRJEB36226, PRJEB36229, PRJEB41045 and PRJEB41053. Files describing the annotation of insertion sites copies in the hg38 release are supplied as supplementary Tables 1–3.

Acknowledgments

This work was supported by the C.N.R.S., the I.N.R.A., and the GDR CNRS 2157. It also received funds from a research program grants from the Ligue Nationale Contre le Cancer, the Merck foundation, and the French National Society of Gastroenterology. Laura Helou holds a PhD fellowship from the Région Centre Val de Loire. We acknowledge the high-throughput sequencing facility of I2BC for its sequencing and bioinformatics expertise. Alex Kentsis is a consultant for Novartis and is supported by the National Cancer Institute grants R01 CA214812 and P30 CA008748. Yves Bigot, who was in charge of the achievement of this project does not have to thank the French National Research Agency for its financial support but he kindly thanks it for the excellent reviews embellished with arguments based on scientific and cultural novelties in the expertise of his yearly application file during the last decade.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jmb.2020.166805.

Received 26 November 2020;
Accepted 29 December 2020;
Available online 13 January 2021

Keywords:
transposon;
DNA cleavage;
neuron;
insertion preference;
vertebrate

Abbreviations:
CRD, cysteine-rich domain; gDNA, genomic DNA; GFP, green fluorescent protein; ISCR, insertion sites-containing regions; NeoR, neomycin resistance; NLS, nuclear localisation signal; ORF, open reading frame; PB, *piggyBac* transposase; *pble*, *piggyBac*-like element; PGBD or *pgbd*, "*piggyBac* derived transposase" protein or gene; STIR, sub-terminal inverted repeats; SV40, simian

virus 40; TE, transposable element; TIR, terminal inverted repeats; TSD, target site duplication

References

1. Piégu, B., Bire, S., Arensburger, P., Bigot, Y., (2016). A survey of transposable element classification systems–a call for a fundamental update to meet the challenge of their diversity and complexity. *Mol. Phylogenet. Evol.*, **86**, 90–109.
2. Arensburger, P., Piégu, B., Bigot, Y., (2016). The future of transposable element annotation and their classification in the light of functional genomics - what we can learn from the fables of Jean de la Fontaine?. *Mob. Genet. Elements*, **6**, e1256852.
3. Arkhipova, I.R., (2017). Using bioinformatic and phylogenetic approaches to classify transposable elements and understand their complex evolutionary histories. *Mob. DNA*, **8**, 19.
4. Goerner-Potvin, P., Bourque, G., (2018). Computational tools to unmask transposable elements. *Nat. Rev. Genet.*, **19**, 688–704.
5. Bouallègue, M., Rouault, J.D., Hua-Van, A., Makni, M., Capy, P., (2017). Molecular evolution of piggyBac superfamily: From selfishness to domestication. *Gen. Biol. Evol.*, **9**, 323–339.
6. Cary, L.C., Goebel, M., Corsaro, B.G., Wang, H.G., Rosen, E., Fraser, M.J., (1989). Transposon mutagenesis of baculoviruses: analysis of Trichoplusia ni transposon IFP2 insertions within the FP-locus of nuclear polyhedrosis viruses. *Virology*, **172**, 156–169.
7. Yusa, K., (2015). piggyBac transposon. *Microbiol. Spectr.*, **3** MDNA3-0028-2014.
8. Chen, Q., Luo, W., Veach, R.A., Hickman, A.B., Wilson, M. H., Dyda, F., (2020). Structural basis of seamless excision and specific targeting by piggyBac transposase. *Nat. Commun.*, **11**, 3446.
9. Morellet, N., Li, X., Wieninger, S.A., Taylor, J.L., Bischerour, J., Moriau, S., Lescop, E., Bardiaux, B., Mathy, N., et al., (2018). Sequence-specific DNA binding activity of the cross-brace zinc finger motif of the piggyBac transposase. *Nucl. Acids Res.*, **46**, 2660–2677.
10. Sharma, R., Nirwal, S., Narayanan, N., Nair, D.T., (2018). Dimerization through the RING-Finger domain attenuates excision activity of the piggyBac transposase. *Biochemistry*, **57**, 2913–2922.
11. Keith, J.H., Fraser, T.S., Fraser Jr., M.J., (2008). Analysis of the piggyBac transposase reveals a functional nuclear targeting signal in the 94 c-terminal residues. *BMC Mol. Biol.*, **9**, 72.
12. Hong, J.B., Chou, F.J., Ku, A.T., Fan, H.H., Lee, T.L., Huang, Y.H., Yang, T.L., Su, I.C., et al., (2014). A nucleolus-predominant piggyBac transposase, NP-mPB, mediates elevated transposition efficiency in mammalian Cells. *PLoS One*, **9**, e89396.
13. Luo, W., Galvan, D.L., Woodard, L.E., Dorset, D., Levy, S., Wilson, M.H., (2017). Comparative analysis of chimeric ZFP-, TALE- and Cas9-piggyBac transposases for integration into a single locus in human cells. *Nucl. Acids Res.*, **45**, 8411–8422.
14. Henssen, A.G., Koche, R., Zhuang, J., Jiang, E., Reed, C., Eisenberg, A., Still, E., MacArthur, I.C., et al., (2017). PGBD5 promotes site-specific oncogenic mutations in human tumors. *Nat. Genet.*, **49**, 1005–1014.
15. Wang, H., Mayhew, D., Chen, X., Johnston, M., Mitra, R. D., (2012). "Calling cards" for DNA-binding proteins in mammalian cells. *Genetics*, **190**, 941–949.
16. Li, M.A., Pettitt, S.J., Eckert, S., Ning, Z., Rice, S., Cadiñanos, J., Yusa, K., Conte, N., Bradley, A., (2013). The piggyBac transposon displays local and distant reintegration preferences and can cause mutations at noncanonical integration sites. *Mol. Cell. Biol.*, **33**, 1317–1330.
17. Thorvaldsdóttir, H., Robinson, J.T., Mesirov, J.P., (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinformatics.*, **14**, 178–192.
18. Elick, T.A., Lobo, N., Fraser Jr., M.J., (1997). Analysis of the cis-acting DNA elements required for piggyBac transposable element excision. *Mol. Gen. Genet.*, **255**, 605–610.
19. Mitra, R., Fain-Thornton, J., Craig, N.L., (2008). piggyBac can bypass DNA synthesis during cut and paste transposition. *EMBO J.*, **27**, 1097–1109.
20. Pavelitz, T., Gray, L.T., Padilla, S.L., Bailey, A.D., Weiner, A.M., (2013). PGBD5: a neural-specific intron containing piggyBac transposase domesticated over 500 million years ago and conserved from cephalochordates to humans. *Mob. DNA*, **4**, 23–39.
21. Henssen, A.G., Henaff, E., Jiang, E., Eisenberg, A.R., Carson, J.R., Villasante, C.M., Ray, M., Still, E., et al., (2015). Genomic DNA transposition induced by human PGBD5. *Elife*, **4**, e10565.
22. Lohe, A.R., Timmons, C., Beerman, I., Lozovskaya, E.R., Hartl, D.L., (2000). Self-inflicted wounds, template-directed gap repair and a recombination hotspot. Effects of the mariner transposase. *Genetics*, **154**, 647–656.
23. Woodard, L.E., Li, X., Malani, N., Kaja, A., Hice, R.H., Atkinson, P.W., Bushman, F.D., Craig, N.L., Wilson, M.H., (2012). Comparative analysis of the recently discovered hAT transposon TcBuster in human cells. *PLoS One*, **7**, e42666.
24. Wilson, M.H., Coates, C.J., George Jr., A.L., (2007). PiggyBac transposon-mediated gene transfer in human cells. *Mol. Ther.*, **15**, 139–145.
25. Huang, X., Guo, H., Tammana, S., Jung, Y.C., Mellgren, E., Bassi, P., Cao, Q., Tu, Z.J., et al., (2010). Gene transfer efficiency and genome-wide integration profiling of Sleeping Beauty, Tol2, and piggyBac transposons in human primary T cells. *Mol. Ther.*, **18**, 1803–1813.
26. Gogol-Döring, A., Ammar, I., Gupta, S., Bunse, M., Miskey, C., Chen, W., Uckert, W., Schulz, T.F., et al., (2016). Genome-wide profiling reveals remarkable parallels between insertion site selection properties of the MLV retrovirus and the piggyBac transposon in primary human CD4(+) T cells. *Mol. Ther.*, **24**, 592–606.
27. Landry, J.J., Pyl, P.T., Rausch, T., Zichner, T., Tekkedil, M. M., Stütz, A.M., Jauch, A., Aiyar, R.S., et al., (2013). The genomic and transcriptomic landscape of a HeLa cell line. *G3 (Bethesda)*, **3**, 1213–1224.
28. Adey, A., Burton, J.N., Kitzman, J.O., Hiatt, J.B., Lewis, A.P., Martin, B.K., Qiu, R., Lee, C., Shendure, J., (2013). The haplotype-resolved genome and epigenome of the aneuploid HeLa cancer cell line. *Nature*, **500**, 207–211.
29. Lin, Y.C., Boone, M., Meuris, L., Lemmens, I., Van Roy, N., Soete, A., Reumers, J., Moisse, M., et al., (2014). Genome dynamics of the human embryonic kidney 293 lineage in

response to cell biology manipulations. *Nat. Commun.*, **5**, 4767.
30. Liu, Y., Mi, Y., Mueller, T., Kreibich, S., Williams, E.G., Van Drogen, A., Borel, C., Frank, M., et al., (2019). Multi-omic measurements of heterogeneity in HeLa cells across laboratories. *Nat. Biotechnol.*, **37**, 314–322.
31. Negi, S.K., Guda, C., (2017). Global gene expression profiling of healthy human brain and its application in studying neurological disorders. *Sci. Rep.*, **7**, 897.
32. Zylka, M.J., Simon, J.M., Philpot, B.D., (2015). Gene length matters in neurons. *Neuron*, **86**, 353–355.
33. Zhang, Y., Cheng, T.C., Huang, G., Lu, Q., Surleac, M.D., Mandell, J.D., Pontarotti, P., Petrescu, A.J., et al., (2019). Transposon molecular domestication and the evolution of the RAG recombinase. *Nature*, **569**, 79–84.
34. Guizard, S., Piégu, B., Arensburger, P., Guillou, F., Bigot, Y., (2016). Deep landscape update of dispersed and tandem repeats in the genome model of the red jungle fowl, Gallus gallus, using a series of de novo investigating tools. *BMC Genom.*, **17**, 659.
35. Kapusta, A., Suh, A., (2017). Evolution of bird genomes-a transposon's-eye view. *Ann. N. Y. Acad. Sci.*, **1389**, 164–185.
36. Bire, S., Ley, D., Casteret, S., Mermod, N., Bigot, Y., Rouleux-Bonnin, F., (2013). Optimization of the piggyBac transposon using mRNA and insulators: toward a more reliable gene delivery system. *PLoS One*, **8**, e82559.
37. Travnickova-Bendova, Z., Cermakian, N., Reppert, S.M., Sassone-Corsi, P., (2002). Bimodal regulation of mPeriod promoters by CREB-dependent signaling and CLOCK/BMAL1 activity. *Proc. Natl. Acad. Sci. USA*, **99**, 7728–7733.
38. Bire, S., Casteret, S., Piégu, B., Beauclair, L., Moiré, N., Arensbuger, P., Bigot, Y., (2016). Mariner transposons contain a silencer: Possible role of the polycomb repressive complex 2. *PLoS Genet.*, **12**, e1005902.
39. Demattei, M.V., Hedhili, S., Sinzelle, L., Bressac, C., Casteret, S., Moiré, N., Cambefort, J., Thomas, X., et al., (2011). Nuclear importation of Mariner transposases among eukaryotes: motif requirements and homo-protein interactions. *PLoS One*, **6**, e23693.
40. Bartholomae, C.C., Glimm, H., von Kalle, C., Schmidt, M., (2012). Insertion site pattern: global approach by linear amplification-mediated PCR and mass sequencing. *Meth. Mol. Biol.*, **859**, 255–265.
41. Bolger, A.M., Lohse, M., Usadel, B., (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.
42. Li, H., Durbin, R., (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, **26**, 589–595.
43. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., et al., (2009). The sequence alignment/map (SAM) format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
44. Quinlan, A.R., Hall, I.M., (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.
45. Layer, R.M., Chiang, C., Quinlan, A.R., Hall, I.M., (2014). LUMPY: a probabilistic framework for structural variant discovery. *Gen. Biol.*, **15**, R84.
46. Bindea, G., Mlecnik, B., Hackl, H., Charoentong, P., Tosolini, M., Kirilovsky, A., Fridman, W.F., Pagès, F., et al., (2009). ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics*, **25**, 1091–1093.
47. Mlecnik, B., Galon, J., Bindea, G., (2018). Comprehensive functional analysis of large lists of genes and proteins. *J. Proteom.*, **171**, 2–10.
48. Ambrosini, G., Groux, R., Bucher, P., (2018). PWMScan: A fast tool for scanning entire genomes with a position-specific weight matrix. *Bioinformatics*, **34**, 2483–2484.

Supplementary data 1 : Transposon donor sequences and primers used in the study.

a. Primers used to amplify Mm523 cDNA by RT-PCR.
Mn524 (Accession N°XM_006530804.1)
Forward primer (flanked with *Eco*RI) : 5'-CCGAAGAATTCATGGCCGAGGGCGGTGG-3'
Reverse primer (flanked with *Xba*I) : 5'-CCGTTTCTAGATCAGTGGGCTGGCGATGAG-3'

b. Sequence of the *Ifp2*-TIR5'-NeoR-TIR3 donor (the cassette containing the promoter SV40 - NeoR ORF - sv40 terminator is highlighted in yellow)

TTAACCCTAGAAAGATAATCATATTGTGACGTACGTTAAAGATAATCATGCGTAAAATTGACGCAT
GTGTTTTATCGGTCTGTATATCGAGGTTTATTTATTAATTTGAATAGATATTAAGTTTTATTATAT
TTACACTTACATACTAATAATAAATTCAACAAACAATTTATTTATGTTTATTTATTTATTAAAAAA
AAACAAAAACTCAAAATTTCTTCTATAAAGTAACAAAACTTTTGGTATCGATGTCGACGAATTCTG
TGGAATGTGTGTCAGTTAGGGTGTGGAAAGTCCCCAGCTCCCCAGGCAGGCAGAAGTATGCAAAGC
ATGCATCTCAATTAGTCAGCAACCAGGTGTGGAAAGTCCCCAGGCTCCCCAGCAGGCAGAAGTATG
CAAAGCATGCATCTCAATTAGTCAGCAACCATAGTCCCGCCCCTAACTCCGCCCATCCCGCCCCTA
ACTCCGCCCAGTTCCGCCCATTCTCCGCCCCATGGCTGACTAATTTTTTTTATTTATGCAGAGGCC
GAGGCCGCCTCGGCCTCTGAGCTATTCCAGAAGTAGTGAGGAGGCTTTTTTGGAGGCCTAGGCTTT
TGCAAAAAGCTCCCGGGAGCTTGGATATCCATTTTCGGATCTGATCAAGAGACAGGATGAGGATCG
TTTCGCATGATTGAACAAGATGGATTGCACGCAGGTTCTCCGGCCGCTTGGGTGGAGAGGCTATTC
GGCTATGACTGGGCACAACAGACAATCGGCTGCTCTGATGCCGCCGTGTTCCGGCTGTCAGCGCAG
GGGCGCCCGGTTCTTTTTGTCAAGACCGACCTGTCCGGTGCCCTGAATGAACTGCAGGACGAGGCA
GCGCGGCTATCGTGGCTGGCCACGACGGGCGTTCCTTGCGCAGCTGTGCTCGACGTTGTCACTGAA
GCGGGAAGGGACTGGCTGCTATTGGGCGAAGTGCCGGGGCAGGATCTCCTGTCATCTCACCTTGCT
CCTGCCGAGAAAGTATCCATCATGGCTGATGCAATGCGGCGGCTGCATACGCTTGATCCGGCTACC
TGCCCATTCGACCACCAAGCGAAACATCGCATCGAGCGAGCACGTACTCGGATGGAAGCCGGTCTT
GTCGATCAGGATGATCTGGACGAAGAGCATCAGGGGCTCGCGCCAGCCGAACTGTTCGCCAGGCTC
AAGGCGCGCATGCCCGACGGCGAGGATCTCGTCGTGACCCATGGCGATGCCTGCTTGCCGAATATC
ATGGTGGAAAATGGCCGCTTTTCTGGATTCATCGACTGTGGCCGGCTGGGTGTGGCGGACCGCTAT
CAGGACATAGCGTTGGCTACCCGTGATATTGCTGAAGAGCTTGGCGGCGAATGGGCTGACCGCTTC
CTCGTGCTTTACGGTATCGCCGCTCCCGATTCGCAGCGCATCGCCTTCTATCGCCTTCTTGACGAG
TTCTTCTGAGCGGGACTCTGGGGTTCGAAATGACCGACCAAGCGACGCCCAACCTGCCATCACGAG
ATTTCGATTCCACCGCCGCCTTCTATGAAAGGTTGGGCTTCGGAATCGTTTTCCGGGACGCCGGCT
GGATGATCCTCCAGCGCGGGGATCTCATGCTGGAGTTCTTCGCCCACCCCAACTTGTTTATTGCAG
CTTATAATGGTTACAAATAAAGCAATAGCATCACAAATTTCACAAATAAAGCATTTTTTTCACTGC
ATTCTAGTTGTGGTTTGTCCAAACTCATCAATGTATCTTATCATGTCTGGATCCCCAgctagcTGG
CTCGAGCAATTGGCAGATCTAAGCTTGATATCTATAACAAGAAAATATATATATAATAAGTTATCA
CGTAAGTAGAACATGAAATAACAATATAATTATCGTATGAGTTAAATCTTAAAAGTCACGTAAAAG
ATAATCATGCGTCATTTTGACTCACGCGGTCGTTATAGTTCAAAATCAGTGACACTTACCGCATTG
ACAAGCACGCCTCACGGGAGCTCCAAGCGGCGACTGAGATGTCCTAAATGCACAGCGACGGATTCG
CGCTATTTAGAAAGAGAGAGCAATATTTCAAGAATGCATGCGTCAATTTTACGCAGACTATCTTTC
TAGGGTTAACCGCGGTGGAGCTCCAGCTTTTGTTCCCTTTAGTGAGGGTTAA

c. Primers to amplify loci containing *Ifp2* insertions
First linear amplification (B = biotin)
(B)-NeoR 5' : (B)-TTCTTCGCCCACCCCAACTTGTTTATTGC-3'
(B)-NeoR 3' : (B)-GACTTTCCACACCCTAACTGACA-3'

Linkers
Oligo I : 5'-GACCCGGGAGATCTGAATTCAGTGGCACAGCAGTTAGG-3'
Oligo II DpnII : 5'-GATCCCTAACTGCTGTGCCACTGAATTCAGATC-3'
Oligo II PNB : 5'-CATGCCTAACTGCTGTGCCACTGAATTCAGATC-3'

First round of nested PCR
Primer anchored in linker
LC1 : 5'-GACCCGGGAGATCTGAATTC-3'

Primers anchored in the 5’ or 3’ end of the *Ifp2* vector
(B)-ITR-UTR 5’ PB III : (B)-CGATAAAACACATGCGTCAAT -3’
(B)-ITR-UTR 3’ PB III : (B)-GCTATTTAGAAAGAGAGAGCA -3’

Second round of nested PCR
Primer anchored in linker
LC2 : 5’-GATCTGAATTCAGTGGCACAG-3’

Primers anchored in the 5’ or 3’ end of the *Ifp2* vector
ITR-UTR PB 5’III : 5’-CGATAAAACACATGCGTCAAT -3’
ITR-UTR PB 3’III : 5’-GCTATTTAGAAAGAGAGAGCA-3’

d. Sequences used as baits to identify transposon breakpoints in analyses done with lumpy.

Sequence of the *Ifp2*-TIR5'-NeoR-TIR3 sequence flanked upstream and downstream by 100-bp plasmid regions. The cassette containing the promoter SV40 - NeoR ORF - sv40 terminator was highlighted in yellow. TTAA TSD were bolded while 100-bp plasmid backbone sequences were typed in blue.

ATTAAGTTGGGTAACGCCAGGGTTTTCCCAGTCACGACGTTGTAAAACGACGGCCAGTGAGCGCGC
GTAATACGACTCACTATAGGGCGAATTGGGTACC**TTAA**CCCTAGAAAGATAATCATATTGTGACGT
ACGTTAAAGATAATCATGCGTAAAATTGACGCATGTGTTTTATCGGTCTGTATATCGAGGTTTATT
TATTAATTTGAATAGATATTAAGTTTTATTATATTTACACTTACATACTAATAATAAATTCAACAA
ACAATTTATTTATGTTTATTTATTTATTAAAAAAAAACAAAAACTCAAAATTTCTTCTATAAAGTA
ACAAAACTTTTGGTATCGATGTCGACGAATTCTGTGGAATGTGTGTCAGTTAGGGTGTGGAAAGTC
CCCAGCTCCCCAGGCAGGCAGAAGTATGCAAAGCATGCATCTCAATTAGTCAGCAACCAGGTGTGG
AAAGTCCCCAGGCTCCCCAGCAGGCAGAAGTATGCAAAGCATGCATCTCAATTAGTCAGCAACCAT
AGTCCCGCCCCTAACTCCGCCCATCCCGCCCCTAACTCCGCCCAGTTCCGCCCATTCTCCGCCCCA
TGGCTGACTAATTTTTTTTATTTATGCAGAGGCCGAGGCCGCCTCGGCCTCTGAGCTATTCCAGAA
GTAGTGAGGAGGCTTTTTTGGAGGCCTAGGCTTTTGCAAAAAGCTCCCGGGAGCTTGGATATCCAT
TTTCGGATCTGATCAAGAGACAGGATGAGGATCGTTTCGCATGATTGAACAAGATGGATTGCACGC
AGGTTCTCCGGCCGCTTGGGTGGAGAGGCTATTCGGCTATGACTGGGCACAACAGACAATCGGCTG
CTCTGATGCCGCCGTGTTCCGGCTGTCAGCGCAGGGGCGCCCGGTTCTTTTTGTCAAGACCGACCT
GTCCGGTGCCCTGAATGAACTGCAGGACGAGGCAGCGCGGCTATCGTGGCTGGCCACGACGGGCGT
TCCTTGCGCAGCTGTGCTCGACGTTGTCACTGAAGCGGGAAGGGACTGGCTGCTATTGGGCGAAGT
GCCGGGGCAGGATCTCCTGTCATCTCACCTTGCTCCTGCCGAGAAAGTATCCATCATGGCTGATGC
AATGCGGCGGCTGCATACGCTTGATCCGGCTACCTGCCCATTCGACCACCAAGCGAAACATCGCAT
CGAGCGAGCACGTACTCGGATGGAAGCCGGTCTTGTCGATCAGGATGATCTGGACGAAGAGCATCA
GGGGCTCGCGCCAGCCGAACTGTTCGCCAGGCTCAAGGCGCGCATGCCCGACGGCGAGGATCTCGT
CGTGACCCATGGCGATGCCTGCTTGCCGAATATCATGGTGGAAAATGGCCGCTTTTCTGGATTCAT
CGACTGTGGCCGGCTGGGTGTGGCGGACCGCTATCAGGACATAGCGTTGGCTACCCGTGATATTGC
TGAAGAGCTTGGCGGCGAATGGGCTGACCGCTTCCTCGTGCTTTACGGTATCGCCGCTCCCGATTC
GCAGCGCATCGCCTTCTATCGCCTTCTTGACGAGTTCTTCTGAGCGGGACTCTGGGGTTCGAAATG
ACCGACCAAGCGACGCCCAACCTGCCATCACGAGATTTCGATTCCACCGCCGCCTTCTATGAAAGG
TTGGGCTTCGGAATCGTTTTCCGGGACGCCGGCTGGATGATCCTCCAGCGCGGGGATCTCATGCTG
GAGTTCTTCGCCCACCCCAACTTGTTTATTGCAGCTTATAATGGTTACAAATAAAGCAATAGCATC
ACAAATTTCACAAATAAAGCATTTTTTTCACTGCATTCTAGTTGTGGTTTGTCCAAACTCATCAAT
GTATCTTATCATGTCTGGATCCCCAgctagcTGGCTCGAGCAATTGGCAGATCTAAGCTTGATATC
TATAACAAGAAAATATATATATAATAAGTTATCACGTAAGTAGAACATGAAATAACAATATAATTA
TCGTATGAGTTAAATCTTAAAAGTCACGTAAAAGATAATCATGCGTCATTTTGACTCACGCGGTCG
TTATAGTTCAAAATCAGTGACACTTACCGCATTGACAAGCACGCCTCACGGGAGCTCCAAGCGGCG
ACTGAGATGTCCTAAATGCACAGCGACGGATTCGCGCTATTTAGAAAGAGAGAGCAATATTTCAAG
AATGCATGCGTCAATTTTACGCAGACTATCTTTCTAGGG**TTAA**CCGCGGTGGAGCTCCAGCTTTTG
TTCCCTTTAGTGAGGGTTAATTGCGCGCTTGGCGTAATCATGGTCATAGCTGTTTCCTGTGTGAAA
TTGTTATCCGCT

Chapitre II. La protéine dérivée de piggyBac 5 (PGBD5) transpose à la fois les éléments proches et éloignés de *piggyBac*, *Tcr-pble* et *Ifp2*

1. Contexte et objectifs de l'article

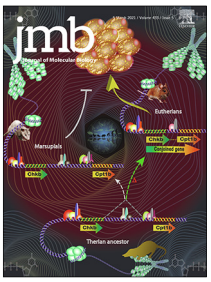
Le premier volet de mes travaux de thèse a été d'identifier au moins un *pble* appartenant à la même lignée phylogénétique que *pgbd5*. Atteindre cet objectif requiert en préalable de vérifier et mettre à jour les informations concernant l'origine des autres *pgbd* chez les vertébrés. Cette étape est indispensable dans la recherche d'un *pble* proche de *pgbd5*. En effet, une connaissance approfondie des caractéristiques des membres de la famille *pgbd* et de leurs *pble* est nécessaire afin de sélectionner les séquences *pble* candidates du plus proche parent de *pgbd5*. Dans un premier temps, il a été nécessaire d'actualiser l'inventaire des gènes *pgbd* chez les vertébrés et d'en vérifier leurs signatures de domestication. Dans un second temps, en utilisant les bases de données publiques, nous avons identifié les *pble* les plus proches de chacun des neuf gènes *pgbd identifiés*. Nous avons profité de cet inventaire pour vérifier comment les séquences des TIR des *pble* sont conservées dans chacun des huit clades à l'origine de chacun des gènes *pgbd* chez les vertébrés.

2. Résumé des principaux résultats

Une fois le plus proche *pble* identifié, le second objectif a été de vérifier si PGBD5 transpose aussi efficacement que l'élément *ifp2*. *Ifp2* est l'un des *pble* phylogénétiquement les plus éloignés de *pgbd5*.

Neuf gènes *pgbd* sont présents chez les vertébrés, nous en avons complètement caractérisé trois nouveaux (*pgbd7, 8* et *9*) à l'occasion de cette étude. En utilisant les bases de données publiques, nous avons identifié un à trois *pble* parents les plus proches de chacun des neuf gènes *pgbd*. Nous avons constaté que sept d'entre eux, *pgbd* 3 à 9, provenaient de sept lignées distinctes de *pble*. *Pgbd1* et *pgbd2* semblent partager le même *pble* « ancestral ». Nous avons identifié le *pble* le plus proche de *pgbd5*, *Tcr-pble*. L'analyse des extrémités des *pble* dans les huit clades de gènes *pgbd* a révélé qu'au-delà du motif TTAACNNNT, aucune conservation de séquence primaire n'est observée entre les clades. Enfin, nous avons testé si PGBD5 était capable de mobiliser l'élément ancestral *Tcr-pble* aussi ou plus efficacement qu'il est capable de mobiliser *ifp2*. Nous avons constaté que PGBD5 pouvait médier des transpositions propres et impropres des deux *pble*.

Research Article



The *piggyBac*-derived protein 5 (PGBD5) transposes both the closely and the *distantly* related *piggyBac*-like elements *Tcr-pble* and *Ifp2*

Laura Helou [1], Linda Beauclair [1], Hugues Dardente [1], Benoît Piégu [1], Louis Tsakou-Ngouafo [2,3], Thierry Lecomte [4], Alex Kentsis [5,6,7], Pierre Pontarotti [1,3] and Yves Bigot [1]*

*1 - **UMR INRAE 0085,** CNRS 7247, Physiologie de la Reproduction et des Comportements, Centre INRA Val de Loire, 37380 Nouzilly, France*
*2 - **UMR MEPHI D-258,** I, IRD, Aix Marseille Université, 19-21 Boulevard Jean Moulin, 13005 Marseille, France*
*3 - **CNRS SNC 5039,** 13005 Marseille, France*
*4 - **EA GICC 7501,** CHRU de Tours, 37044 TOURS Cedex 09, France*
*5 - **Molecular Pharmacology Program,** Sloan Kettering Institute, Memorial Sloan Kettering Cancer Center, New York, NY, USA*
*6 - **Weill Cornell Medical College,** Cornell University, New York, NY, USA*
*7 - **Department of Pediatrics,** Memorial Sloan Kettering Cancer Center, New York, NY, USA*

***Correspondence to Yves Bigot:** yves.bigot@inrae.fr (Y. Bigot)*
https://doi.org/10.1016/j.jmb.2021.166839
Edited by M.F. Summers

Abstract

The vertebrate *piggyBac* derived transposase 5 (PGBD5) encodes a domesticated transposase, which is active and able to transpose its distantly related *piggyBac*-like element (*pble*), *Ifp2*. This raised the question whether PGBD5 would be more effective at mobilizing a phylogenetically closely related *pble* element. We aimed to identify the *pble* most closely related to the *pgbd5* gene. We updated the landscape of vertebrate *pgbd* genes to develop efficient filters and identify the most closely related *pble* to each of these genes. We found that *Tcr-pble* is phylogenetically the closest *pble* to the *pgbd5* gene. Furthermore, we evaluated the capacity of two murine and human PGBD5 isoforms, Mm523 and Hs524, to transpose both *Tcr-pble* and *Ifp2* elements. We found that both *pble*s could be transposed by Mm523 with similar efficiency. However, integrations of both *pbles* occurred through both proper transposition and improper PGBD5-dependent recombination. This suggested that the ability of PGBD5 to bind both *pbles* may not be based on the primary sequence of element ends, but may involve recognition of inner DNA motifs, possibly related to palindromic repeats. In agreement with this hypothesis, we identified internal palindromic repeats near the end of 24 *pble* sequences, which display distinct sequences.

© 2021 Elsevier Ltd. All rights reserved.

Introduction

Transposable elements (TEs) are mobile genetic elements present in almost all organisms. Three main types of TEs are found in eukaryote genomes: 1) LTR-retrotransposons and 2) nonLTR-retrotransposons that encode a reverse-transcriptase to replicate an RNA copy into a DNA intermediate which thereafter integrates at a new locus; and 3) transposons with terminal inverted repeats (TIR transposon) at their extremities that code for and use a transposase to move from one locus to another using a DNA molecule as an intermediate.[1,2]

Ifp2 is an autonomous TIR transposon, 2475 bp long, commonly called *piggyBac*.[3] It is the reference element of the *piggyBac* family and originated from the cabbage looper (*Trichoplusia ni*) genome. *Ifp2*

and the other members of its family, the *piggyBac*-like elements (*pbles*), insert into a tetranucleotide TTAA that is duplicated upon insertion.[4,5] They are flanked at their extremities by TIRs approximately 13 bp long. Depending on the *pble* "species" they can also display sub-terminal inverted repeats (STIR) which are 19 bp in *Ifp2* and located internally at 3 and 31 nucleotides distance of 5' and 3' TIR inner ends, respectively.[6,7] An open reading frame (ORF) in the inner region of the element, about 1.8 kb long, encodes a *pble* transposase approximately 600 amino acids long (594 aa in *Ifp2* transposase, a.k.a. PB).

TEs can have deleterious effects on the genome of their host through mutagenic consequences of their mobility. TEs can also have a positive effect on their host at the evolutionary scale. They have contributed to the adaptation of gene expression by being a source of novelties to rewire networks of cis-regulating elements. ORFs contained in TEs may also lead to novel genes and functions in the host,[8,9] through a process known as molecular domestication or exaptation. While the nucleic acid sequence of the original copies of TEs diverge and degenerate by accumulating mutations over time, the domesticated ORFs derived from TEs are preserved due to selection on their function. Such domesticated ORFs are usually present as single gene copies and conserved at orthologous loci in all related species in which the gene has been domesticated in the common ancestor of the clade.

In vertebrates there are several cases of domesticated genes originating from TIR transposon ORFs[8] and some still encode a transposase that retains its recombinase activity. The RAG1/RAG2 DNA recombinase derived from *Transib/protoRAG* transposases constitutes a prototypical example.[10,11] RAG1/RAG2 are both required as key actors in vertebrate V(D)J recombination.[11–13] A recent inventory of *piggyBac* sequences in eukaryotes identified 6 well-characterized *pgbd* genes in vertebrate genomes[6]: *pgbd1-5* genes and the *Kobuta* gene (hereafter called *pgbd6*). *Pgbd5* appears to be the most ancient case of *piggyBac* ORF domestication and is present in all chordates including cephalochordates and vertebrates, while being absent in tunicates.[14,15] This supports the hypothesis that *pgbd5* was the result of a domestication event in the common chordate ancestor of cephalochordates and vertebrates, but was likely lost in the common ancestor of tunicates.[16,17] The activities and roles of *pgbd* genes in vertebrates remain widely unexplored. For *pgbd5*, two PGBD5 isoforms of ~523 and 409 amino acids are expressed in vertebrate species (Mm523 and Mm409 in *Mus musculus*). At least four isoforms are predicted in human cells (Hs554, Hs524, Hs455, Hs435). Hs524 was found to function as a mutator in various human cancers, but its role in healthy tissues remains enigmatic.[18] Interestingly, human Hs524 and Hs455 were found to be transposase isoforms that display a very unusual property: they transpose *Ifp2,* a *pble* that is one of the most distantly related element to the vertebrate *pgbd5* gene.[6,19] This raises the question of whether human PGBD5 may be more efficient at transposing a *pble* more closely related to itself than to insect *Ifp2*.

Here, we investigated this possibility. We first identified the closest *pble* relative of the *pgbd5* gene. We then compared its integration features with those of *Ifp2* using PGBD5 as a transposase source. Because no *pble* close to *pgbd5* gene has been deposited so far into public databases[6] and due to the presence of ambiguities among *pgbd* genes, the closest *pble* relative of *pgbd5* gene was identified in four steps (supplementary data 1). Indeed, identifying the closest *pble* relatives of PGBD5 was challenging because this gene was the most ancient case of *pgbd* gene domestication in chordate genomes (~600 million years). In-depth knowledge of features of other *pgbd* genes and their *pble* relatives were therefore required to accurately filter sequences. We updated the landscape of *pgbd* genes in vertebrates using comparative analysis and verification of their domestication signatures. We found 9 *pgbd* genes, 3 of them being new with reproducible features. Using public databases we identified one to three of the closest *pble* relatives for each of the 9 *pgbd* genes. We found that 7 of them originated from distinct *pbles*, except for *pgbd1* and *2* genes. We identified the first *pble* belonging to the same clade as the *pgbd5* gene, *Tcr-pble*. Finally, we tested whether PGBD5 could transpose or integrate *Tcr-pble* more efficiently than *Ifp2*.

Results

Inventory of pgbd genes

Two criteria are used as signatures of domestication: i) the dN/dS ratio of each gene and ii) the interspecific syntenic conservation of their flanking chromosomal regions. The dN/dS ratio indicates whether a gene evolved under neutral selection (ratio close to 1), purifying selection (ratio closer to 0 than 1) or under positive selection (ratio above 1).[20] For a domesticated gene with a physiological function, some form of selection is expected, most likely purifying selection. Interspecific syntenic conservation of the flanking chromosomal regions of a gene is used to indicate whether its genomic environment was conserved. For a domesticated gene, synteny supported chromosomal conservation between species belonging to the same clade and in which it was originally domesticated.

Six *pgbd* genes were inventoried in chordates. Their gene organization places them in two distinct classes: i) genes involving the apparent fusion of an ORF coding a host protein (Zn-SCAN in *pgbd1* gene, an uncharacterized domain in

pgbd2 gene, CSB in *pgbd3* gene; Figure 1(a)) with an ORF coding a transposase-derived protein, and ii) genes derived only from one ORF coding for a transposase derived protein (*pgbd4*, *5* and *6* genes; Figure 1(a)). These 6 *pgbd* genes had different numbers of introns. All have been domesticated.[6] The dN/dS ratios for *pgbd1* through *6* genes were very low (Figure 1(a)), indicating that these genes were under purifying selection. Graphic representations of syntenic environments of *pgbd1* to *5* genes were found at http://www.genomicus.biologie.ens.fr. Although three amphibian genomes were sequenced, orthologous regions containing *pgbd6* gene in *Xenopus tropicalis* were not available in the *Xenopus laevis* and *Nanorana parkeri* genome models, which prevented us from assessing the syntenic conservation of this gene.

Five other potential *pgbd* gene candidates were reported in vertebrates.[6] Some of them were inappropriately identified in databases as genes coding for PGBD2 or PGBD4-like proteins; their domestication signatures have not, so far, been fully verified. Here we have renamed these 5 genes as follows: *pgbd7* gene is specific to Afrotherians, *pgbd8* gene is specific to Stepsirrhini, and *pgbd9*, *10* and *11* genes are specific to the Actinopterygii species (called G3 or NeoPGBD, G2 and G1 in Boullègue *et al.*[6], respectively; all references are supplied in supplementary Table 1). We studied the domestication signatures of these five candidate genes. One criteria that is sometimes used to identify a domesticated TE is the loss of TIRs. However, presence of TIRs in the region containing the *pgbd* gene is not sufficient to exclude domestica-

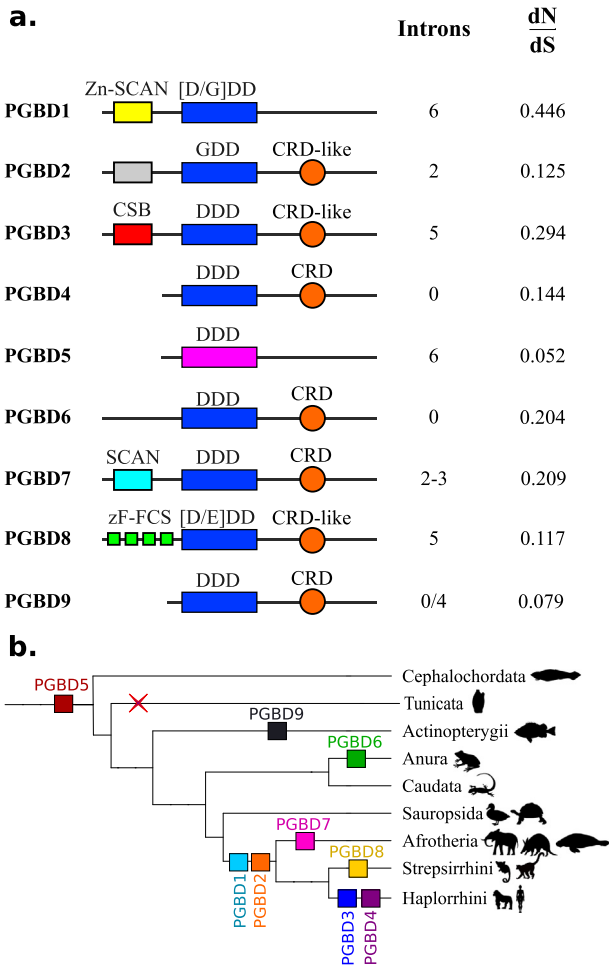


Figure 1. Characteristics of PGBDs in vertebrates. (a) Domain organization in PGBD proteins. For each domesticated sequence, colored boxes represent domains. Yellow is for Zn-SCAN domain, gray is for an unknown domain, red is for CSB domain, light blue is for SCAN domain and green is for zF-FCS. The orange circle is for CRD or CRD-like domains. Blue is for the catalytic triad with the composition of the residues. The catalytic triad is in pink for PGBD5 because it is not aligned with other domesticated sequences. (b) Distribution of PGBDs in vertebrates. Each coloured square represents a PGBD. PGBD1 and PGBD2 are found in mammals, PGBD3 and PGBD4 are found in haplorrhinis, PGBD5 is present in all vertebrates and cephalochordates but not in tunicates. PGBD6 is specific to anura, PGBD7 is found in Afrotheria, PGBD8 is specific to Strepsirrhini and PGBD9 is found in Actinopterygii.

tion. Indeed, the *pgbd3* and *pgbd4* genes were both flanked by sequences corresponding to 5′ and 3′ regions of MER85 and MER75 *pble*, respectively. This suggests that each shares a common origin with one of these two 2 *pbles* (supplementary data 2).

dN/dS ratios of PGBDs. Results showing domain conservation as well as dN/dS ratios are summarized in Figure 1(a). The dN/dS ratio of the *pgbd7* gene was calculated using 4 afrotherian protein sequences (for references see supplementary Table 1) yielding a value of 0.209. The dN/dS ratio of the *pgbd8* gene was 0.117 and was calculated from 6 protein sequences (for references see supplementary Table 1), while the values of actinopterygian-specific *pgbd* 9, 10 and 11 genes were 0.079, 0.146, 0.171 respectively. Two genes had a very low dN/dS ratio, 0.052 for the *pgbd5* gene and 0.079 for the *pgbd9* gene. These genes were the two oldest cases of *piggyBac* domestication in our analysis. In summary, our results indicated that the 11 *pgbd* genes had a dN/dS ratio closer to 0 than to 1 indicating purifying selection.

Syntenic conservation of pgbd7 through 11 gene candidates. Public databases were sufficient to analyze *pgbd7, 8* and *9* genes. Despite a lack of proper annotation and limited number and sizes of contigs, we found that *pgbd7* gene was located between two genes that code for SCAND3 and TRIM27 in 4 related species (Figure 2(a)). For two species (*T. m. latirosis* and *C. asiatica*) three more genes (coding for ZSCAN23, GPX6, and GPX5) were found upstream of the *pgbd7* gene and two other genes (coding for ZNF311 and HLA) were found downstream. *pgbd8* contigs displayed very different sizes as well as many gaps filled with Ns. In two species (*M. murinus* and *P. coquereli*), we succeeded in identifying four genes upstream and downstream of the *pgbd8* gene (Figure 2(b)). In the other species for which shorter contigs were available, we found that the *pgbd8* gene was

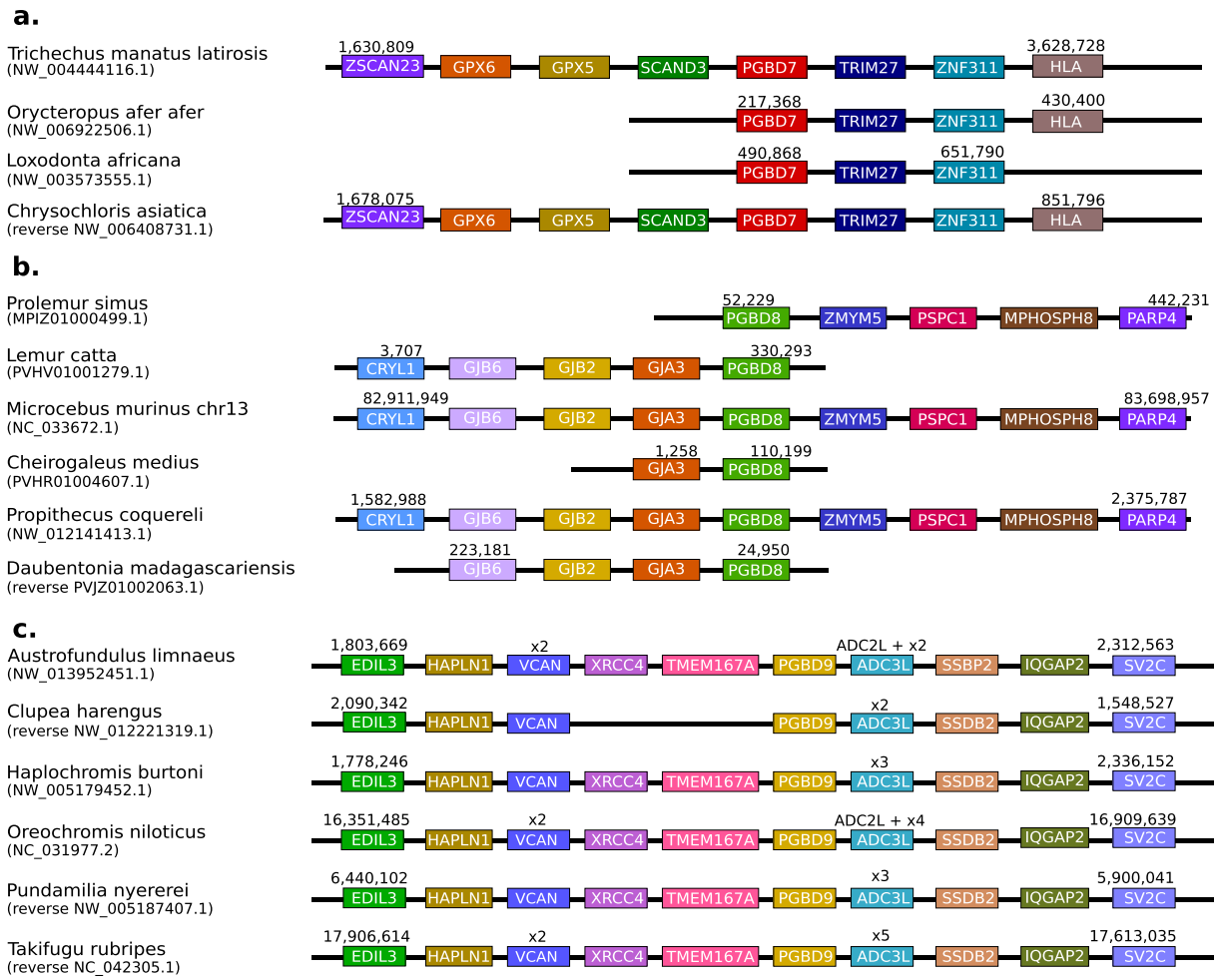


Figure 2. Syntenic conservation around pgbd genes 7, 8 and 9. Schematic representation of syntenic conservation in the neighbouring genes of *pgbd7* (a), *pgbd8* (b), and *pgbd9* (c) across genomes. The names are those of the proteins encoded by the genes. Orthologous genes are colour-coded similarly. Some genomic regions have been drawn in reversed order for visual purposes – designated “reverse”.

located between the same genes. For the *pgbd9* gene four genes were identified in upstream and downstream regions (Figure 2(c)). Some of these were duplicated or even triplicated. *Pgbd9* seemed to be the only gene among the three potential actinopterygians *pgbd* genes to display a well-conserved genomic environment. Our analyses of public databases for the *pgbd10* and *pgbd11* gene candidates were not consistent enough to achieve similar analyses. It should also be noted that the complex situation found in fish genomes, particularly in teleosts, might be due to the presence of giant *pbles* such as "Teratorn"[21,22] (a fusion of a herpesvirus and a *pble*). Indeed, these giant *pbles* contain numerous host and viral genes that are comprised of mobile repeated copies found in the host genome. These results indicated that the *pgbd7, 8* and *9* genes all originated from separate domesticated *pble* ORFs.

Conclusions. To date nine vertebrate *pgbd* genes were shown to display domestication signatures (Figure 1(b)). We have provided evidence supporting the existence of three new *pgbd* genes. *Pgbd7* gene is specific to afrotherian species and encodes a protein characterized by the presence of a SCAN domain at its N-terminal and a CRD at its C-terminal (Figure 1(a)). The *pgbd8* gene is specific to strepsirrhinian species and encodes a PGBD that displays a zF-FCS domain (MYM-type Zinc finger with FCS sequence motif) at its N-terminal and a partial CRD at its C-terminal (noted as CRD-like in Figure 1(a)). The *pgbd8* gene has been annotated as a MYM2 protein in only one species, *Microcebus murinus*. While the *pgbd8* gene was specific to Strepsirrhini species, our BLAST analyses showed that genes coding for PGBD3 and PGBD4 were not present in all primate species but only in a specific sub-order, the Haplorrhini. No trace of these two genes was found in the current Strepsirrhini data, although the percentage of protein identity between PGBD8 and PGBD4 was about 30% (similarity = 49%). The *pgbd9* gene was specific to actinopterygian species.

Identifying the closest pble transposase relative(s) to PGBD5

We focused on *pble* transposases close to each PGBD. We analysed databases with reciprocal BLAST searches using as a query each PGBD sequence. We then verified the relationship of each *pble* candidate by plotting phylogenetic trees. Figure 3 shows a phylogenetic tree summarizing our final results. The protein sequence alignment used to calculate this tree and their references are provided in supplementary data 3 and supplementary Table 1.

We did not find *pbles* closely related to *pgbd1* and *pgbd2* genes but we found one element, *Ota-pble*, located at the root of both genes in the phylogenetic tree (Figure 3). This suggested that the *pgbd1* and *pgbd2* genes resulted from a single domestication event or that a single (or two highly related) *pble* was at the origin of both genes via two closely timed domestication events. For the *pgbd3* gene, our analysis confirmed that *Smi7-pble*,[6] *Hvu-pble* and *Lfu-pble* were the closest relatives of the *pgbd3* genes. For *pgbd4* gene, two *pbles*, *Oab1-pble* and *Pva-pble* were identified as the closest relatives. For the *pgbd5* gene, we identified a new *pble, Tcr-pble,* from the genome of the walking stick, *Timema cristinae* (a phasmid species). Phylogenetic investigations showed that *Tcr-pble* was the only relative close to the *pgbd5* gene found in public databases and its lineage shares the same root as these genes (Figure 3). Because the *pgbd5* gene is the most ancient case of *pble* domestication, the distance between *Tcr-pble* and PGBD5 is greater than that observed between other *pble*/PGBD couples. For PGBD6, two *pble*s were already identified, *Uribo1* and *Uribo2*.[23] For PGBD7, two *pbles* were identified, *Smi2-pble* and *Nvi-pble.* For PGBD8, three *pbles* were identified, *Api1-pble*, *Atr1-pble* and *Api3-pble*. Finally, we identified a new *pble* from a hemichordate, *Ptychodera flava* (*Pfl-pble*), which was the closest relative to PGBD9 together with *Cgi-pble* and *Bgl-pble*.

In all, we identified *pble*s sharing a close common ancestor with each of the 9 *pgbd* genes, *Tcr-pble* being currently the *pble* encoding a transposase that is closest relative of PGBD5 available in databases.

Mobilisation of Ifp2 and Tcr-pble for transposition mediated by PGBD5

Ectopic expression of PGBD5 isoforms can be cytotoxic depending on the cell line and/or their expression levels.[24] In order to manage this issue, two integration assays were used to investigate whether PGBD5 would integrate *Tcr-pble* more efficiently than *Ifp2*, both in terms of integration rates and the precision of integration events at their insertion sites. The first assay was done in HeLa cells transiently transfected with two plasmids: a source of transposon (pBSK-*Ifp2*-TIR5′-NeoR-TIR3′, pBSK-*Tcr-pble*-TIR5′-NeoR-TIR3′ (supplementary data 4a and b), or pBSK-NeoR as a control), and a source of PGBD5, the murine isoform Mm523 (94% identical to its orthologous human isoform, Hs524). In this assay, the expression of Mm523 was driven by a strong CMV promoter, which can be cytotoxic for transfected HeLa cells. However, NeoR clones could be still obtained at a rate lower than that of the control done with the GFP as a source of protein. The second assay was done by transiently transfecting one plasmid source of transposon in a human rhabdoid tumor G401 cell line which expresses Hs524 endogenously.[24] Two clonal G401 cell lines were used. The first line was lentivirally transduced to constitutively express specific shRNA that suppress the expression of Hs524

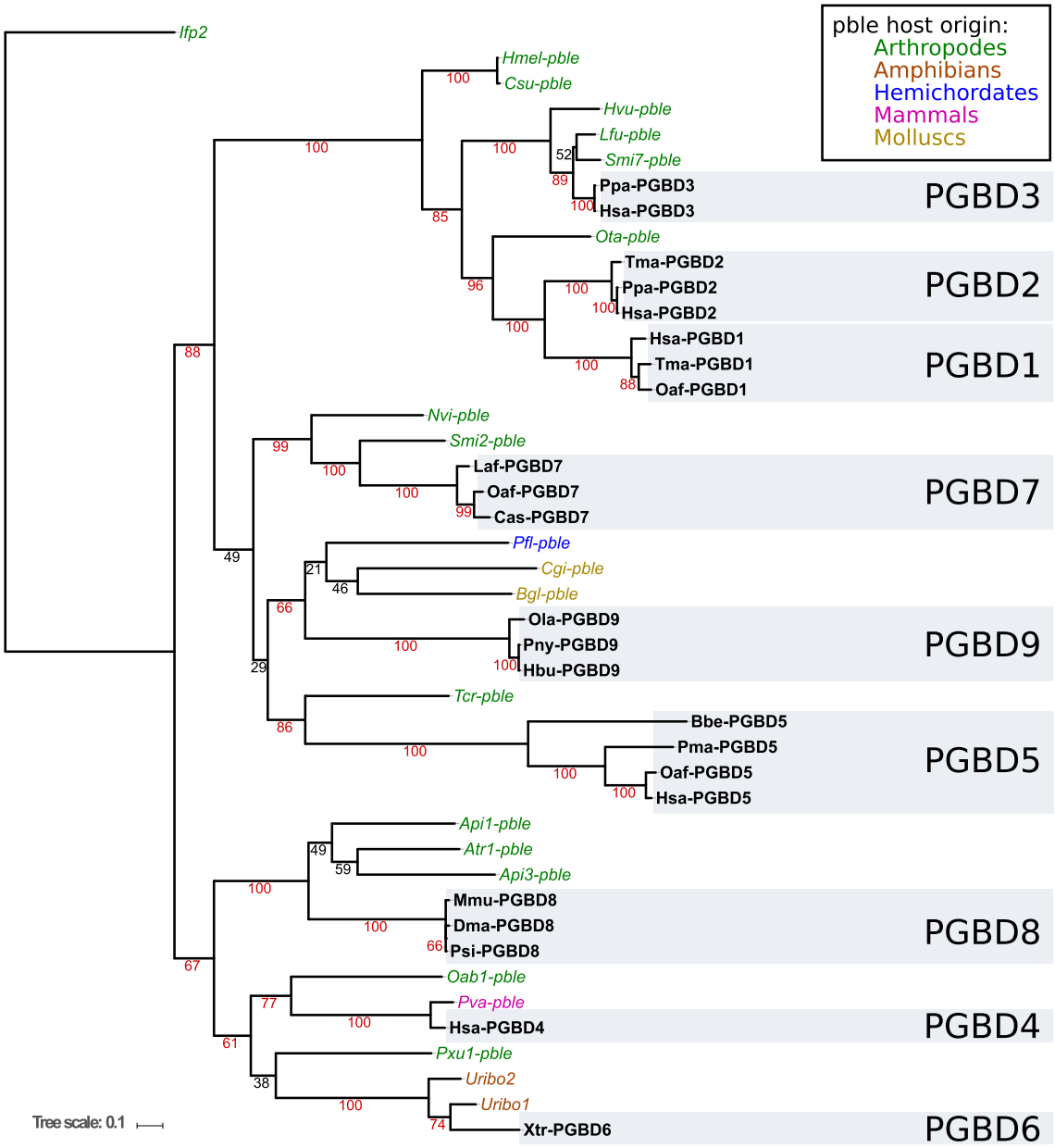


Figure 3. Phylogenetic tree of PBLE and PGBD transposases. Named in bold refers to domesticated *piggyBac* transposases (PGBD), names in italics to *pbles*. The colour of each *pble* refers to the classification and the grey blocks refer to the PGBDs. Bootstraps above 60 are indicated in red.

PGBD5.[13] The second line was modified to constitutively express shRNA to target GFP which is not expressed, thereby preserving the endogenous expression of Hs524 and its recombination activities.

Transposon integration rates in both assays. Integration assays in HeLa cells were monitored separately using transposon-donor plasmids. Two controls were done: the first was done in the presence of GFP instead of Mm523; the second using a source of NeoR cassette void of any transposon sequences (pBSK-NeoR plasmid).

Results confirmed that Mm523 isoform had a negative effect on obtaining NeoR clones since the numbers of clones was one third of the GFP control with pBSK-NeoR (Figure 4(a)), and half of GFP controls when *Ifp2* and *Tcr-pble* were used (Figure 4(b) and (c)). However, a significant difference was found between each of both transposons and the NeoR cassette void of any transposon sequences (t-test, $p < 0.00001$). In spite of cytotoxicity induced by ectopic expression of PGBD5, this suggested that a residual integration activity could be detected under these experimental conditions.

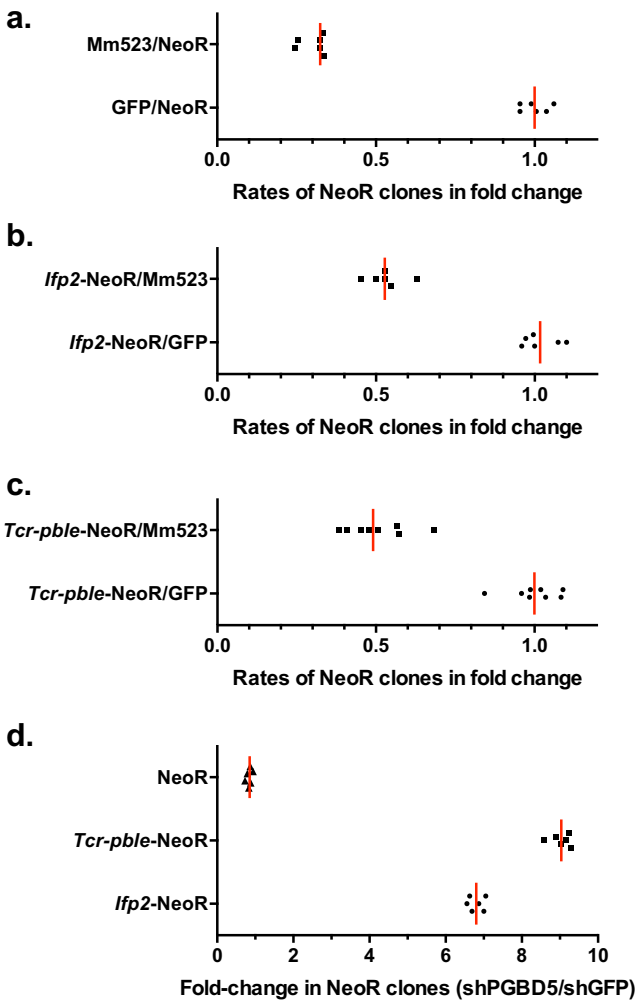


Figure 4. Graphic representations of integration assay results. (a) rates of chromosomal integration in HeLa cells that were mediated by GFP or Mm523 for a NeoR cassette void of transposon sequences; (b) rates of chromosomal integration in HeLa cells that were mediated by GFP or Mm523 for a NeoR cassette contained or not in *Ifp2*; (c) rates of chromosomal integration in HeLa cells that were mediated by GFP or Mm523 for a NeoR cassette contained or not in *Tcr-pble*; (d) rates of chromosomal integration between both G401 lines (shGFP/shPGBD5) that were mediated by Hs524 for a NeoR cassette contained or not in *Ifp2 or Tcr-pble*. In each plot, the red lines represented the median and black symbols, the value of each replicate.

Results obtained with the two clonal G401 cell lines and both transposons as donor plasmids were monitored separately. They revealed that the rates of integration events for *Ifp2* and *Tcr-pble* were seven-fold and nine-fold more elevated than that obtained with a source of NeoR cassette void of any transposon sequences, respectively (Figure 4(d)). This confirmed that PGBD5 actively integrates DNA fragments containing these elements into chromosomes. This also suggested that the sequence of *Tcr-pble* ends would be a bit more efficient than those of *Ifp2* ends to stimulate integration events (t-test: $p < 0.0001$). We thereafter characterized integration events into chromosomes using populations of NeoR clones from both assays.

Landscapes of insertion junctions among neointegrated transposons. To prepare chromosomal DNA samples we used ~ 1800, ~1600, ~1200 and ~ 1000 NeoR clones from integration assays performed with *Ifp2*-NeoR and Mm523, *Tcr-pble*-NeoR and Mm523, *Ifp2*-NeoR and Hs524, and *Tcr-pble*-NeoR and Hs524, respectively. Fragment populations of integration junctions were produced by LAM-PCR from NeoR clones chromosomal DNA, sequenced using Illumina Miseq technology. About 3 million paired Miseq reads (250 nucleotides per read) were obtained for each of the 4 samples. Using computer analyses as described in materials and methods we characterized 1,461 and 1,051 transposon/chromosome junctions produced by

integration events mediated by Mm523 or Hs524 using *Ifp2*-NeoR as a transposon source (supplementary Table 3a and 4a) and 1766 and 210 junctions with *Tcr-pble*-NeoR (supplementary Table 5a and 6a). Junctions at the 5′ and 3′ ends of both transposons were not equally represented in sequence data, perhaps because of different efficiency during DNA fragment amplification in the LAM-PCR.

We further characterized sequence junctions while taking into account the conservation of TSD and TIR sequences, two features required to keep the capacity of neo-inserted elements to be re-remobilized for a new excision and insertion, i.e. to remain "active in mobility".[4,25] Four kinds of junctions were observed: those displaying i) a full TIR sequence and a putative TTAA TSD (Red bars in Figure 5), ii) a region containing an intact TIR and a TTAA TSD juxtaposed with a small piece of plasmid backbone (black bars in Figure 5), iii) no TTAA TSD but a full TIR sequence (blue bars in Figure 5), and iv) no apparent TTAA TSD and a TIR sequence lacking one or several nucleotides at its outer end (grey bars fin Figure 5). In each assay, the relative proportions of each of these 4 categories for each transposon end were evaluated. Because of the read length (250 nucleotides), the parameter of minimal alignment used in bwa and lumpy (20 nucleotides) and the window used for both transposons ends (180 and 170 nucleotides at the 5' end and 175 and 150 nucleotides at the 3' ends of *Ifp2* and *Tcr-pble*, respectively), these proportions could

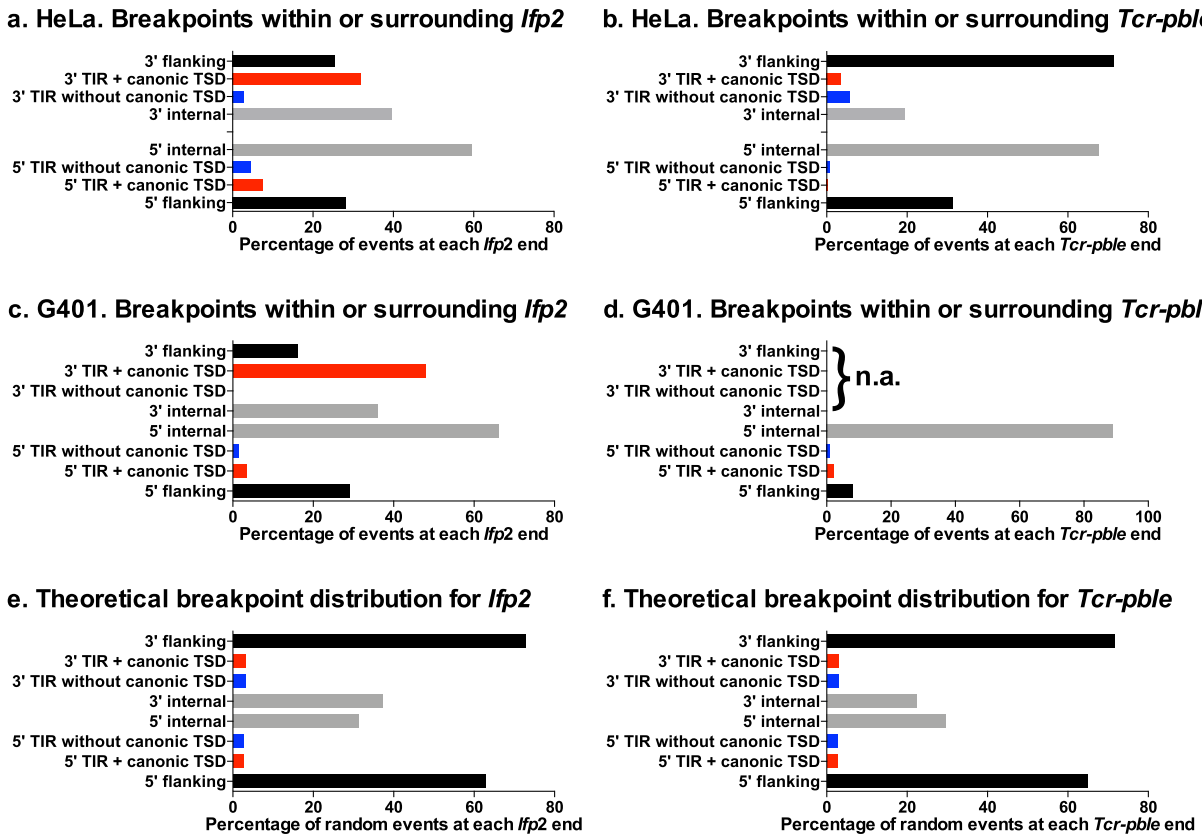


Figure 5. Features of transposon breakpoints. Distributions in *Ifp2* and *Tcr-pble* inserted in the presence of Mm523 (a and b) and Hs524 (c and d), and theoretical distribution whether insertions of fragments containing *Ifp2*-NeoR or *Tcr-pble*-NeoR occur at random (e and f). Black and grey bars correspond to percentages of breakpoints located within the plasmid backbone flanking the transposon and those located within inner transposon regions (from the position 2 in TIR to the primer used for the LAM-PCR), respectively. Four kinds of junctions were observed: those displaying i) a full TIR sequence and a putative TTAA TSD (Red bars from positions 101 to 104 and 2222 to 2225 in supplementary Table 3a and 4a and 101 to 104 and 731 to 734 in supplementary Table 5a and 6a), ii) a region containing an intact TIR and a TTAA TSD juxtaposed with a small piece of plasmid backbone (black bars from positions 1 to 100 and 2226 to 2301 in supplementary Table 3a and 4a and 1 to 100 and 735 to 833 in supplementary Table 5a and 6a), iii) no TTAA TSD but a full TIR sequence (blue bars from positions 102 to 105 and 2218 to 2221 in in supplementary Table 3a and 4a and 102 to 105 and 727 to 730 in supplementary Table 5a and 6a), and iv) no apparent TTAA TSD and a TIR sequence lacking one or several nucleotides at its outer end (black bars from positions 107 to 178 and 2147 to 2217 in supplementary Table 3a and 4a and 106 to 173 and 669 to 726 in supplementary Table 5a and 6a). In (d), n. a. indicates that data were "not available" because the number of transposon/chromosome junctions was too small to perform statistics.

not be biased by the detection procedure for junctions.

Using our analysis procedure for detecting transposon/chromosome junctions, we have re-analysed integration assays previously performed in HEK293 cells with transiently transfected plasmid sources of transposon (*Ifp2*) and PGBD5 (Hs455).[18] We found 202 breakpoints (supplementary Table7), 66% of them displayed features corresponding to *Ifp2* integration into chromosomes by canonical transposition (i.e. with a perfect duplication of the "TTAA" TSD and a full conservation of the TIR sequences). The 33% remaining integration events could be considered as resulting from improper transposition events or alternative mechanisms of chromosomal integration (PGBD5-dependent integration by recombination based on its nuclease activity). When PB was used as a transposase source to transpose *Ifp2* into chromosomes, 96–98 % of breakpoints displayed a perfect duplication of the "TTAA" TSD and conservation of the TIR sequences.[25–27]

The number of perfect transposon/chromosome junctions in Mm523 assays done with *Ifp2* and *Tcr-pble* was 6- to 40-fold lower than when *Ifp2* was transposed by PB (Figure 5(a) and (b)). A similar observation was obtained in G401 cells expressing endogenous Hs524 (Figure 5(c) and (d)), but with a lower rate of perfect transposon/chromosome junctions. This indicated that numerous integration events did not result from canonical transposition events. In order to verify the possibility that these profiles mainly resulted from integration events due to non-canonical activity, the breakpoint distribution at each end was compared to that of a theoretical breakpoint distribution (Figure 5(e) and (f)) using Chi-squared tests. Whatever the profile of transposon/chromosome junctions, p-values were all $\ll 0.01$, indicating that the profiles did not result from random integration events. Interestingly, the numbers of breakpoint junctions were found to be dramatically more elevated within transposon sequences than in the flanking plasmid regions, except for the 3′ end of *Tcr-pble.* This suggested that numerous events might result from non-canonical transposition events during which wounds at transposon ends would occur (Figure 6). To confirm the occurrence of such improper transposition events, we identified 5′ and 3′ junctions of integration events that occurred exactly into the same chromosomal insertion site in each dataset. We found 24, 5, 12 and 1 chromosomal insertion sites in which pairs of transposon/chromosome junctions occurred at both ends and in inverse orientations for *Ifp2*-NeoR and Mm523, *Tcr-pble*-NeoR and Mm523, *Ifp2*-NeoR and Hs524, and *Tcr-pble*-NeoR and Hs524, respectively (supplementary Table 3b to 6b). Each of these junction pairs putatively described both junctions of a single insertion event. Detailed examination of read alignment in the bam file using IGV[28] revealed four cases of putative single integration events in which the *pble*-NeoR transposon was inserted into a duplicated TSD corresponding to duplicated GATC (2 events), CATG (1 event) or C (1 event) motifs.

Finally, we observed that the rate of junctions resulting from canonical transposition events (Figure 5, red bars) was lower with *Tcr-pble* than *Ifp2* in both transposition assays. This suggested that *Tcr-pble* was not a better substrate for canonical transposition with Mm523 and Hs524 than *Ifp2*. Nevertheless, results of transposition assays (Figure 4) supported that *Tcr-pble* was a better substrate than *Ifp2* to actively stimulate the integration of the NeoR cassette by non-canonical transposition and/or alternative mechanisms of DNA integration.

Discussion

The murine and human PGBD5 isoforms Mm523 and Hs524 are able to stimulate the integration of *Ifp2* and *Tcr-pble* in HeLa and G401 cells with no apparent differences, at least under our experimental conditions. They can do it by canonical transposition, but in most cases by other mechanisms, such as PGBD5-dependent recombination based on its nuclease activity. Our findings provide new insights into two distinct questions. The first concerns the identity of the closest *pble* relatives at the origin of each *pgbd* gene in vertebrates, with specific attention to those related to *pgbd5* gene. The second concerns the features of *Ifp2* and *Tcr-pble* ends that stimulate the recombination activities of Mm523 and Hs524.

Domestication of pble transposase genes during evolution of vertebrates

Our study of the origins of *pgbd* genes showed that different elements of the *piggybac* family were domesticated at least 8 or 9 different times over a period spanning over 600 million years of evolution in vertebrates. All *pgbd* genes are under purifying selection pointing at physiological relevance for species in which they were domesticated. New cases of *pble* domestication should be characterized as further vertebrate genomes are sequenced. We do not anticipate any discovery of domestication cases as ancient as *pgbd1*, *2* and *5* genes in vertebrate genomes since sufficient data are available to characterize such ancient genes. However, more recent domestication cases might well be discovered in specific lineages. This was exemplified here with *pgbd3* and *pgbd4* genes that were previously considered specific to the primate lineage,[6] while they were in fact specific to a sub-clade of the Haplorhini. Interestingly, we identified the Strepsirrhini

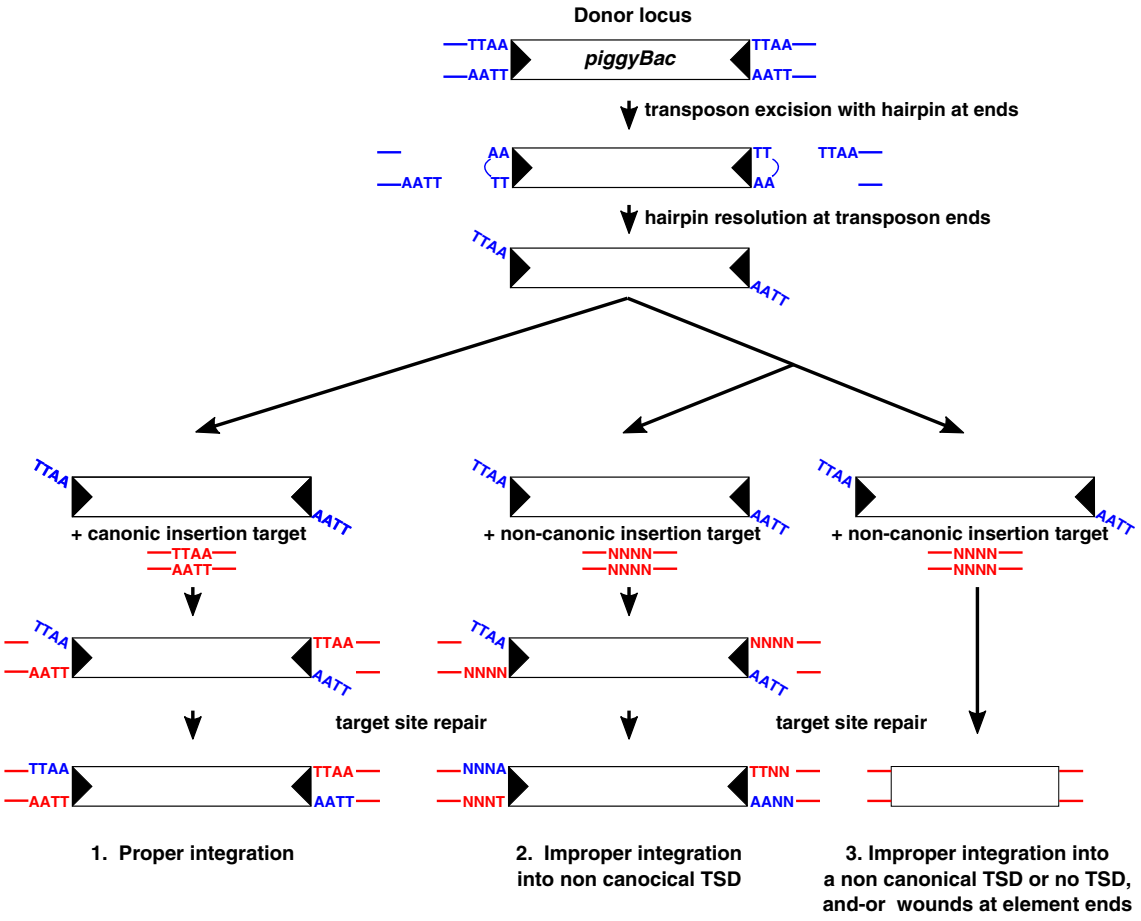


Figure 6. Signatures of proper and improper DNA integration in a *pble* transposition system (graphic derived from Figure 1 in https://doi.org/10.1101/015289). *piggyBac* transposition is initiated by nicks at the transposon ends. The exposed 3′OHs then attack the complementary strand 4 nt inside the flanking donor DNA to form the hairpins on the transposon end. The hairpins on the transposon ends are nicked at the 3′ transposon ends. In a proper integration event (pathway 1), transposon ends thereafter attack the TTAA target sequence at staggered positions, forming covalent links between the 3′ end of the transposon and the 5′ ends of the target site. The single strand gap between the 3′ ends of the target DNA and the 5′ ends of the transposon are repaired to generate the four bp TTAA target sequence duplication. In an improper integration event, transposon ends thereafter attack a random target sequence at staggered positions, forming covalent links between the 3′ end of the transposon and the 5′ ends of the target site. The single strand gap between the 3′ ends of the target DNA and the 5′ ends of the transposon are repaired to generate a non-canonical target sequence duplication (pathway 2). This process can also lead to wounds at transposon ends (pathway 3).

as another primate sub-clade in which *pgbd3* and *pgbd4* genes were absent. We provided evidence for three new *pgbd* genes, *pgbd7*, *pgbd8*, and *pgbd9* and identified *pbles* close to the common ancestor of all *pble*s. We identified *Tcr-pble* as the closest relative of *pgbd5* gene. To summarize, Figure 3 provides a comprehensive analysis of phylogenetic relationships between all *pbles* available to date.

The paradigm of specific interactions of PGBD5 with Ifp2 and Tcr-pble ends

In spite of very limited sequence identity between the terminal motifs (Figure 7) and regions at ends of *Ifp2* and *Tcr-pble*, both *Ifp2* and *Tcr-pble* ends can stimulate PGBD5 for recombination, including transposition. Since specific recognition by PGBD5 of both *pble* ends may not rely on their primary sequence, it might be guided by another feature located within the terminal regions. For more than a decade, it has been known that there is another conserved motif in each *pble* terminal region, an inner inverted repeat (IIR).[29–30] Here, we observed in the 24 *pbles* used in Figure 3 that they display little identity between their terminal motifs (supplementary data 5a), but 100% of them have IR of at least 6 nucleotides and separated by at least 4 nucleotides at both ends (supplementary data 5b). The distance between these IRs and TIR varied at both ends from 0 to 360 nucleotides,

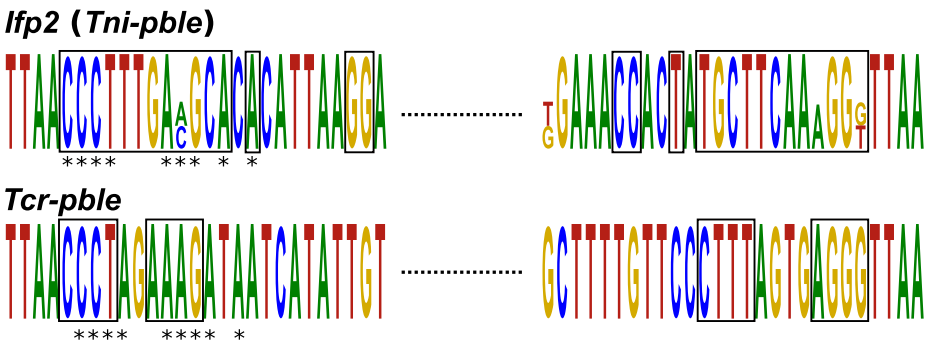


Figure 7. Gapped local alignment of the 5′ and 3′ extremities of *Tcr-pble* and *Ifp2*. The alignment was performed by GLAM2, and the logo was generated by Skyline. The black rectangles indicate similar nucleotides between the two ends. The stars indicate the similarities between the two *pble*.

with one of both IRs per *pble* always being close to TIR. Each *pble* could display the same IR sequence at both *e*nds, thus corresponding to what was previously called subterminal inverted repeats (STIR; Figure 8(a)), or different IR sequence corresponding to internal IR (IIR; Figure 8(b)). Previous studies have demonstrated that the regions spanning from the outer end to the complete STIR at both ends of *Ifp2* were required to obtain a minimal *Ifp2* element that can be actively transposed by PB.[31–33] These STIR and IIR might therefore be an important component for the binding of *pble* and PGBD transposases to their cognate DNA targets. An interesting property of these IRs is that they could extrude a non-B cruciform structure when a DNA torsion, such as a negative super-helicity, was applied to DNA.[34–36] This could be correlated with the fact that *Ifp2* has substantially decreased transposition efficiency when integration assays were monitored using a linear *Ifp2* donor plasmid rather than in a circular and negatively superhelical configuration.[37]

Further investigation will be required to determine whether other *pbles* are able to stimulate transposition and recombination events mediated by PGBD5, to define the critical role of inner IRs within *pble* ends, and whether STIR and IIR can switch between a B and non-B DNA configuration. If non-*pbles* sequences displaying such features are interspersed in chromosomes, a part of them might be bound by PGBD5 and trigger various recombination events such as deletions, inversions or fusion of chromosomal arms. Elucidating how IRs and their secondary structure affect the binding of PGBD5 to chromosomal DNA will require detailed structural and biochemical studies.

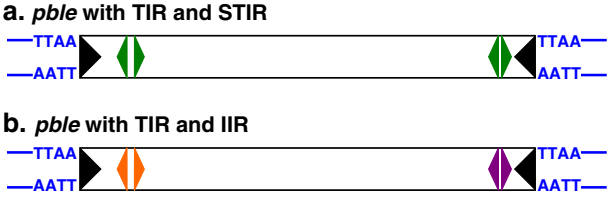


Figure 8. IR organization at *pble* ends. (a) IR organization in *pbles* displaying TIRs (black triangles) and STIR (double triangles in green). (b) IR organization in *pbles* displaying TIRs (black triangles) and IIR (double triangles in orange or purple).

Materials and methods

Data mining

The identification of *pble*s was done with a suite of tools. To annotate *pbles,* we used reciprocal blast searches to find homologous sequences, taking into account both the length of the alignment, its identity and e-value to filter results. This was used to launch BLASTP and TBLASTN. The query set consisted of 7 sequences: human PGBD1-5, PGBD5 from the cephalochordate Branchiostoma floridae (Bfl-PGBD5), PGBD5 from the lamprey Petromyzon marinus (Pma-PGBD5), and KOBUTA from the Xenopus sp. We performed BLASTP against the NCBI nr and transcriptome shotgun assembly (TSA) nr, RepBase[38] and protein data supplied in Bouallègue *et al.*, 2017. TBLASTN was used on NCBI WGS, NT/NR and TSA databases. Due to the large number of sequenced genomes, an important number of orthologues and paralogues of each PGBD was present in databases. In order to manage this issue, two queries per search were used for TBLASTN analyses: all species except mammals for *pgbd1-2* and all species except primates for *pgbd3-4*. To identify the closest pble to *pgbd5*, we targeted as subject all species except vertebrate genomes. For TBLASTN only sequences with an e-value between 0 and 1e-80 and with a complete transposase were retained. To verify if sequences belong to the piggyBac family, we performed HMM searches (hmmsearch) with HMMER 3.2.1 and translation following by Pfam domain searches. After verification, nucleotide hits were extended to 3 kb on each side and extracted with EFetch. TIRs were identified using custom scripts. These sequences were translated using the ExPASy Translate tool (https://web.expasy.org/translate/).

Sequence alignment and phylogeny

In order to determine the relationships between *pbles* and PGBDs, and because the DD[D/E] domain of the Tnp_1_7 transposase displayed an elevated divergence between the most distantly related *piggyBac* transposases, the alignment was done taking into account the secondary structure, residue type, position conservation, position reliability and residue hydrophobicity of the protein sequences using the PRALINE pipeline.[39] We choose PRALINE because structure is more conserved than sequence.[40–41] Junctions between conserved blocks in the alignment were verified and corrected in some case by hand. For the visualization (supplementary data 3), the alignment was shaded using BoxShade server (http://www.ch.embnet.org/software/BOX_form.html). The phylogenetic tree is based on an alignment covering approximately 400 residues. The tree was generated by a Pthreads version called raxmlHPC-PTHREADS-SSE3 Version 8.2.12 of RaXML.[42] One thousand bootstraps were used to infer branch support and we used the model of amino acid substitution PROTGAMMABLOSUM62.

Substitution rate analyses

Amino acid alignments were performed using MAFFT [43] Version 7.407 (--reorder --adjustdirection settings). Nucleotide sequences of protein-coding genes were extracted and were aligned by using PAL2NAL Version 14[44] according to the corresponding amino acid alignment. A phylogenetic tree was constructed using the protein sequence alignments with RaxML. To estimate the selection pressure on *pgbd* genes, we estimated the ratio (ω; so-called dN/dS) of the rate of nonsynonymous substitutions to the rate of synonymous substitutions using PAML version 4.9.[45] We used different parameters in the CODEML control file: CodonFreq = 2 for estimating the codon frequencies using F3X4 model, runmode = 0 for evaluation of the tree topologies and model = 0 for a single ω value across all branches, model = 0 is for one omega ratio for all sites.

Synteny analyses

The annotations available at NCBI were used to recover annotated genes around *pgbd* genes in the selected species. From these genes, proteins were recovered to run TBLASTN against selected contig/genomes. Only hits displaying more than 80% of the protein length with an identity percentage greater than 60 were used. For *pgbd7* gene, we used the genome annotations of *Trichechus manatus latirosis* (GCF_000243295.1, TriManLat1.0), *Orycteropus afer afer* (GCF_000298275.1, OryAfe1.0), *Loxodonta africana* (GCF_000001905.1, Loxafr3.0) and *Chrysochloris asiatica* (GCF_000296735.1, ChrAsi1.0). For *pgbd8* gene, only two genomes were annotated: *Microcebus murinus* (GCF_000165445.2, Mmur_3.0) and *Propithecus coquereli* (GCF_000956105.1, Pcoq_1.0). For *pgbd9* gene we used genome annotations of *Austrofundulus limnaeus* (GCF_001266775.1, AstBur1.0), *Clupea harengus* (GCF_000966335.1), *Haplochromis burtoni* (GCF_000239415.1), *Oreochromis niloticus* (GCF_001858045.2), *Pundamilia nyererei* (GCF_000239375.1, PunNye1.0), and *Takifugu rubripes* (GCF_901000725.2, fTakRub1.2).

Searches of introns and loss of TIR

The presence of introns was analysed by BLAST analyses between the proteins available in public databases and their coding genes. The analysis of the presence of TIR was done with an in-house script (https://github.com/Leelouh/G2TIR).

Sequence analysis of the pble

For each PGBD, *pble* TIR were analysed. For PGBD3 we used TIRs from *Smi7-pble*, *Hvu-pble*, *Lfu-pble*, and *MER85*, for PGBD4, those of *Oab-pble*, *Pva-pble*, *Pxu1-pble*, and *MER75*. For PGBD5 we compared *Tcr-pble* and *Ifp2* (supplementary data 6). For PGBD7 we used *Smi2-pble* and *Nvi-pble*. For PGBD8 we used *Api1-pble*, *Atr1-pble* and *Api3-pble*. For PGBD9 we used *Pfl-pble*, *Cgi-pble*, and *Bgl-pble*. For each element, 70 bp were extracted from both *pble* extremities. These were aligned using GLAM2[46] a tool that can discover conserved sequence motifs containing indels. The ends of each *pble* were also analysed with non-B DB[47] that predicts the presence of DNA structure from a primary sequence. The identity between all pairs of *pble* sequences (supplementary Table 2) was calculated by global pairwise alignment "end to end" using stretcher.[48]

Cloning of cDNA coding for the Mm523 PGBD5 isoform

A single mouse brain (strain C57Bl6) was used for total RNA extraction using Tri-reagent (Sigma-Aldrich, St-Louis, MO, USA). cDNA synthesis was carried out using Omniscript RT kit and oligo dT primers (Qiagen, Valencia, CA, USA). PCR primers with appropriate flanking restriction sites were synthesized by Eurofins Genomics, Ebersberg, Germany. PCR was performed with Phusion High-Fidelity PCR Master Mix (ThermoScientific). Following agarose gel electrophoresis, PCR fragments were extracted (QIAquick gel extraction kit, Qiagen), submitted to enzyme restriction (*Eco*RI/*Xba*I for the long N-term isoform and *Eco*RI/*Xho*I for the short N-term isoform), purified (QIAquick PCR purification kit, Qiagen) and kept for cloning. Their sequence identity was verified by Sanger sequencing (Eurofins Genomics, Ebersberg, Germany). The

primers used to amplify Mm523 (Accession N°: XM_006530804.1) were supplied in supplementary data 4d.

Integration assays

Plasmid expression for transposases. The plasmid pCS2-Mm523 encodes two myc tagged PGBD5 isoforms different in size (523 and 409 amino acid residues, respectively). The cDNA was inserted into the multi-cloning site of a modified pCS2 vector with an in-frame N-term 5XMyc tag.[49]

Plasmids donors of transposons. The plasmid pBSK-*Ifp2*-TIR5′-NeoR-TIR3′ (supplementary data 4b) was built by introducing the Ifp2 5′ and 3′ terminal regions (262 and 400 bp, respectively) into the pBluescript SK plasmid (pBSK). A cassette (NeoR) containing a simian virus 40 (SV40) promoter, the neomycin phosphotransferase ORF and a terminator was cloned between both transposon ends as described.[50] The transposon *Tcr-pble* (2467 bp) was synthesized by ATG:biosynthetic, Merzhausen, Germany (supplementary data 4c). NeoR was cloned using a *Bam*HI site that was added into its sequence during the DNA synthesis. The plasmid pBSK-NeoR was built by cloning the NeoR cassette into the multi-cloning site of a pBSK plasmid was as described.[51]

Integration assay in HeLa cells. Assays were monitored as described.[50] Briefly, each sample of 100,000 cells in a well of a 24-well plates of plaque assays was co-transfected with JetPEI (Polyplus-transfection, Illkirch-Graffenstaden) and 400 ng DNA plasmid and with equal amounts of donor of NeoR cassette included or not within a transposon and transposase sources (1:1 ratio). Two days post-transfection, each cell sample was transferred to a cell culture dish (100 mm diameter) and selected with a culture medium containing 800 μg/mL G418 sulfate (Eurobio Scientific, Les Ulis) for 15 days. After two washing with 1X saline phosphate buffer, cell clones were fixed and stained overnight with 70% EtOH-0.5% methylene blue and colonies > 0.5 mm in diameter were counted. Experiments were performed at least twice in triplicate.

Integration assay in G401 cells. Assays were monitored as described.[13] Two clonal cell G401 lines were used. The first line was lentivirally transduced to constitutively express specific shRNA suppressing the expression of Hs524 PGBD5.[13] The second line was modified as a control to constitutively express shRNA to target GFP which is not expressed, thereby preserving the endogenous expression of Hs524 PGBD5. Briefly, each sample of 100,000 cells in a well of a 24-well plates of plaque assays was transfected with jetOptimus and 500 ng DNA plasmid (pBSK-IFP2-TIR5′-NeoR-TIR3′, pBSK-*Tcr-pble*-TIR5′-NeoR-TIR3′ or pBSK-NeoR) as recommended by the supplier (Polyplus- transfection, Illkirch-Graffenstaden). Two days post-transfection, each cell sample was transferred to a cell culture dish (100 mm diameter) and selected with a culture medium containing 2 mg/mL G418 sulfate (Eurobio Scientific, Les Ulis) for 15 days. After two washing with 1X saline phosphate buffer, cell clones were fixed and stained overnight with 70% EtOH-0.5% methylene blue and colonies > 0.5 mm in diameter were counted. Experiments were performed at least twice in triplicate.

Recovery of integration sites

LAM-PCR and Illumina libraries. Integration assays were done to produce cell populations containing integrated copies of the donor transposon. Fifteen days post-transfection, cell clones were harvested for genomic DNA preparation using the DNeasy kit (Qiagen, Hilden, Germany). Linear amplification-mediated PCR (LAM-PCR) was performed to amplify the vector-genomic DNA junctions of *Ifp2* and *Tcr-pble* vectors as described.[52] All PCRs were done using the high fidelity Q5 DNA Polymerase (New England Biolabs, Ipswich, MA). For both approaches, 1 μg DNA was used for two rounds of 50 linear amplification using a biotinylated primer anchored near one end of the NeoR cassette to enrich DNA species containing transposon-chromosomal DNA junctions (for sequences of (B)-NeoR 5′ and 3′ primers, supplementary data 4e). One reaction was done per ends. The single-stranded products were immobilized on streptavidin-coated magnetic beads (Dynabeads M-280 Streptavidin, Invitrogen, Carlsbad, CA). All subsequent steps were performed on the magnetic bead-bound DNA. Two washes with water followed each step. Second strand synthesis was performed with random hexamer primers (Roche, Basel, Switzerland) using Klenow DNA polymerase (New England Biolabs, Ipswich, MA). The double-stranded DNA was split into two batches and subjected to restriction digests with *Dpn*I for the first one and *Pci*I, *Nco*I and *Bsp*HI for the second one using restriction enzymes. The DNA fragments with a CG-3′ or a CATG-3′ overhang ends were ligated to linkers displaying appropriate overhang ends and made from annealed oligonucleotides (supplementary data 4e).

To increase the specificity of the full process, a first PCR was done using one biotinylated primer anchored within the 5′ or 3′ region of the transposon donor and one primer anchored within the linker (for sequences of (B)-TIR-UTR 5′ and 3′, and LC1 primers, see supplementary data 4e). PCR products were immobilized on streptavidin-coated magnetic beads and purified as described above. Next, the bead-bound DNA was subjected to a nested PCR using nested primers anchored within transposon ends and within linkers (supplementary data 4e). Final PCR products were purified, quantified and gathered in equimolar DNA amounts for each transposon

vector (4 populations of LAM-PCR products) before to be used to make Illumina libraries using NEBNext® Ultra™ II DNA Library Prep Kit for Illumina® and NEBNext Multiplex Oligos for Illumina (New England Biolabs, Ipswich, MA). Fragment size selection, library-quality control and Illumina sequencing (MiSeq 250 nucleotides, TruSeq SBS Kit v3) were achieved at the Plateforme de Séquençage Haut Débit I2BC (Gif-sur-Yvette, France). DNA quantities were monitored at various steps in the procedure with the Qubit® dsDNA (Molecular Probes, Eugene, USA).

Computer analysis. Trimmomatic[53] was used to filter Miseq reads using default parameters, excepted for SLIDINGWINDOW:5:20 and MINLEN:100. The purpose of the following steps was to recover chromosome-inserted DNA fragment junctions taking into account the plasmid backbone regions located 100 bp upstream and downstream the *pble*-NeoR transposon. Filtered reads were first mapped against the sequence of plasmid backbone less the 100 bp regions flanking on both sides the *pble*-NeoR transposon with bwa-mem using default parameters.[54] The unmapped reads were then extracted using SAMtools view with parameters -b -f 4[55] and bamToFastq from the BEDTools suite using default parameters.[56] Recovered unmapped reads were aligned using bwa-mem against a bwa bank gathering the sequences of hg38 chromosomes plus those of the *Ifp2*-NeoR or *Tcr-pble* transposons flanked by the 100 bp plasmid backbone regions on both sides (supplementary data 4e). Default parameters were used except for -w 1 and -r 1. The bam file resulting from each dataset alignment was analysed with Lumpy in order to identify split reads.[57] The used parameters were -e -mw 2 -tt 0.0 and back_distance:20,weight:1,id:lumpy_v1,min_mapping_threshold:20. Structural variants (SV) characterized by "BND" for the broken end notations and displaying for each of them an SV with two positions, one genomic and one on the transposon, were extracted using a house python program (https://github.com/Lee-louh/lumpy2site). Results were filtered taking into account a difference below 3 between the transposon breakpoint calculated by Lumpy and the maximal spread of read alignments in the transposon donor sequence for each integration event. Each TSD nucleotide motif at the insertion site was obtained after extracting 10 bp sequences before and after the breakpoint in the chromosome sequences.

Statistical analysis

Values in graphics were medians, quartiles 1 and 3 and spread of experiments done in triplicate. Shapiro-Wilk test was used to confirm the normality of each set of samples, t-test to analyse distribution differences between experimental samples, Chi2 tests to analyse differences between an experimental distribution and a theoretical one, and logarithmic distribution test to analyse enrichment using free tools and tutorials available at http://www.anastats.fr/outils.php.

Data repository

All raw and processed data are available through the European Nucleotide Archive under accession number PRJEB36229, PRJEB36230, PRJEB41045 and PRJEB41052. Files describing the annotation of insertion sites copies in the hg38 release are supplied as Supplementary Tables 3 to 7.

CRediT authorship contribution statement

Laura Helou: Conceptualization, Data curation, Formal analysis, Investigation, Resources, Software, Writing. **Linda Beauclair:** Conceptualization. **Benoît Piégu:** Investigation, Resources, Software. **Louis Tsakou Ngouafo:** Data curation. **Thierry Lecomte:** Funding acquisition. **Alex Kentsis:** Writing. **Pierre Pontarotti:** Conceptualization, Investigation, Resources, Software, Writing. **Yves Bigot:** Conceptualization, Formal analysis, Funding acquisition, Investigation, Resources, Software, Writing.

Acknowledgments

This work was supported by the C.N.R.S., the I.N.R.A., and the GDR CNRS 2157. It also received funds from a research program grants from the Ligue Nationale Contre le Cancer, the Merck foundation, and the French National Society of Gastroenterology. Laura Helou holds a PhD fellowship from the Région Centre Val de Loire. We acknowledge the high-throughput sequencing facilities of I2BC for its sequencing and bioinformatics expertize. Alex Kentsis is a consultant for Novartis and is supported by the National Cancer Institute grants R01 CA214812 and P30 CA008748. Yves Bigot, who was in charge of the achievement of this project does not have to thank the French National Research Agency for its financial support but he kindly thanks it for the excellent reviews embellished with arguments based on scientific and cultural novelties in the expertise of his yearly application file during the last decade. Our thanks to Peter Arensburger at California Polytechnic University in Pomona CA (Cal Poly Pomona) for reviewing the manuscript for legibility.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jmb.2021.166839.

Received 14 April 2020;
Accepted 14 January 2021;
Available online 2 February 2021

Keywords:
domestication;
transposition;
DNA binding;
molecular evolution;
transposable element

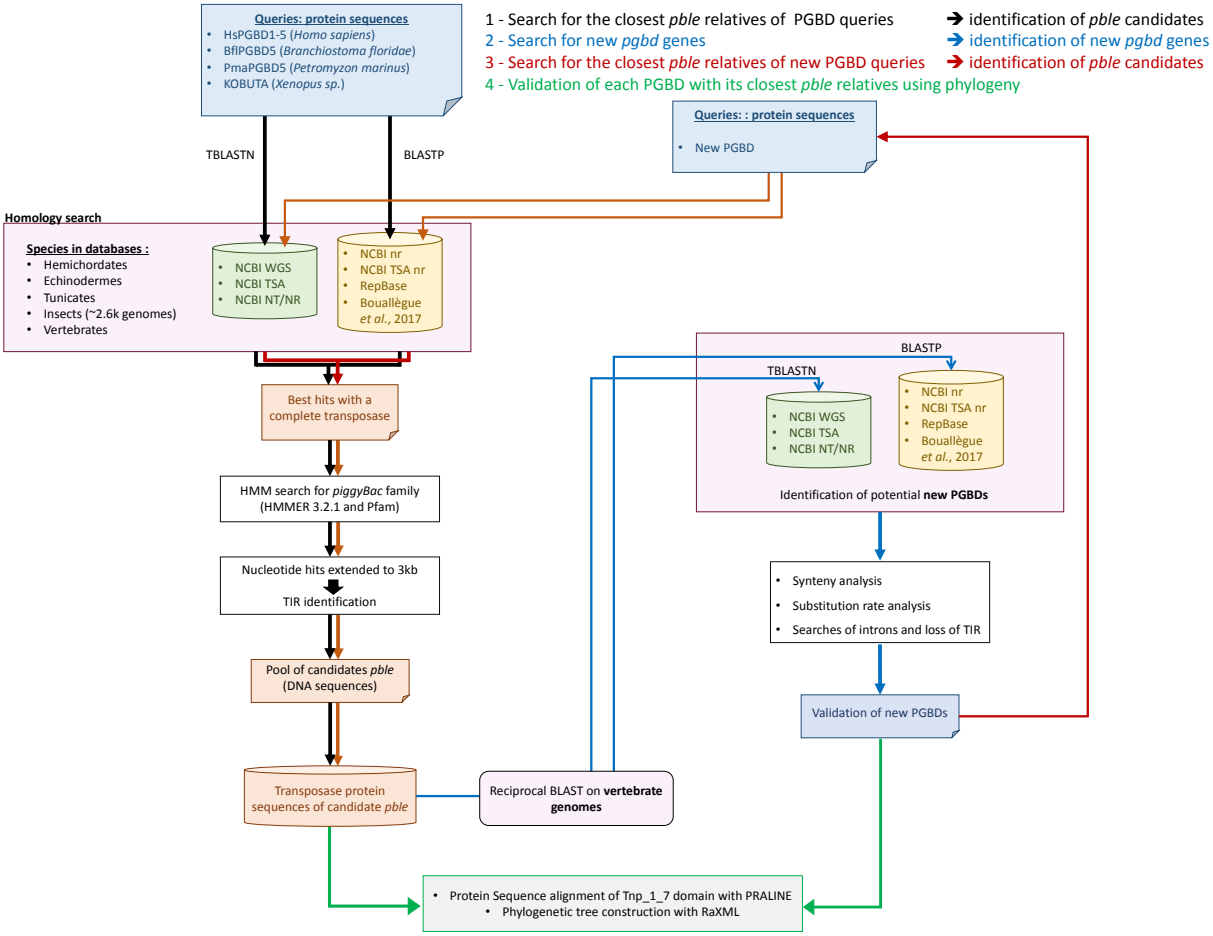
Abbreviations used:
CMV, Human cytomegalovirus; CRD, Cystein-rich domain; GFP, green fluorescent protein; LTR, long terminal repeats; NeoR, Neomycin resistance; ORF, Open reading frame; *pble*, *piggyBac*-like element; PGBD or *pgbd*, "*piggyBac* derived transposase" protein or gene; STIR, sub-terminal inverted repeats; SV40, simian virus 40; TE, transposable element; TIR, terminal inverted repeats; TSD, target site duplication

References

1. Piégu, B., Bire, S., Arensburger, P., Bigot, Y., (2015). A survey of transposable element classification systems - a call for a fundamental update to meet the challenge of their diversity and complexity. *Mol. Phylogenet. Evol.*, **86**, 90–109.
2. Goerner-Potvin, P., Bourque, G., (2018). Computational tools to unmask transposable elements. *Nature Rev. Genet.*, **19**, 688–704.
3. Fraser, M.J., Smith, G.E., Summers, M.D., (1983). Acquisition of host cell DNA sequences by baculoviruses: relationship between host DNA insertions and FP mutants of Autographa californica and Galleria mellonella nu- clear polyhedrosis viruses. *J. Virol.*, **47**, 287–300.
4. Mitra, R., Fain-Thornton, J., Craig, N.L., (2008). piggyBac can bypass DNA syn- thesis during cut and paste transposition. *EMBO J.*, **27**, 1097–1109.
5. Mitra, R., Li, X., Kapusta, A., Mayhew, D., Mitra, R.D., Feschotte, C., Craig, N.L., (2013). Functional characterization of piggyBat from the bat Myotis lucifugus unveils an active mammalian DNA transposon. *Proc. Nat. Acad. Sci. USA*, **110**, 234–239.
6. Bouallègue, M., Rouault, J.D., Hua-Van, A., Makni, M., Capy, P., (2017). Molecular evolution of piggyBac superfamily: from selfishness to domestication. *Gen. Biol. Evol.*, **9**, 323–339.
7. Morellet, N., Li, X., Wieninger, S.A., Taylor, J.L., Bischerour, J., Moriau, S., Lescop, E., Bardiaux, B., et al., (2018). Sequence-specific DNA binding activity of the cross-brace zinc finger motif of the piggyBac transposase. *Nucl. Acids. Res.*, **46**, 2660–2677.
8. Sinzelle, L., Izsvák, Z., Ivics, Z., (2009). Molecular domestication of transposable elements: from detrimental parasites to useful host genes. *Cell. Mol. Life. Sci.*, **66**, 1073–1093.
9. Joly-Lopez, Z., Bureau, T.E., (2018). Exaptation of transposable element coding sequences. *Curr. Opin. Genet. Dev.*, **49**, 34–42.
10. Huang, S., Tao, X., Yuan, S., Zhang, Y., Li, P., Beilinson, H.A., Zhang, Y., Yu, W., et al., (2016). Discovery of an Active RAG Transposon Illuminates the Origins of V(D)J Recombination. *Cell*, **166**, 102–114.
11. Zhang, Y., Cheng, T.C., Huang, G., Lu, Q., Surleac, M.D., Mandell, J.D., Pontarotti, P., Petrescu, A.J., et al., (2019). Transposon molecular domestication and the evolution of the RAG recombinase. *Nature*, **569**, 79–84.
12. Kapitonov, V.V., Jurka, J., (2005). RAG1 core and V(D)J recombination signal sequences were derived from Transib transposons. *PLoS. Biol.*, **3**, e181
13. Hencken, C.G., Li, X., Craig, N.L., (2012). Functional characterization of anactive Rag-like transposase. *Nature Struct. Mol. Biol.*, **19**, 834–836.
14. Sarkar, A., Sim, C., Hong, Y.S., Hogan, J.R., Fraser, M.J., Robertson, H.M., Collins, F.H., (2003). Molecular evolutionary analysis of the widespread piggyBac transposon family and related "domesticated" sequences. *Mol. Genet. Genom.*, **270**, 173–180.
15. Pavelitz, T., Gray, L.T., Padilla, S.L., Bailey, A.D., Weiner, A.M., (2013). PGBD5: a neural-specific intron containing piggyBac transposase domesticated over 500 million years ago and conserved from cephalochordates to humans. *Mob. DNA*, **4**, 23–39.
16. Delsuc, F., Brinkmann, H., Chourrout, D., Philippe, H., (2006). Tunicates and not cephalochordates are the closest living relatives of vertebrates. *Nature.*, **439**, 965–968.
17. Putnam, N.H., Butts, T., Ferrier, D.E.K., Furlong, R.F., Hellsten, U., Kawashima, T., Robinson-Rechavi, M., Shoguchi, E., et al., (2008). The amphioxus genome and the evolution of the chordate karyotype. *Nature*, **453**, 1064–1071.
18. Henssen, A.G., Henaff, E., Jiang, E., Eisenberg, A.R., Carson, J.R., Villasante, C.M., Ray, M., Still, E., et al., (2015). Genomic DNA transposition induced by human PGBD5. *Elife.*, **4**, e10565
19. Henssen, A.G., Koche, R., Zhuang, J., Jiang, E., Reed, C., Eisenberg, A., Still, E., MacArthur, I.C., et al., (2017). PGBD5 promotes site-specific oncogenic mutations in human tumors. *Nature Genet.*, **49**, 1005–1014.
20. Kryazhimskiy, S., Plotkin, J.B., (2008). The Population Genetics of dN/dS. *PLoS. Genet.*, **4**, e1000304
21. Inoue, Y., Saga, T., Aikawa, T., Kumagai, M., Shimada, A., Kawaguchi, Y., Naruse, K., Morishita, S., et al., (2017). Complete fusion of a transposon and herpesvirus created the Teratorn mobile element in medaka fish. *Nature Commun.*, **8**, 551.
22. Arkhipova, I.R., Yushenova, I.A., (2019). Giant transposons in eukaryotes: is bigger better?. *Gen. Biol. Evol.*, **11**, 906–918.

23. Hikosaka, A., Kobayashi, T., Saito, Y., Kawahara, A., (2007). Evolution of the Xenopus piggyBac transposon family TxpB: domesticated and untamed strategies of transposon subfamilies. *Mol. Biol. Evol.*, **24**, 2648–2656.
24. Henssen, A.G., Reed, C., Jiang, E., Garcia, H.D., von Stebut, J., MacArthur, I.C., Hundsdoerfer, P., Kim, J.H., et al., (2017). Therapeutic targeting of PGBD5-induced DNA repair dependency in pediatric solid tumors. *Sci. Transl. Med.*, **9**, eaam9078.
25. Elick, T.A., Lobo, N., Fraser Jr., M.J., (1997). Analysis of the cis-acting DNA elements required for piggyBac transposable element excision. *Mol. Gen. Genet.*, **255**, 605–610.
26. Li, M.A., Pettitt, S.J., Eckert, S., Ning, Z., Rice, S., Cadiñanos, J., Yusa, K., Conte, N., et al., (2013). The piggyBac transposon displays local and distant reintegration preferences and can cause mutations at noncanonical integration sites. *Mol. Cell. Biol.*, **33**, 1317–1330.
27. Wang, H., Mayhew, D., Chen, X., Johnston, M., Mitra, R. D., (2012). "Calling cards" for DNA-binding proteins in mammalian cells. *Genetics.*, **190**, 941–949.
28. Thorvaldsdóttir, H., Robinson, J.T., Mesirov, J.P., (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.*, **14**, 178–192.
29. Wu, M., Sun, Z.C., Hu, C.L., Zhang, G.F., Han, Z.J., (2008). An active piggyBac-like element in Macdunnoughia crassisigna. *Insect. Sci.*, **15**, 521–528.
30. Daimon, T., Mitsuhira, M., Katsuma, S., Abe, H., Mita, K., Shimada, T., (2010). Recent transposition of yabusame, a novel piggybac-like transposable element in the genome of the silkworm, Bombyx mori. *Genome*, **53**, 585–593.
31. Solodushko, V., Bitko, V., Fouty, B., (2014). Minimal piggyBac vectors for chromatin integration. *Gene. Ther.*, **21**, 1–9.
32. Troyanovsky, B., Bitko, V., Pastukh, V., Fouty, B., Solodushko, V., (2016). The Functionality of Minimal PiggyBac Transposons in Mammalian Cells. *Mol. Ther. Nucleic. Acids*, **5**, e369
33. Chen, Q., Luo, W., Veach, R.A., Hickman, A.B., Wilson, M. H., Dyda, F., (2020). Structural basis of seamless excision and specific targeting by piggyBac transposase. *Nature Commun.*, **11**, 3446.
34. Brázda, V., Laister, R.C., Jagelská, E.B., Arrowsmith, C., (2011). Cruciform structures are a common DNA feature important for regulating biological processes. *BMC Mol. Biol.*, **12**, 33.
35. Kaushik, M., Kaushik, S., Roy, K., Singh, A., Mahendru, S., Kumar, M., Chaudhary, S., Ahmed, S., et al., (2016). A bouquet of DNA structures: Emerging diversity. *Biochem. Biophys. Rep.*, **5**, 388–395.
36. Guiblet, W.M., Cremona, M.A., Cechova, M., Harris, R.S., Kejnovská, I., Kejnovsky, E., Eckert, K., Chiaromonte, F., et al., (2018). Long-read sequencing technology indicates genome-wide effects of non-B DNA on polymerization speed and error rate. *Gen. Res.*, **28**, 1767–1778.
37. Nakanishi, H., Higuchi, Y., Kawakami, S., Yamashita, F., Hashida, M., (2011). Comparison of piggyBac transposition efficiency between linear and circular donor vectors in mammalian cells. *J. Biotechnol.*, **154**, 205–208.
38. Bao, W., Kojima, K.K., Kohany, O., (2015). Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA.*, **6**, 11.
39. Simossis, V.A., Heringa, J., (2003). The PRALINE online server: Optimising progressive multiple alignment on the web. *Comput. Biol. Chem.*, **27**, 511–519.
40. Sander, C., Schneider, R., (1991). Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins*, **9**, 56–68.
41. Rost, B., (1999). Twilight zone of protein sequence alignments. *Prot. Engin.*, **12**, 85–94.
42. Stamatakis, A., (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinform.*, **30**, 1312–1313.
43. Katoh, K., Rozewicki, J., Yamada, K.D., (2019). MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Brief. Bioinform.*, **20**, 1160–1166.
44. Suyama, M., Torrents, D., Bork, P., (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucl. Acids. Res.*, **34**, W609–W612.
45. Yang, Z., (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.*, **24**, 1586–1591.
46. Frith, M.C., Saunders, N.F., Kobe, B., Bailey, T.L., (2008). Discovering sequence motifs with arbitrary insertions and deletions. *PLoS. Comput. Biol.*, **4**, e1000071
47. Cer, R.Z., Donohue, D.E., Mudunuri, U.S., Temiz, N.A., Loss, M.A., Starner, N.J., Halusa, G.N., Volfovsky, N., et al., (2013). Non-B DB v2.0: a database of predicted non-B DNA-forming motifs and its associated tools. *Nucl. Acids. Res.*, **41**, D94–D100.
48. Myers, E.W., Miller, W., (1988). Optimal alignments in linear space. *Comput. Appl. Biosci.*, **4**, 11–17.
49. Travnickova-Bendova, Z., Cermakian, N., Reppert, S.M., Sassone-Corsi, P., (2002). Bimodal regulation of mPeriod promoters by CREB-dependent signaling and CLOCK/BMAL1 activity. *Proc. Natl. Acad. Sci. USA*, **99**, 7728–7733.
50. Bire, S., Ley, D., Casteret, S., Mermod, N., Bigot, Y., Rouleux-Bonnin, F., (2013). Optimization of the piggyBac transposon using mRNA and insulators: toward a more reliable gene delivery system. *PLoS. One.*, **8**, e82559
51. Bire, S., Casteret, S., Piégu, B., Beauclair, L., Moiré, N., Arensbuger, P., Bigot, Y., (2016). Mariner transposons contain a silencer: possible role of the polycomb repressive complex 2. *PLoS. Genet.*, **12**, e1005902
52. Bartholomae, C.C., Glimm, H., von Kalle, C., Schmidt, M., (2012). Insertion site pattern: global approach by linear amplification-mediated PCR and mass sequencing. *Meth. Mol. Biol.*, **859**, 255–265.
53. Bolger, A.M., Lohse, M., Usadel, B., (2014). Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*, **30**, 2114–2120.
54. Li, H., Durbin, R., (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, **26**, 589–595.
55. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., 1000 Genome Project Data Processing Subgroup, et al., (2009). The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
56. Quinlan, A.R., HallI, M., (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics.*, **26**, 841–842.
57. Layer, R.M., Chiang, C., Quinlan, A.R., Hall, I.M., (2014). LUMPY: a probabilistic framework for structural variant discovery. *Gen. Biol.*, **15**, R84.

Supplementary Data 1: Graphic representation of the data mining process to inventory *pgbd* genes and their closest *pble* relatives.



Supplementary data 2. : Localization of PGBD3 and PGBD4 ORFs and respectively MER85 and MER75.
Fasta sequences of localization of PGBD3 and PGBD4 genes. PGBD3 sequence is reversed (complementary strand). ORF of genes is underlined in green and TIR of the element are in bold. For PGBD3, the first bold sequence corresponds to position 1 to 43 of MER85 and the second to the positions 39 to 140. For PGBD4, the first sequence is for positions 68 to 341 of MER75 and the second one is the positions 379 to 514. TSD were not identified for MER75.

>NC_000010.11:49515105-49524281 Homo sapiens chromosome 10, GRCh38.p13 Primary Assembly (PGBD3)

ACATCTTTAA**CCCATTTATG**T**CTGAGGTT**A**CAATTTTTTGAATTTTTGCATGA**AATTCATATATGTTACC
TAGAAAAATTCAAGGAACAAAAATTTGAAGAGAAAAAAAAATTTATTATGTTCCATTGTTATGTGACGTT
CCAAAATTGGAACACTAGGCAAAACCACTGCTACACTGTACATAATGTATAAAACTATGTTTCTTATCTC
AGGAGGTGGTGACACCTGCTATTCAGTGTGATATTCAACGGAACACTTCACATGTAAGGCAACATCACAT
TTTTCACATCGAAAAGTTGTGTTCTTATGACATTCAGCGCATCGCGTTTGCTTTCCCTGTTTGACTATCA
CATGATTTATGCCATCATAACGTGAGTCAATGTTACGCTTCTGAGGTCTTCCTTTTTGGCCAGGTTCTGG
AGGATGACCATGGGTCTCCAGATAATGGCATACCACACGTCGACGAAACTCCAGAAAATCCACTGGTTTC
TCATCATATGTTTTATGCAATTGCCAAGCATTTTGTAAGACCAGTTCGAAACAGAACAAAAGAGGGCTTG
AATACCATTTCTTTCCACGGATTGATGCCCGATACTTATCAATGTTTTCATCAGCTCTGTCTACGCCTCC
CATGAACTGGTTATACACTTTGATCATGTTTGGCTGCTGAACTTGTATCTTCTTTTTCAGTTTCTGGGAG
TAACGACTGACAAGACACAGGGGATGGATACCAGCACCAGATGAGGCAACAGTGACAACACTGTTATCAT
TCCATCTGCAGACAATATTGCCTTTGCCATCAATTCGATAATCAAATGTGCCTCTTTCTTTTTTCTTTAA
AGCTACATCTGATTCCAGTGGAACTCTGTCAATGTGATCCTTTCTCACTGTACCTGTTGCCTGATGTCCC
ATTGAACTGAGCTTATCAAGAAGTGCAATACTGGTGAAAAAGTTATTGAATACAAAATGGTATTGTCCAG
GGTGTGCCTCTGTAAGTGCCTCACTAAACTGAAGGACAAGTGACGCACCGACACCATATTCCTCATGTTT
AGTATTTGGGTTTTTACCCTGATACGGCTGAAACCAGCAAATGTAGCCCAGACAGGTGGCACCACACCAA
AACTTATAGCCAAACCGAATGGGCTTTCCCCGAATAAATTGTTTGCACCCGTGACGACCAAAATAAGGAA
CCATGAATTCATCAAAGCTGAAATATGTTTCATTTGGAACAAATTTCATGCATCTCTCATTAAGTTTGCT
TATGAGAGGTCGCAATTTGGAAAATTTGTCCACTGGATCCAAATTTGCATTGTCAGCAACATGCAAATTA
GAAAATATAGTTTCAAACCGGTCACGTCTCATGGCAGCACTAACCAGTACATTATGCACATCTGTTCTTT
GTTCCCAAAACATACGCCTTCTAGGAACTGAGACATAACCACTCAGAAAAATAATTCCCAGAAAACATTT
GAATTCAGAGCTAGTCAAGCCAAGATGTACACCTTTACTGCAAGCATATAAGTTGGAGTACTTGACAATG
AGTTCAATGACCTCGTCATCAAGAAAAAGTTCAAGAATTTCTGTGGGAGTTCTCATTACGGTGAAGAAAT
CGTTTGGTGGTGCTGTAACTCTACCTGCTACGGGTTGTACAGTTAGGTCGGCTTTTTTCCATTTGCAAAG
AATTTTTGTCATTTTCCTCCTCCTTGATGGTGGAGGTTGCTGCACAGTAAACGTAGATGGAACTTCATCA
GGAGAGTCATCTTTAGGTGCGTATGAGGGATCATCTGAGTCAGACTCAGCATCAGAGCCATCTTGAATAA
GATACGCAGCTGTGTGCAACAAAGAACCTGGCAGATTATTTATTGTTCCACCTTCTTCATCTCCTGATTC
CTCATCAGAAACAGGTGCTGTAGCATTTTCAGGTGGTTGTATCACTATAGCACTTGCTTCTATGCTGTCA
TCTGTCTCTAAAAGGTCAGTTATTTCATGTAAACTTAGTGTTCGAGGCATCTTGGGACTAAAACGGAAAA
AAGGTATTATTAGTCCTAGTGGTAGTGGTTTTGCCTAGTGTTCCAAAAATGGAACATGAAAAATGTCAAA
CCCATAACTAATGCAATAATAATACTTTCTGTATTTTCTTTTTACCGCACAATAGCATAACTAGTAGAGA
AGTTATTTCTTAAAAATTACTAGTATAGGATGTAGTGCAGTAAAGTTATATGTACCTGTCTTTGGATACT
CAAAAATGTTTCTGAAAACTCACATCACTTTTTGACAAGCTCCCAGCATGCCAGTATCTGTAGCAGTAAC
TAACTCTGTGCACCAAGTGATAACATGTACCCTCAGGAGGCAGA**CATG**T**GTGAAAAATCAGACCTTGGC**A
ATGACCTTGAGCAGTAGGGC**AT**G**AATAACTCCCACATGCTTAGCGTTCCAATAATGGAACACTAGGCATA
AATA**GG**TTTTAAAAGTGTAACTTTTTCTTTCAACTTATAGATTCCTAAAAGAAAATGAACACAGTCTTCT
TTCTTTTTTTTTTTTCTCTCAGGCTCCAGTGATCCACCCACCTCAGCCTCCTGGGTAGCTGGGACTACAG
GCGCACATCACCACACCCAGCTAACTTTTTCTATTTTTGGGGTTTCACCATGTTGCCCAGGCTGGTCTCG
AACTCTTGAGCTCAAGCAATCCTCCTGCCTTGGCCTCCTAAAGTGCTGGGATTACAGGTATGAGCCATTG
TGCCCGGCCAGGCTTATTTCTTGAGAAAAATGAACTCTCAGAGATGTTAAGAACATTCTTAAAATAATTG
TATTTAGGTTTCTGGATAAAATCTCAAACTTTTACTAAATTCAGATTTTGGGAGAGTTTTGTTTTTTCTC
TTTAATTTGGGCTTTAAAAAAAAGGCACAATCAAATCAACCTCCACATTGTGAGTTGGCCTTTGCCGAAG
CACAGCCTAAAATTGCTAAATCCTCATCGGGTGCAAACCAGGATCCAACTGAAGGACAGACAGGACAATC
CCAGTGCAAACAGGTGCTAGACAGAAGGGAGAAGTTCCCAGAAGACAATAAAATGGAGACAGGCTGTGAG
GTTGGTGGTTCTCAGGGCACTGCCTTTAGAAGCATTTCAGGATAGGAATGGGACATTGGGTTGGAGTTGG
CCCCCATCAAGTGGAGGGAGAAACCAGGCACTGAAACAATGCTTGTCCGAGTCTGCTGCAGCAAACCAAG
CTGCGAGATCACTCTCATGTGTCCCACACCACCTTCAGACCCCTTCTCCTTCTCCAAGCCCTTCCACTGA
CTGCTGCCTTGAGCAAATGACTTCATCTCTCTGGGCTCAAAGTCCTATCTCCTGGGTGGCTGTAAGGATT
AAACGAATTAACCTATATAACAAGAAGAACCACAAGAACCACTAAGTCGTACGCATTCAAACATGTCAGC
CATCATTGTTCTTGTGCACCACTCCCGCCCCTGCCCCCACCTCCACTTGATGCCACCTTTGCTCAGGCCC
TAAAGGCTGCAACGTTCCAAATCCACAAGACTCTCTTCATAACAGTCATTCTACTTGGCTTGACCCCTAG
CATCTGAAATGCTGACAGCTCCCTCCTCAGAACAGTGGATGTGCACAACCCAACCCTCCCAAGCCAGGGT
TCAATGGGGCCCTGGCCACAGAGGCCCAGTGCTCAGGGGAAGGACAATGAAGAGTATTATCTTCGCAGGA
GATTATGTCTCACAGACTATTAGAAGACATTCTGTAAGACTTACACAATGAATGGAATCTTTATCACCCC
TTTACATCTCAGAGTTACTCATGTAAAAAAGGTAAAAAATAGGGCACAAGAAAAACCTTTAGGATCAAAA
ATTTATTAAAAACCAATTATATCAACAGGCATCAAGTCTACAGATTCAGGTTACACCAGACCATGAAGTA

AATTCTGTCCCCATCCACACCATACTTGCCAGGTCTTCTAGACTCCTGAGCCATCTCCCTATATCCTCAT
CCAAATTCCCAGATTACAGGCTTAGGTTTTTGGTTTGGTTTTCAGTCCAATGGAGGTGGGGGCAGCTATG
TGTTGATTTTTGGCACCACCCTGTGGTCATACCTAAATATTGCACCTTCTACTCAATCCCACAAAGGGAA
GGAAACATGTATGTAGAGGGCCCAGCCGACTTCAAATCTCAGCTCTACTTCACAGTAAAACTGTGGGCAA
GCCAATTAACTGGAGGGGCCGTGATTATAACCATGTAAAGATGTGTATTCACAGAACCAAAGACTGGAAT
GACATGTAGAAAATAAAAACATAAAAGAGGTCATTTTTTCATCTTAATATTGTTGTCATACTTTTACAAT
AAAAATAGGGACTCATCAGAAGTGCTGAAAGGTGGCGAGGGAGCCTTCAAACCATGCAGAGGAACCAACT
CAAATCCACTTCAGGCTATGCATTTAGCAGCTTAGCCCAATGCACCAGGTCATCTGAGGAGACTCAGGCA
CCGCCTACTGAAGAGTCCGTTTCTGAACTCCCCACTATGTTTCCCTGATCGGGGTGGGCATGGGTCGGGG
ATTTATCCCAGAGGCAATTATTTGCTGAGTGTGGGCCATAGTAAAGAAAAAAGCATTACATTCTCAAGCC
TCAACATCCTTACCAACAATGCTTCCCAACTGCCTCAAAGCTCTCCTAAATGAGAACATAGTTCTTTCTG
AGCAAGGTCCTGTGGACCATGAAGAATGTCACCAAGCTCCCCTCAGAGTCAGCGGGAGCTCAGCCAAAGC
ACAAGTGCAGTGCCCAGCTCCTCCCACTCTGCACCTGCTGCCTCAGACTCCCCACGCTGAGCCCAGGCCC
CTACCCTCTGAAGGTGTTTCCCATGTGATTCTGACACACACACCCCACAAGAACCAGATGATCTATGACA
TACAGCATTTAGCTAAGGCTCCCTGCCAGGCCCTCCCCTGTGTACATACAACTGTGCAGGGCACATAAAA
GAGGATTTTTGAGAAACTGATATACAAACTGCATCAGCCAAAAGCTGCAGCCCCAAAAAAGAGAAATACA
TAAGACTGACTTATTTCCATCCCTAGCTCCCAAACCTGACAGTGATCAGAGCCATCTGGGAACATTTTTT
AAAAGTCCTGCCCGGTGACCCCACGCTACACTTCCTAAGTCAGCGTCTCCAGTGTTAGGGGCTCCCAGTG
ACACTGCATAATCAGTCTGTGGGTGGGGAAGCTTTGCTCTCGTCTGTGCTCGTTGCCCTCACTGTAGTCT
GCGAGGGCAGCAACTTAAATAGTAACTGCCGATACATGAAAATAGAGTAAGAACTTGGGGGAATGAGTTT
AAGTCAAATGATAAACACTAACCCTAGTCCTGCCAGCAAGAAGAGACAAAAACATATCTGGAAATGCATT
TTCTCACTCAGGATGCAACCCATCTATGCCCTGCAGCCTGACCAAGCAGTGCTGACCACTGCTGGACAAA
GAGGCAGGAGTCAGTGACATCTGTTTTATTATGTCAAGGCTGGCAAATACCTAGCATGAATGTTGCTAGT
TCCCCACTGTGCCTATGATAGACATCACTAAGGAACTGCACACTCTTCCCAATCCACCCAAACAGGGCCT
CGGAGCCCCCTGAATAGACTACCACTCCCCAGAAATCAGAGCTGGCTGGAAAGTGAAGCCCACCTGCCCT
CCCTGTCCTCAGCCTCTCTGCACATTGTGGCATCACACCAACCTCCAAGGCAATCCAAATGAGAACCCTG
AGGAAATAATGTGCCTCCAAAGCAAAGCAAATTCCTTCCCTTCTCAATGATGCACCTGGAGTTTAGGGGC
AAATGCGTAACATTGTTAATTTGGATAAGGGTCATATTAATATATCTGAGCCACCCAACAAGACAAGGCC
CTGCACTGGATGAACTTCCCCAACAGTCTCTCCCAAATGCCTCAGTCTGAATTCATTATTCCCTTCACAC
TTGGTTAGGTCCTATTTGTGCCCAAGTCTATTAACAGACTGCAAATTTTTTGCCAGAAACCCTTGAAGCC
AGTCCAAGGATGCAGACATGCATTGTTGACATGAAAGCAAAGATAAAGCTATGGAGGCTTACCTCGAATA
TGGCAACTGGGAAGTGAATCAATAACATGGAGGTAACAAGGACTGGGGACAAGCGACCTGATCTTTATCA
CCCTGAAGGGGAAAGGTTCACATGAAATGTTTGGTCTTCTGGGCCACACAGGGCTGTCTTGGCTGCACAG
CCTGTACTAGGTAAGCAGCATGTAGGTGCTCACCGGGAACCCACATCACTCAAACACTGAAGGTGTGTTA
GTCCTAATCTTTATTTGGCACCCGGAGCTCCCAGGCCCACAATGTGCAAAGCAGTAGTATATTTTAAGAA
AGCACTAGACCTCACAACCAACAGCCATAAGCAAATGGAAACCGTGGTAGCACAAAGTAAAATGAAAAAA
TAATGTATGTTTTACAAGACCTGTATACCACTGACCATAGAACCGACAGAACATTTACTTCTAAGCAGTG
GCTCATCCATCCTGTATGACATGAAAGGGTATGAAAACCAGGGAACCCACTGAGGAGTCAGCCTCCAGAT
ACAGTTCAGTGCAGAGGAGGAGGCAGGAGAAACACATGAGAAGGGGGAATACACGGTTTGAAAAATGGTT
AAAACTTCAAAGACTGAGCTGAGGGCTTGAGCCAAATGAAGGAAAGCAGGATCCAAAGGGCGACTGCCCA
GTTCCATTAAAAAAGCTATGGAAAGGGCTTGACGGAAAATTCCTAGTAAGTATCAGCTAAAACAAAGGCA
CCTAGTTAGGGTGAACCTAGTGCCAGTGAAAGGCCGATGGCTCTCAAGCTGAGAGAATAAAGAGGATGAG
ATATCCAGAAGTCACCCACTCTCACTTCCTAGAGTTGAAAAGAGCATTATAATTAACAGGCTATACCCTC
AGTAAAAGGTGTTAAGTTCAGCAAAGGGTGCCTACCCTGGGAACTACTCCTAAATTCTAACCACTAAAAA
AAAACTAATTTTCTATGTTGTAAACTCTATGACTTTATTCTAATCATCCATCGTCAAACCACTGCTAAGA
CAAGCAACCCTAAAGGTCTTTCTGTTTCTCAAGGGAATAGCATTACCATTTCAATTAGTCTACAAAGATC
AGCAATGAGTAACAAAATATAGAGAGACTGGCAGTTCTAATATGCTTCTCAATCCTGAGCATTTTATTTT
CTTTCATCCCCCAAACCAAACAGCACCAAAGGAGATGTAATAACAAATATTTCAATGTGGATATCTGCAG
TTTAACAGCCTTAATAACTGTAGCTATAAAATGTTAAGAGCTATCGTATTTTGGTGGCAATTAAGAGTTA
TTGATATTTCTGAAATTCATTTGAGATTTCACTGTAGCTTTAGCAAAATATTGGCTCTGGAAATAATAAA
TTGTTTTACAATCTTAACATTATTAAGAAAAACTTATTGCCATCAAATTAACTAAACCCACATTTAAAAT
TCAATCAAAATAAATTGCTCTTGATCAAAATACTGATATATTTTAAACATATTTAATATTGACCCAATTT
AATTGGAAAAATATATATTGCTCCATTTATAATCAATATGTTGAGATTAACACTTTTATTAATTCCTTGC
AGTAGTTAAGAAAAAGTCTGACACCGCTCACTGTTCAAGAGTACTGGGCAACATTCCTCAAACAACTTCA
CAGGTAATAAACGTGCACAGAAACGAAAGCTGTGCTGAGGGATGACAAATATAAGCCAGCCCTCATCTGT
GCATTACTACATAAGGTGCTACTAGAATGTGAGTGCCGCAACTCTAAAGAACAGGATTTAAGTCACAGGT
CCTCAAATTACTGTTTAAAAGAATTGACACATCTATTTCGGAAAGGTTATGGAATGTCTGTAGTGTCTGG
ACGGGGGTTAATGAGCAGTTCTAAGTAACAAGAAAAAACACCAGGTCAACCTTCACTCACAAAGGCCAAT
AAATACAACAAACTAGGTCTCTTCACCTTGAGAAATAAGAAAAAACTGTTGAGGATTGGGAAAAAGCCAT
CCCACTATCAAGTATTTTAAAAGTAATAAGGTTCCTTTTTCCCTTTGACTTCCTTAGAGCTAAGTGAGCT
CTAAGTCAAGGAGACTTGAAATGCGCGGCAGTAGCACAGTGTTCAGTGGGCCACCATGGGGAGGCCTGCA
ATACAGATTATTTTTCCTGGCTCTCAGAAAACGTGTGGCAATATGCAATTCAGACTTCATGTTTCACATG
TTCACCTTCCCTTTTATGGTGCTTGCTTTATACAATGACAATGTCATGGAGCTAAAAGCCAAAGTCAGTC
TCTAACATAAAAAACAACAGTGAAAAGAAAGGTATGCTGTGCATACAAAGTACACTTTTTGCATGATGGT

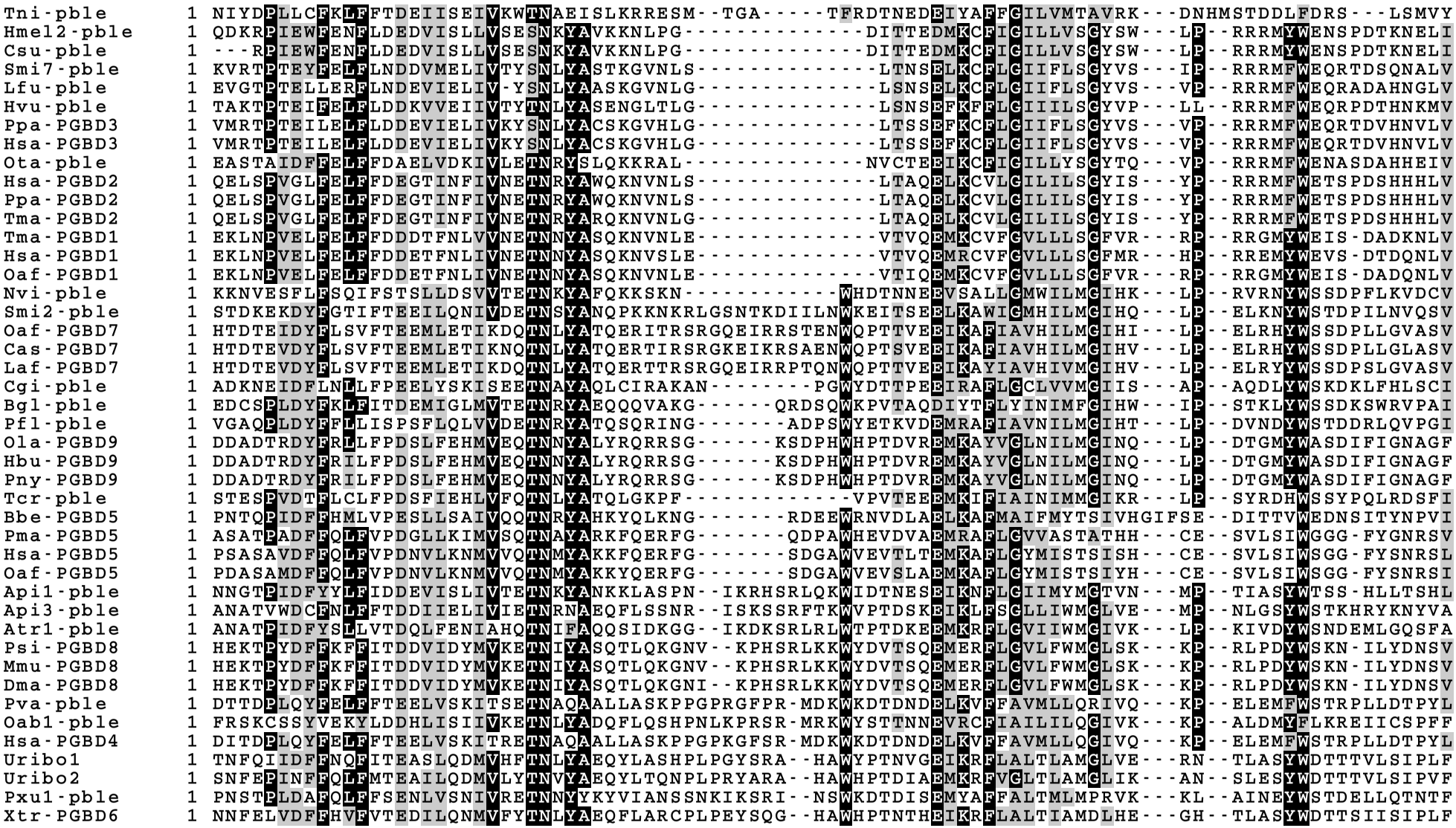
GGCTTATAATTATGCAGATGACTATTTTCTCAAGACAAAGGAAGGATTTACAAAAGCAAGCTTATTCTTT
CTATGCCAAAGCTTAAGTTTCAGCTCTTTTTTGAAGAAAAGTTCTCTAAAATTACTTTCCTGCTTCAAAA
GGGAAAACAGAAACCCCTATCCAGTATGCATTAATTCACCCTTTTCTTTTTCCTGATACTCATTCTAGGG
TGACTGTCGGTATAAAGGCGACACAGCTCACACTCAGCCCTCTGTATGCCTCTTGCCCCAGCCTTCCCTT
AAGTTTTTCATCTTAGACTTGCCTCCAAATCCAACTAAAGACCTAGAAGTACATGCATTTGGGGCCAGCA
TCATTAAGGACTCAGGGGCCCCACTGCTTCTAGCAGGTTAAGGCTGCAGAAATCCAACCTCTGTATAATC
GGGGGGGTCTAATATATTAAGCTGGATCTAGAATGCTTACATATTAATAAACCTGTCACCTAAGATAAAG
CAAACAGTACAATGTATTTATAATCCCCACAGACCGACCTTAACCGCTGCTTATAATAATCTTCATCTCC
ATCATCTCGGTATCTTCCCACTTTCCGACCTCCTCCTCCTCCTTCTCCTACAGAAGCAGCTTCAGCTTCT
TCCCCAGAACTTGGGAAAAAGTCATCATCAATCTCCTGCACTGGCACTTTCTTCTGCCGTTTCCCGCCCT
TGGGCAGAGGCTTCAGCTCATAGTCAGTACCATCTCCAGACAGGTCCGCCTCTGCCCCCTCCACCTCGTC
ATCTTCC

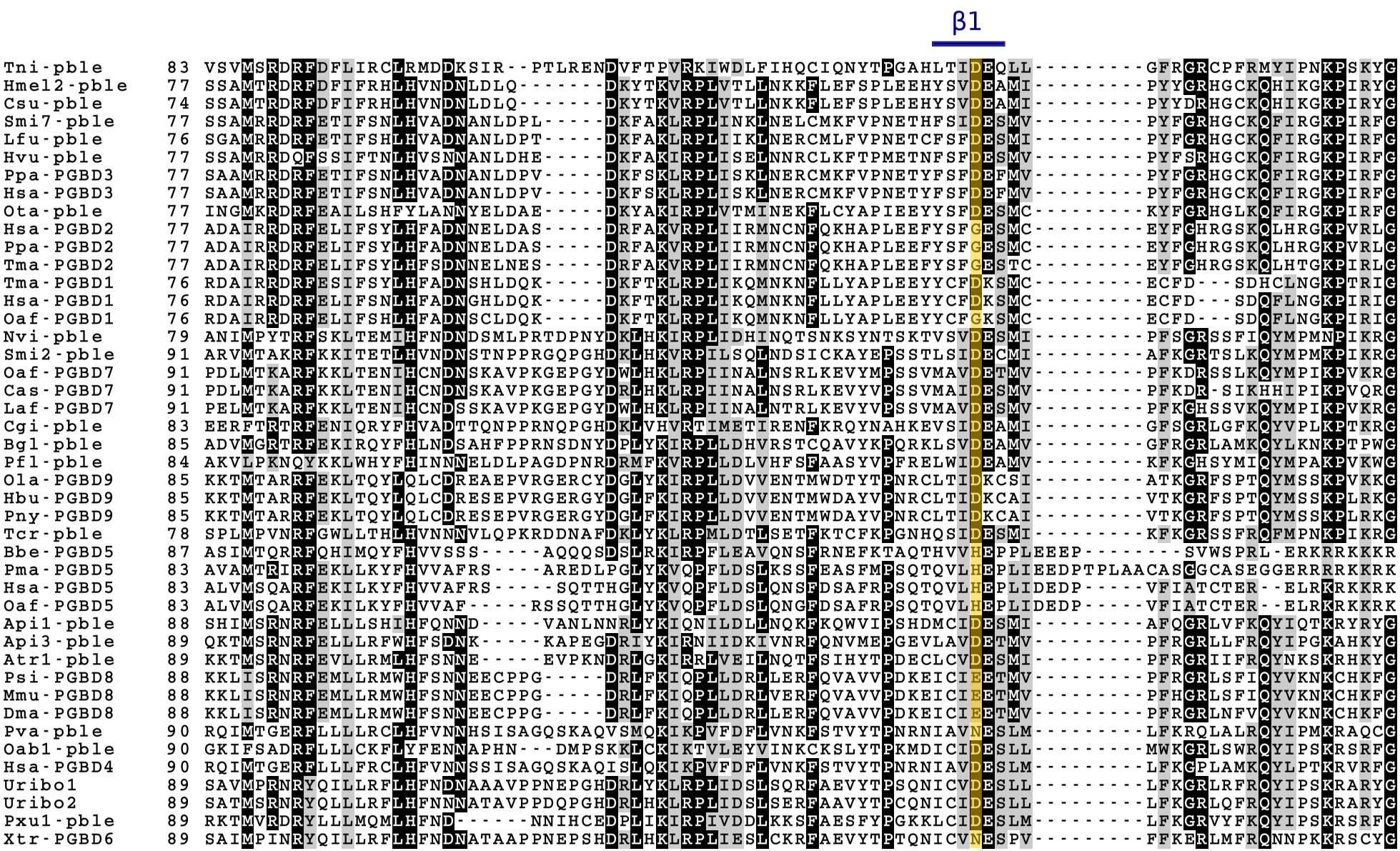
>NC_000015.10:34102083-34108686 Homo sapiens chromosome 15, GRCh38.p13 Primary Assembly (PGBD4)

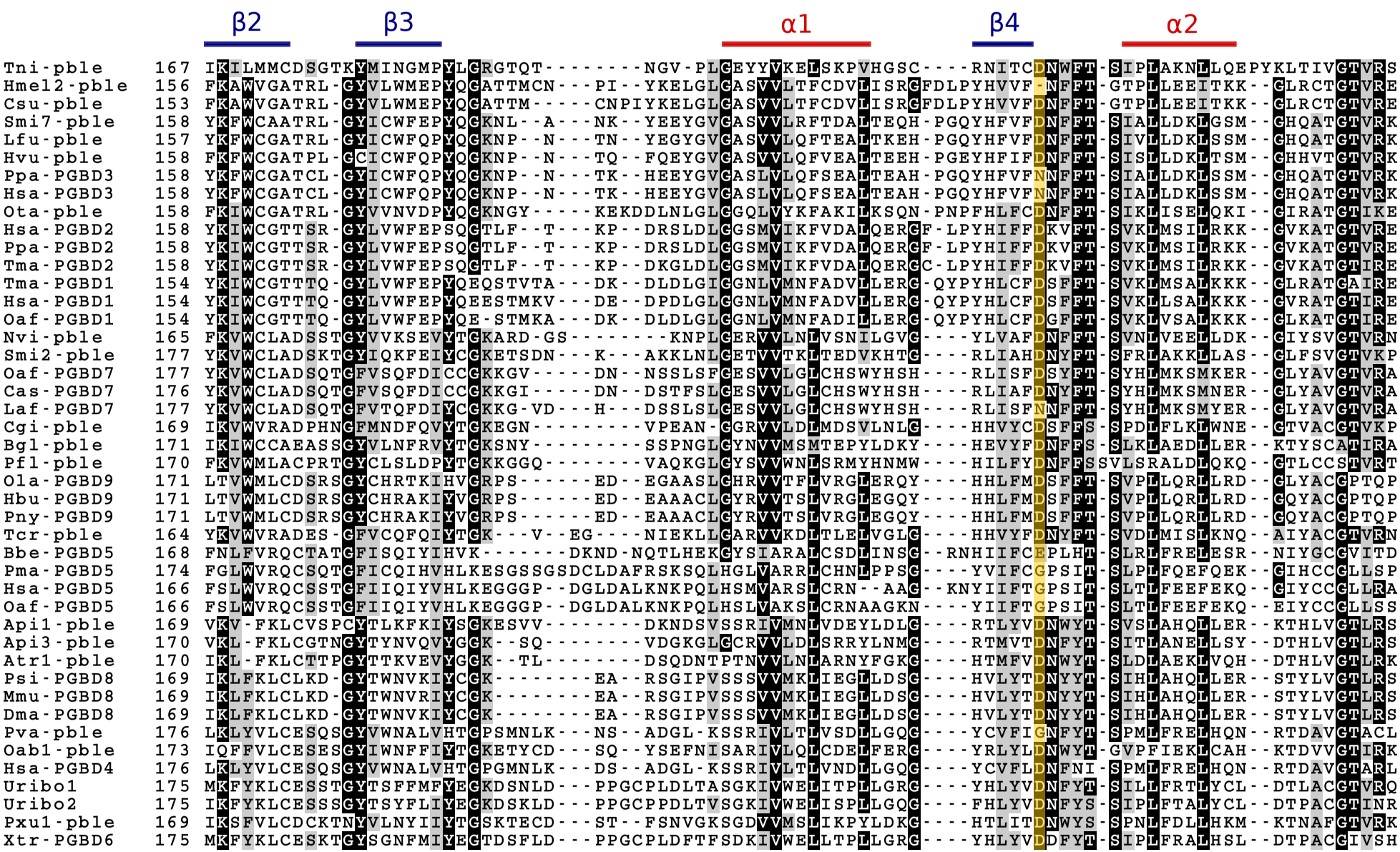
GCCTATGAAGGGGACTGCCCATGAAGTGAAAGTCAAGTGTGTGTTGCTGCGGCAGCCACGGAGGCCAAGG
ACCTCACGGGAGTAAAAGATGACGAGACTGGCTTCGGGAGAAACACCATCCAGAAGAGACCTTTCAAAAA
ACTTCTAGAGACTCCCCAAGACGTATGAGATGAAAGGCTTCTTCTGTCTGTGTGCGTTTAACTCATTTCT
CCTTAGCCCCGAGATTACGCGCTGCTGTGCCTGCGACTGCAGCGTTTACGCCGAGATAACTCGTGGATTA
CAGTGCCAACCTTACTCCCA**AA**G**TTTGCCA**C**GAAATAT**C**T**C**G**CTTC**T**G**T**TTATTTTCGCAT**G**GTTCT**G**GTA**
TATT**GACTTTT**G**AAACAAAAGACATC**A**TTCT**G**TTTATAGCATTCTGTTTTTAGTAGTGG**G**ATTTCCAT**C**T**
ACAAAATATAGTAATTCTCGATC**GCTGAA**ATGTCAAATCCTAGAAAACGTAGCATTCCTA**T**GCGTGAT**AG**
TAAT**A**C**GG**TCTCGAACAGTTGTTGGC**T**GAAGATTCATTTGATGAATC**T**GATTTTTC**G**GAAATAGA**T**GAT**
TCTGAT**A**ATT**TTTCGGATAGTGCTTTAGAAGCCGATAAGATCAGGCCTCTGTCCCATTTAGAATCTGATG
GAAAGAGCTCTACATCAAGTGACTCAGGGCGCTCCATGAAATGGTCAGCTCGTGCTATGATTCCACGTCA
AAGGTATGACTTTACCGGCACACCTGGCAGAAAAGTCGATGTCAGTGATATCACTGACCCATTGCAGTAT
TTTGAACTGTTCTTTACTGAGGAATTAGTTTCAAAAATTACTAGAGAAACAAATGCCCAAGCTGCCTTGT
TGGCTTCAAAGCCACCGGGTCCGAAAGGATTTTCGCGAATGGATAAATGGAAAGACACTGACAATGACGA
GCTCAAAGTCTTTTTTGCAGTAATGTTACTGCAAGGTATTGTGCAGAAACCTGAGCTGGAGATGTTTTGG
TCAACAAGGCCTCTTTTGGATACACCTTATCTCAGGCAAATTATGACTGGTGAAAGATTTTTACTTTTGT
TTCGGTGCCTGCATTTTGTCAACAATTCTTCTATATCTGCTGGTCAATCAAAGGCCCAGATTTCATTGCA
GAAGATCAAACCTGTGTTCGACTTTCTTGTAAATAAATTTTCCACTGTATATACTCCAAACAGAAACATT
GCAGTTGATGAATCACTGATGCTGTTCAAGGGGCCATTAGCTATGAAGCAGTACCTCCCGACAAAACGAG
TACGATTTGGTCTGAAGCTATATGTACTTTGTGAAAGTCAGTCTGGTTATGTGTGGAATGCGCTTGTTCA
CACAGGGCCTGGCATGAATTTGAAAGATTCAGCGGATGGCCTGAAATCATCACGCATTGTTCTTACCTTG
GTCAATGACCTTCTTGGCCAAGGGTATTGTGTCTTCCTCGATAACTTTAATATATCTCCCATGCTTTTCA
GAGAATTACATCAAAATAGGACTGATGCAGTTGGGACAGCTCGTTTGAACAGAAAACAGATTCCAAATGA
TCTGAAAAAAAGGATTGCAAAGGGGACGACTGTAGCCAGATTCTGTGGTGAACTTATGGCACTGAAATGG
TGTGACGGCAAGGAGGTGACAATGTTGTCAACATTCCACAATGATACTGTGATTGAAGTAAACAATAGAA
ATGGAAAGAAAACTAAAAGGCCACGTGTCATTGTGGATTATAACGAGAATATGGGAGCAGTGGACTCGGC
TGATCAAATGCTTACTTCTTATCCATCTGAGCGCAAAAGACACAAGGTTTGGTATAAGAAATTCTTTCAC
CATCTTCTACACATTACAGTGCTGAACTCCTACATCCTGTTCAAGAAGGATAATCCTGAGCACACGATGA
GCCATATAAACTTCAGACTGGCATTGATTGAAAGAATGCTGGAAAAGCATCACAAGCCAGGGCAGCAACA
TCTTCGAGGTCGTCCTTGCTCCGATGATGTCACACCTCTTCGTCTGTCTGGAAGACATTTCCCCAAGAGC
ATACCAGCAACGTCCGGGAAACAGAATCCAACTGGTCGCTGCAAAATTTGCTGCTCCCAATACGACAAGG
ATGGCAAGAAGATCCGGAAAGAAACGCGCTATTTTTGTGCCGAATGTGATGTTCCGCTTTGTGTTGTTCC
GTGCTTTGAAATTTACCACACGAAAAAAAATTATTAAATACCGATCATCATACACATCTGTTCCATTAGG
ATTAGAGACAAGTTCTGTTTAGAAATAACTCCAAGATTGGGAGGCCGACCTGGGTGGATCACCTGAGATC
AGGAATTCAAGACCAGCCTGGCCAACATGGTGAAACCCTGTCTCTACTAAAAATACAACAATTAGCTGGA
TGTGGTGGCACATGCCTGTAATCCAAGCTACTTGGGAGGCTGAGGCAGGAGAATTGCTTGAACCCATGAG
GCGGAGGTTGCAGTGAGCTGAGATCACACCACTGCACTCCAGCCTGGGTGACATAGCGAGACTCCATCTC
AAAAAAAAGAAATAACTCCAAGAACAGTTTTGTTGTTTTTGTTTTGTTTTGTTTTGTTTTGAGACGGAGT
TTCACTCTTGTTGCCCAGGCTGGAGTGCAATGGCATGATATTGGCTCACTGTAACCTCCACCTCCCAGGT
TCAAGCGATTCTCGTGCCTCAGCCTCCCTAGTAGCTGGGATTACAGGCATGTGCCACCACACCCACCTAA
TTTTGCATTTTTAGTAGAGACTGGGTTTATCCATGTTGGTCAGGCTAGTCTTGACCTCCCGACCTCAGGT
GATCCGCCTATCTGGGCCTCACAAAGTCCTGGGATTATAGGCATGAGCCACTGCGCCTAGCA**CCAAGAAC**
AGTTTTTATATTTTATTTTCACATTGAAAATCAGTCAGATTTGCTTCAGCCTCAAAGAGG**GTGTTTATGT**
AAAAC**TAAATGAG**T**GC**A**GGCAGCGAGCT**A**CACTTTTTTTTTTT**C**CTAAA**T**GG**AAA**T**GGG**AAGGAGATACT
GTCTCTGAACAAAAATAGCTGGAAAAACACATAACTGTTAACTAAGGGTTGCCTCTAGTGAAAAGGATTG
GAAGACAATTTATCTGATGCTACACTATTCTTCCTGTTGTGATTGTTTTACTGTGAACATAACAAGGTTG
TATTGGGGAGGAGGGCAGGGGTTGATTGCTGCCAATGTTGATTGCTGAGGCCTTAGCCAAGATTTCAGGC
ATCATTTGCTCCAGTCCTTTCATTTTCTTTATGGTGGAAAGTGAAGTGACTTGCCCAAGATCTTTGTTGG
CCCTAAACATCGCTGTCTCTGCTACAGATTCCTATTGAGAGCAATGCATGTTTTAGCCTTGTGCACTCTT
AGAGTCAGCATCATGGGAAAGAGTAACAGCAATTGAAAAAAGAGATTCAAGAGCCAAGTACTCAGCCCTG
GCAGGGGATGGGGACATGATGCCCAGAGCCACCCCTATACTCCAGTCTCCCTGGCATCAGCTACTGCAGT
ACTGGGGTACTTGCCTTCATCTTTTTTCTGAAGATAGTTACTTCTTCATATTTGATCATCGGAAGGAGCA
GCATTGCTCAGAAACACACACACACACACACACACTTGACTTTCTTGTCTGCCAAATAAATTTCTTATTT
GCTAGGGAACTGGTAATCTAGGCTCAAGAAGTGTGTGGAGAAGGTTTATTTTCATTCTTTTTACTTTAAA
ATATGCCTTAGTTTTTGAATAGATAAGACATTCAAGTGTCAGAAAGTAAAAGTCAGTTCAGGACTGAGAA
AGGAAAGATCAGTGCTGATTTTAATCAGACTGTAATGTACCAACTAGATACAGACAGGGCCAGTGGCAAG
GCCATCAGAAAACTTCTTCCCAAGGCCACACCACACCTGCAAAATGCCTAACAACCTTCTTCTTTTCAGT
GATTACCACTTTCTTACCAATGACTTCGCTTTGTTCCTCGTGCCTCTGAACAGACTGCTGAGATACCCAA
TCAAGAGCCAATCCTTGCTTCCCTAGACTCTCCTCAATCATTCAACAGAAGCCCAAATCTTGTAAAAGTT
CCTTTCTGGGCCGGGCACAGTGGCTCACGCCTGTAATCCCAGCACTTTGGAAGGCCGAGATGGGCGAATC
ACAAGGTCAGGAGTTTGAGACTAGCCTGGCCAACATGGTGAAACCCTGTCTCTACTAAAAATACAAAAAA
AAATTAGCTGGGCATAGTGGTGGGTGCCTGTAATCCCAGCTACTTGGGAGGCTGAGACAGGAGAATCACT
TGAACCCGGGAGGCAGAGGTTGCAGTGAGCCAAGATCGCACCACTGCACTCCAGCCCTGGCGACAGAGTG

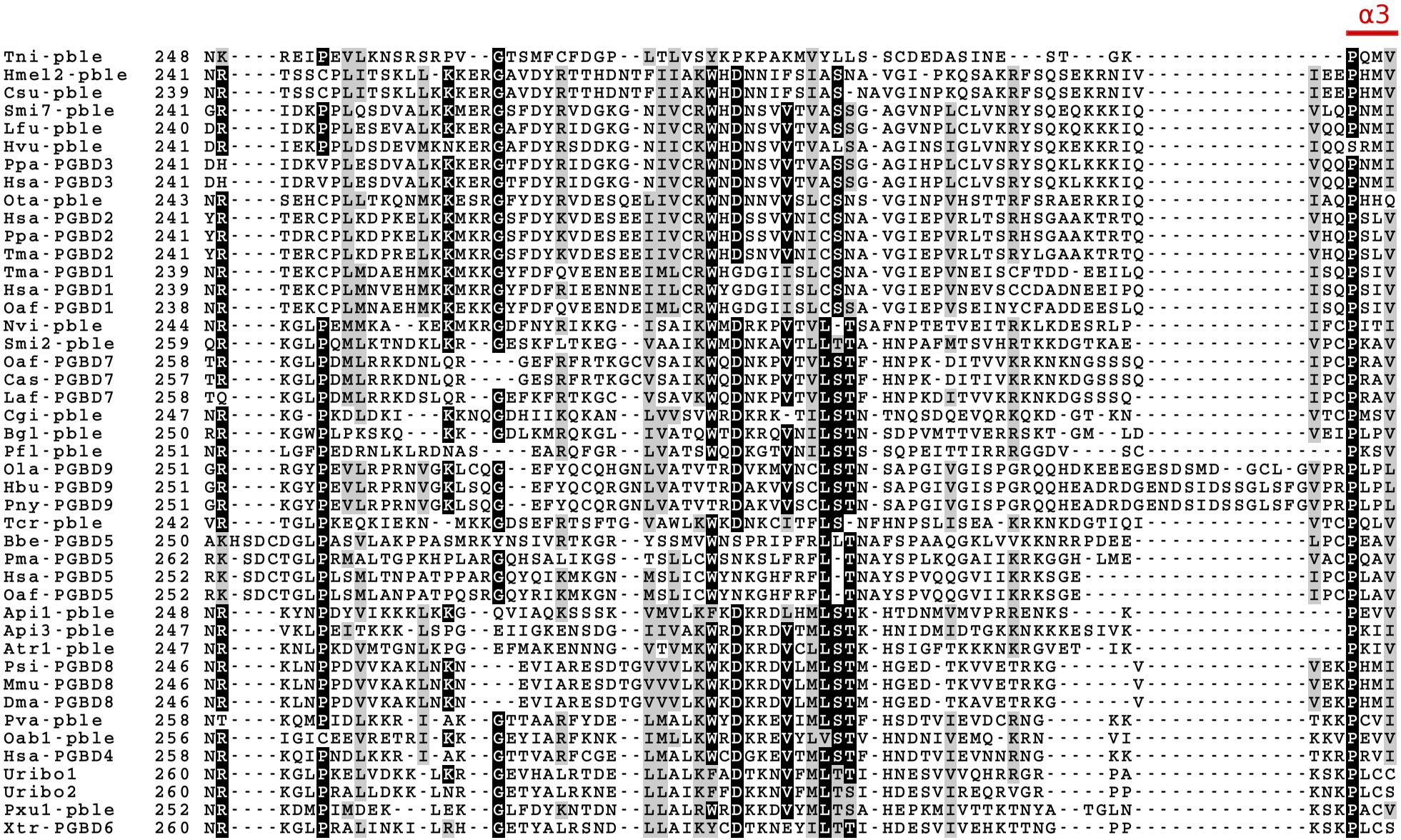
AGACTCCATCTCAAAAAATAAAAAAGTTTCCTTCTGTTCCCTCTTACTGTTATATCCTTGGTTTCTCTGG
TGTGGGGTCTTCCTTACAGCAGCAAGCTCATAAACTTAACTTGACCGCAGGTGTGTTCCTTGGTGGTCTT
TATCTGATAGACTTTGTCACATGGTACAATATTCAAAAGGCACATGGTACAATATTCAAAAGGCACATGG
TACAATATTCAAAAGGTATGAGTATACAAAAAAAATCTCCCACCTACTTCTCTCCCTAGTCACCCAATTC
TCAAGAGGCAATCACTGTTGCCATTTTCCTGAGTATTCTTCCACAGATAGTCAACTATCTACAGAAAAAT
ATATACATTAAAAATAAGGCCAGGCACAGTGGCATGCCTGTAGTCCCAGCTACTCAGAAGGTTGAAGCAG
GAGGATCACTTGAGGCCAGGAGGGCCAGGAATTTGAAGCTGTAGTGTGCTATGATAGCAGGCCTGTGAAT
AGCCATTGCACTCCATCCTGGGCAACATAGCCAGACCCTGTCTCTAAAAAAAAAAAAAAAAAAAAAAAAG
ACAAAAAAACAAATCCATATATACATATATATGTATATATATAGTATGCTACCAGTTGTGTGAAAAAAGG
GGCAAAATATATATTTATTTAACTTATATACAAATAACATTCTAGAAATAAATTGTTTAATATAAAATAC
ACTAATATATAATAATGTATTACCTAACATATGATTATATATAACTATAATGTGTACTGTTTTACATACA
TATTTCCAAAGTATACTATAAATGCACTTCCGCACTTTGCTCTTTTTACTAAATATATCTTGGAAATCAT
CCTTTATTCGTACATAAAAAGCTTCATAGTTCCTTTTTATGGCTGCAAAATGTTCCAGCTTATGGATGGA
CTGATTCTCTATCGAGCAACATTAAGATTGTGTCCTATTTTACTATTCCTAATTTTGCTGAAGTGAATTT
CTTTTGCCATGTGATTTCCACAGGTGTATATATGTAGCGTAATTAGTACTAGTAGAAAGTAGAATTGCTA
GATCAAAGAGTATGTGCCTTTGTAATTTTGATGATATTGTGAAATCTTCTTCCACAGAAGTTGTTGCAAC
TTTCACTCTCACCAGCAATGTGTAAGATAACTTGCTTCTTCATACCCTCACCAACAATGAGCTATCAAAT
GTTATTTTTGCCAAGCAAATAAATTAAAAGTGATATCTCAGTTTTTGTTTGTTTGTTTGTTTTTTGAGAC
AGAGTTTCACTCTTTGTTGCCCAGGCTGGAGTGCAGTGGCGTGATCTCGGCTCACTGTAACCTCCGCCTC
CCAGGTTCAAACGATTCTCCTGTCTCACCCTCCCGAGTAGCTGGGATTACAGGCACCTGCCACCACACCC
AGTTAATTTTTGTATTTTTAGTAGAGACAGGGTTTCACCATGTTGGCTAGGCTGACCTCGAACTCCTAAC
CTCAGGTGATCCACCTGCCTCGGCCTCCCAAAGTGCTGGAATTACAGGCGTGAGCCACCATGCCTGGCTC
TTCTTGCTTGTTTGTTTTGAGACAGAGTCTTGCTCTGTCAGCCAGGCTGGAGTGCAGTGGCACGATATTG
GCTCACTGCAACCTCTGTCTCCCAGGCTCAAGCAATTCTCCTGCCTCAGCCTCCTGAGTAGCTAGGATTA
CAGATGTGTGCCACCACGCCTGGCTAATTTTTGTACTTTTAGTAGAGACGGAGTTTCACCATGTTGGCCA
GGCTGGTCTCAAACTCCTGACCTCAGGTAATCCGCCTGCCTTGGCCTCCCAAAGTGCTGGGATGACAGGC
GTGAGCCACCACGCTGGGCCTCTTCTTGGTTTTTTAATCCATCTTCTGCCATATGGTGAATTAAATCTAA
GTTGTATTGTAGGATTAATGGAGAGCAGCTAAAGAGGAATAAAGCCATGTGTTCAGCCAAGGATAGTTAT
CATGAAGGATAAAAGCCTCAAAATCCTGATTTTGAAATACAAAGTGCTGATTTTAAAATCAGTACATGTT
GAGCTGAGCCCTGAGGCCTTTGTTTCAGTCCCAACCAGAGCAGAAGTAAGCATTGTTGAAACCGTAGTAA
AAAGATTATTTGAGAAATAATGGAATGGGACAAAGGATTATAATGTGCTACAGCCCTCTGACCATTACAC
TTTCTAAACTATTAATAAAATTTTACTTTTCAAAAATGATTTCAAAAATTGCAAAATTTAGAAATATGCT
GAAATAAAAACAGCTTCATCCCAA

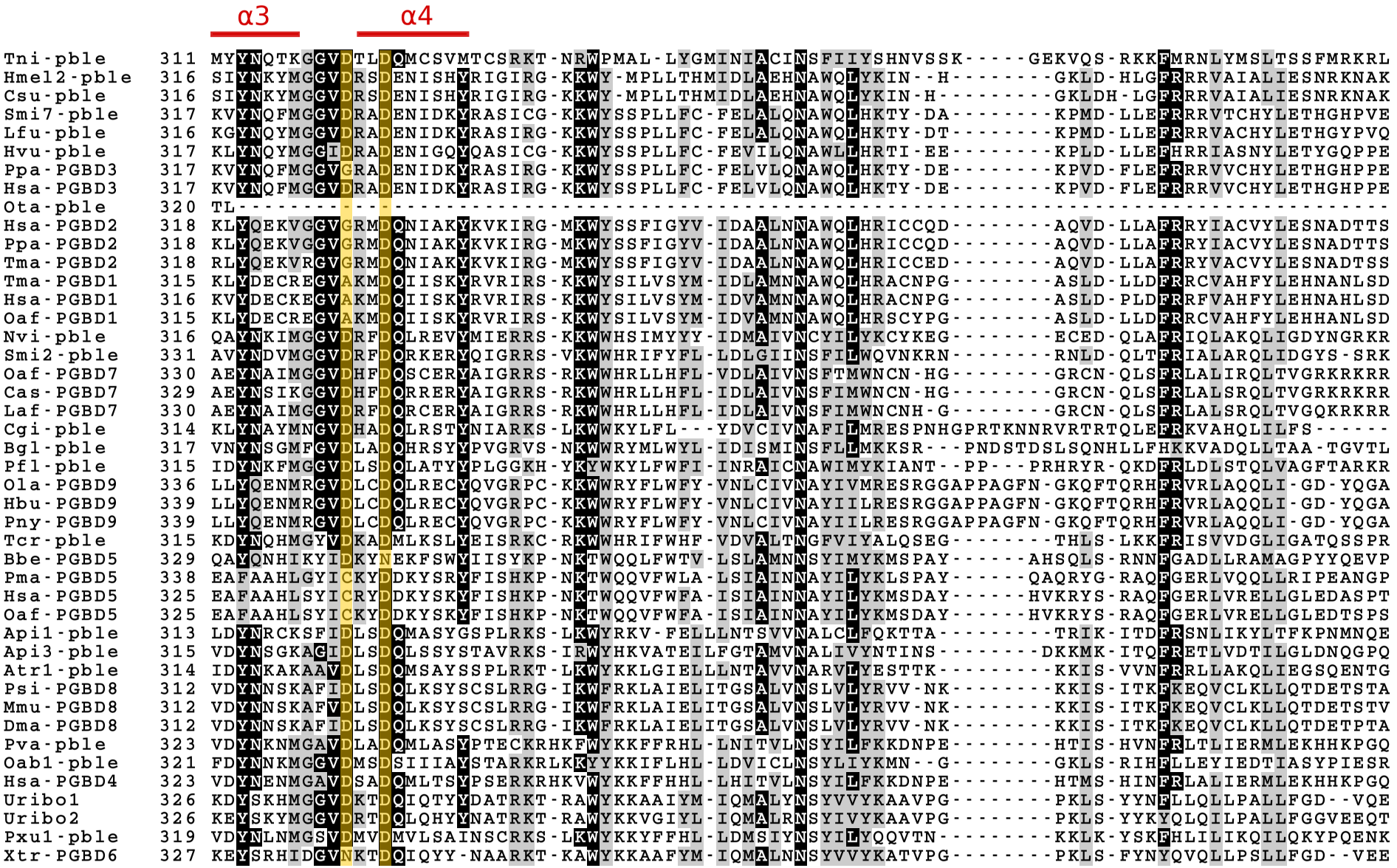
Supplementary data 3. Sequence alignment of *pble* transposases and PGBD proteins used to calculate the phylogenetic tree shown in Figure 4. The black shade is for residues which are present in more than 60% at a position. Yellow is for the position of the catalytic triad. The blue line is for β-sheet and red line for α-helix.


Tni-pble 1 NIYDPLLCFKLFFTDEIISEIVKWTNAEISLKRRESM--TGA-----TFRDTNEDEIYAFFGILVMTAVRK---DNHMSTDDLFDRS---LSMVY
Hmel2-pble 1 QDKRPIEWFENFLDEDVISLLVSESNKYAVKKNLPG--------------DITTEDMKCFIGILLVSGYSW---LP--RRRMYWENSPDTKNELI
Csu-pble 1 ---RPIEWFENFLDEDVISLLVSESNKYAVKKNLPG--------------DITTEDMKCFIGILLVSGYSW---LP--RRRMYWENSPDTKNELI
Smi7-pble 1 KVRTPTEYFELFLNDDVMELIVTYSNLYASTKGVNLS--------------LTNSELKCFLGIIFLSGYVS---IP--RRRMFWEQRTDSQNALV
Lfu-pble 1 EVGTPTELLERFLNDEVIELIV-YSNLYAASKGVNLG--------------LSNSELKCFLGIIFLSGYVS---VP--RRRMFWEQRADAHNGLV
Hvu-pble 1 TAKTPTEIFELFLDDKVVEIIVTYTNLYASENGLTLG--------------LSNSEFKFFLGIIILSGYVP---LL--RRRMFWEQRPDTHNKMV
Ppa-PGBD3 1 VMRTPTEILELFLDDEVIELIVKYSNLYACSKGVHLG--------------LTSSEFKCFLGIIFLSGYVS---VP--RRRMFWEQRTDVHNVLV
Hsa-PGBD3 1 VMRTPTEILELFLDDEVIELIVKYSNLYACSKGVHLG--------------LTSSEFKCFLGIIFLSGYVS---VP--RRRMFWEQRTDVHNVLV
Ota-pble 1 EASTAIDFFELFFDAELVDKIVLETNRYSLQKKRAL--------------NVCTEEIKCFIGILLYSGYTQ---VP--RRRMFWENASDAHHEIV
Hsa-PGBD2 1 QELSPVGLFELFFDEGTINFIVNETNRYAWQKNVNLS--------------LTAQELKCVLGILILSGYIS---YP--RRRMFWETSPDSHHHLV
Ppa-PGBD2 1 QELSPVGLFELFFDEGTINFIVNETNRYAWQKNVNLS--------------LTAQELKCVLGILILSGYIS---YP--RRRMFWETSPDSHHHLV
Tma-PGBD2 1 QELSPVGLFELFFDEGTINFIVNETNRYARQKNVNLG--------------LTAQELKCVLGILILSGYIS---YP--RRRMFWETSPDSHHHLV
Tma-PGBD1 1 EKLNPVELFELFFDDDTFNLVVNETNNYASQKNVNLE--------------VTVQEMKCVFGVLLLSGFVR---RP--RRGMYWEIS-DADKNLV
Hsa-PGBD1 1 EKLNPVELFELFFDDETFNLIVNETNNYASQKNVSLE--------------VTVQEMRCVFGVLLLSGFMR---HP--RREMYWEVS-DTDQNLV
Oaf-PGBD1 1 EKLNPVELFELFFDDETFNLIVNETNNYASQKNVNLE--------------VTIQEMKCVFGVLLLSGFVR---RP--RRGMYWEIS-DADQNLV
Nvi-pble 1 KKNVESFLFSQIFSTSLLDSVVTETNKYAFQKKSKN------------WHDTNNEEVSALLGMWILMGIHK---LP--RVRNYWSSDPFLKVDCV
Smi2-pble 1 STDKEKDYFGTIFTEEILQNIVDETNSYANQPKKNKRLGSNTKDIILNWKEITSEELKAWIGMHILMGIHQ---LP--ELKNYWSTDPILNVQSV
Oaf-PGBD7 1 HTDTEIDYFLSVFTEEMLETIKDQTNLYATQERITRSRGQEIRRSTENWQPTTVEEIKAFIAVHILMGIHI---LP--ELRHYWSSDPLLGVASV
Cas-PGBD7 1 HTDTEVDYFLSVFTEEMLETIKNQTNLYATQERTIRSRGKEIKRSAENWQPTSVEEIKAFIAVHILMGIHV---LP--ELRHYWSSDPLLGLASV
Laf-PGBD7 1 HTDTEVDYFLSVFTEEMLETIKDQTNLYATQERTTRSRGQEIRRPTQNWQPTTVEEIKAYIAVHIVMGIHV---LP--ELRYYWSSDPSLGVASV
Cgi-pble 1 ADKNEIDFLNLLFPEELYSKISEETNAYAQLCIRAKAN--------PGWYDTTPEEIRAFLGCLVVMGIIS---AP--AQDLYWSKDKLFHLSCI
Bgl-pble 1 EDCSPLDYFKLFITDEMIGLMVTETNRYAEQQQVAKG------QRDSQWKPVTAQDIYTFLYINIMFGIHW---IP--STKLYWSSDKSWRVPAI
Pfl-pble 1 VGAQPLDYFFLLISPSFLQLVVDETNRYATQSQRING-------ADPSWYETKVDEMRAFIAVNILMGIHT---LP--DVNDYWSTDDRLQVPGI
Ola-PGBD9 1 DDADTRDYFRLLFPDSLFEHMVEQTNNYALYRQRRSG------KSDPHWHPTDVREMKAYVGLNILMGINQ---LP--DTGMYWASDIFIGNAGF
Hbu-PGBD9 1 DDADTRDYFRILFPDSLFEHMVEQTNNYALYRQRRSG------KSDPHWHPTDVREMKAYVGLNILMGINQ---LP--DTGMYWASDIFIGNAGF
Pny-PGBD9 1 DDADTRDYFRILFPDSLFEHMVEQTNNYALYRQRRSG------KSDPHWHPTDVREMKAYVGLNILMGINQ---LP--DTGMYWASDIFIGNAGF
Tcr-pble 1 STESPVDTFLCLFPDSFIEHLVFQTNLYATQLGKPF--------------VPVTEEEMKIFIAINIMMGIKR---LP--SYRDHWSSYPQLRDSFI
Bbe-PGBD5 1 PNTQPIDFFHMLVPESLLSAIVQQTNRYAHKYQLKNG-------RDEEWRNVDLAELKAFMAIFMYTSIVHGIFSE--DITTVWEDNSITYNPVI
Pma-PGBD5 1 ASATPADFFQLFVPDGLLKIMVSQTNAYARKFQERFG-------QDPAWHEVDVAEMRAFLGVVASTATHH---CE--SVLSIWGGG-FYGNRSV
Hsa-PGBD5 1 PSASAVDFFQLFVPDNVLKNMVVQTNMYAKKFQERFG-------SDGAWVEVTLTEMKAFLGYMISTSISH---CE--SVLSIWSGG-FYSNRSL
Oaf-PGBD5 1 PDASAMDFFQLFVPDNVLKNMVVQTNMYAKKYQERFG-------SDGAWVEVSLAEMKAFLGYMISTSIYH---CE--SVLSIWSGG-FYSNRSI
Api1-pble 1 NNGTPIDFYYLFIDDEVISLIVTETNKYANKKLASPN--IKRHSRLQKWIDTNESEIKNFLGIIMYMGTVN---MP--TIASYWTSS-HLLTSHL
Api3-pble 1 ANATVWDCFNLFFTDDIIELIVIETNRNAEQFLSSNR--ISKSSRFTKWVPTDSKEIKLFSGLLIWMGLVE---MP--NLGSYWSTKHRYKNYVA
Atr1-pble 1 ANATPIDFYSLLVTDQLFENIAHQTNIFAQQSIDKGG--IKDKSRLRLWTPTDKEEMKRFLGVIIWMGIVK---LP--KIVDYWSNDEMLGQSFA
Psi-PGBD8 1 HEKTPYDFFKFFITDDVIDYMVKETNIYASQTLQKGNV--KPHSRLKKWYDVTSQEMERFLGVLFWMGLSK---KP--RLPDYWSKN-ILYDNSV
Mmu-PGBD8 1 HEKTPYDFFKFFITDDVIDYMVKETNIYASQTLQKGNV--KPHSRLKKWYDVTSQEMERFLGVLFWMGLSK---KP--RLPDYWSKN-ILYDNSV
Dma-PGBD8 1 HEKTPYDFFKFFITDDVIDYMVKETNIYASQTLQKGNI--KPHSRLKKWYDVTSQEMERFLGVLFWMGLSK---KP--RLPDYWSKN-ILYDNSV
Pva-pble 1 DTTDPLQYFELFFTEELVSKITSETNAQAALLASKPPGPRGFPR-MDKWKDTDNDELKVFFAVMLLQRIVQ---KP--ELEMFWSTRPLLDTPYL
Oab1-pble 1 FRSKCSSYVEKYLDDHLISIIVKETNLYADQFLQSHPNLKPRSR-MRKWYSTTNNEVRCFIAILILQGIVK---KP--ALDMYFLKREIICSPFF
Hsa-PGBD4 1 DITDPLQYFELFFTEELVSKITRETNAQAALLASKPPGPKGFSR-MDKWKDTDNDELKVFFAVMLLQGIVQ---KP--ELEMFWSTRPLLDTPYL
Uribo1 1 TNFQIIDFFNQFITEASLQDMVHFTNLYAEQYLASHPLPGYSRA--HAWYPTNVGEIKRFLALTLAMGLVE---RN--TLASYWDTTTVLSIPLF
Uribo2 1 SNFEPINFFQLFMTEAILQDMVLYTNVYAEQYLTQNPLPRYARA--HAWHPTDIAEMKRFVGLTLAMGLIK---AN--SLESYWDTTTVLSIPVF
Pxu1-pble 1 PNSTPLDAFQLFFSENLVSNIVRETNNYYKYVIANSSNKIKSRI--NSWKDTDISEMYAFFALTMLMPRVK---KL--AINEYWSTDELLQTNTF
Xtr-PGBD6 1 NNFELVDFFHVFVTEDILQNMVFYTNLYAEQFLARCPLPEYSQG--HAWHPTNTHEIKRFLALTIAMDLHE---GH--TLASYWDTTSIISIPLF


β1
Tni-pble 83 VSVMSRDRFDFLIRCLRMDDKSIR--PTLRENDVFTPVRKIWDLFIHQCIQNYTPGAHLTIDEQLL---------GFRGRCPFRMYIPNKPSKYG
Hmel2-pble 77 SSAMTRDRFDFIFRHLHVNDNLDLQ-------DKYTKVRPLVTLLNKKFLEFSPLEEHYSVDEAMI---------PYYGRHGCKQHIKGKPIRYG
Csu-pble 74 SSAMTRDRFDFIFRHLHVNDNLDLQ-------DKYTKVRPLVTLLNKKFLEFSPLEEHYSVDEAMI---------PYYDRHGCKQHIKGKPIRYG
Smi7-pble 77 SSAMRRDRFETIFSNLHVADNANLDPL-----DKFAKLRPLINKLNELCMKFVPNETHFSIDESMV---------PYFGRHGCKQFIRGKPIRFG
Lfu-pble 76 SGAMRRDRFETIFSHLHVADNANLDPT-----DKFAKLRPLIIKLNERCMLFVPNETCFSFDESMI---------PYFGRHGCKQFIRGKPIRFG
Hvu-pble 77 SSAMRRDQFSSIFTNLHVSNNANLDHE-----DKFAKIRPLISELNNRCLKFTPMETNFSFDESMV---------PYFSRHGCKQFIRGKPIRFG
Ppa-PGBD3 77 SAAMRRDRFETIFSNLHVADNANLDPV-----DKFSKLRPLISKLNERCMKFVPNETYFSFDEFMV---------PYFGRHGCKQFIRGKPIRFG
Hsa-PGBD3 77 SAAMRRDRFETIFSNLHVADNANLDPV-----DKFSKLRPLISKLNERCMKFVPNETYFSFDEFMV---------PYFGRHGCKQFIRGKPIRFG
Ota-pble 77 INGMKRDRFEAILSHFYLANNYELDAE-----DKYAKIRPLVTMINEKFLCYAPIEEYYSFDESMC---------KYFGRHGLKQFIRGKPIRFG
Hsa-PGBD2 77 ADAIRRDRFELIFSYLHFADNNELDAS-----DRFAKVRPLIIRMNCNFQKHAPLEEFYSFGESMC---------EYFGHRGSKQLHRGKPVRLG
Ppa-PGBD2 77 ADAIRRDRFELIFSYLHFADNNELDAS-----DRFAKVRPLIIRMNCNFQKHAPLEEFYSFGESMC---------EYFGHRGSKQLHRGKPVRLG
Tma-PGBD2 77 ADAIRRDRFELIFSYLHFSDNNELNES-----DRFAKVRPLIIRMNCNFQKHAPLEEFYSFGESTC---------EYFGHRGSKQLHTGKPIRLG
Tma-PGBD1 76 RDAIRRDRFESIFSYLHFADNSHLDQK-----DKFTKLRPLIKQMNKNFLLYAPLEEYYCFDKSMC---------ECFD---SDHCLNGKPTRIG
Hsa-PGBD1 76 RDAIRRDRFELIFSNLHFADNGHLDQK-----DKFTKLRPLIKQMNKNFLLYAPLEEYYCFDKSMC---------ECFD---SDQFLNGKPIRIG
Oaf-PGBD1 76 RDAIRRDRFELIFSHLHFADNSCLDQK-----DKFTKLRPLIKQMNKNFLLYAPLEEYYCFGKSMC---------ECFD---SDQFLNGKPIRIG
Nvi-pble 79 ANIMPYTRFSKLTEMIHFNDNDSMLPRTDPNYDKLHKIRPLIDHINQTSNKSYNTSKTVSVDESMI---------PFSGRSSFIQYMPMNPIKRG
Smi2-pble 91 ARVMTAKRFKKITETLHVNDNSTNPPRGQPGHDKLHKVRPILSQLNDSICKAYEPSSTLSIDECMI---------AFKGRTSLKQYMPMKPIKRG
Oaf-PGBD7 91 PDLMTKARFKKLTENIHCNDNSKAVPKGEPGYDWLHKLRPIINALNSRLKEVYMPSSVMAVDETMV---------PFKDRSSLKQYMPIKPVKRG
Cas-PGBD7 91 PDLMTKARFKKLTENIHCNDNSKAVPKGEPGYDRLHKLRPIINALNSRLKEVYVPSSVMAIDESMV---------PFKDR-SIKHHIPIKPVQRG
Laf-PGBD7 91 PELMTKARFKKLTENIHCNDSSKAVPKGEPGYDWLHKLRPIINALNTRLKEVYVPSSVMAVDESMV---------PFKGHSSVKQYMPIKPVKRG
Cgi-pble 83 EERFTRTRFENIQRYFHVADTTQNPPRNQPGHDKLVHVRTIMETIRENFKRQYNAHKEVSIDEAMI---------GFSGRLGFKQYVPLKPTKRG
Bgl-pble 85 ADVMGRTRFEKIRQYFHLNDSAHFPPRNSDNYDPLYKIRPLLDHVRSTCQAVYKPQRKLSVDEAMV---------GFRGRLAMKQYLKNKPTPWG
Pfl-pble 84 AKVLPKNQYKKLWHYFHINNNELDLPAGDPNRDRMFKVRPLLDLVHFSFAASYVPFRELWIDEAMV---------KFKGHSYMIQYMPAKPVKWG
Ola-PGBD9 85 KKTMTARRFEKLTQYLQLCDREAEPVRGERCYDGLYKIRPLLDVVENTMWDTYTPNRCLTIDKCSI---------ATKGRFSPTQYMSSKPVKKG
Hbu-PGBD9 85 KKTMTARRFEKLTQYLQLCDRESEPVRGERGYDGLFKIRPLLDVVENTMWDAYVPNRCLTIDKCAI---------VTKGRFSPTQYMSSKPLRKG
Pny-PGBD9 85 KKTMTARRFEKLTQYLQLCDRESEPVRGERGYDGLFKIRPLLDVVENTMWDAYVPNRCLTIDKCAI---------VTKGRFSPTQYMSSKPLRKG
Tcr-pble 78 SPLMPVNRFGWLLTHLHVNNNVLQPKRDDNAFDKLYKLRPMLDTLSETFKTCFKPGNHQSIDESMI---------KFKGRSSFRQFMPNKPIKRG
Bbe-PGBD5 87 ASIMTQRRFQHIMQYFHVVSSS-----AQQQSDSLRKIRPFLEAVQNSFRNEFKTAQTHVVHEPPLEEEP--------SVWSPRL-ERKRRKKKR
Pma-PGBD5 83 AVAMTRIRFEKLLKYFHVVAFRS----AREDLPGLYKVQPFLDSLKSSFEASFMPSQTQVLHEPLIEEDPTPLAACASGGCASEGGERRRRKKRK
Hsa-PGBD5 83 ALVMSQARFEKILKYFHVVAFRS-----SQTTHGLYKVQPFLDSLQNSFDSAFRPSQTQVLHEPLIDEDP-----VFIATCTER--ELRKRKKRK
Oaf-PGBD5 83 ALVMSQARFEKILKYFHVVAF-----RSSQTTHGLYKVQPFLDSLQNGFDSAFRPSQTQVLHEPLIDEDP-----VFIATCTER--ELRKRKKRK
Api1-pble 88 SHIMSRNRFELLLSHIHFQNND-----VANLNNRLYKIQNILDLLNQKFKQWVIPSHDMCIDESMI---------AFQGRLVFKQYIQTKRYRYG
Api3-pble 89 QKTMSRNRFELLLRFWHFSDNK-----KAPEGDRIYKIRNIIDKIVNRFQNVMEPGEVLAVDETMV---------PFRGRLLFRQYIPGKAHKYG
Atr1-pble 89 KKTMSRNRFEVLLRMLHFSNNE-----EVPKNDRLGKIRRLVEILNQTFSIHYTPDECLCVDESMI---------PFRGRIIFRQYNKSKRHKYG
Psi-PGBD8 88 KKLISRNRFEMLLRMWHFSNNEECPPG-----DRLFKIQPLLDRLLERFQVAVVPDKEICIEETMV---------PFRGRLSFIQYVKNKCHKFG
Mmu-PGBD8 88 KKLISRNRFEMLLRMWHFSNNEECPPG-----DRLFKIQPLLDRLVERFQVAVVPDKEICIEETMV---------PFHGRLSFIQYVKNKCHKFG
Dma-PGBD8 88 KKLISRNRFEMLLRMWHFSNNEECPPG-----DRLFKIQPLLDRLLERFQVAVVPDKEICIEETMV---------PFRGRLNFVQYVKNKCHKFG
Pva-pble 90 RQIMTGERFLLLLRCLHFVNNHSISAGQSKAQVSMQKIKPVFDFLVNKFSTVYTPNRNIAVNESLM---------LFKRQLALRQYIPMKRAQCG
Oab1-pble 90 GKIFSADRFLLLCKFLYFENNAPHN---DMPSKKLCKIKTVLEYVINKCKSLYTPKMDICIDESLL---------MWKGRLSWRQYIPSKRSRFG
Hsa-PGBD4 90 RQIMTGERFLLLFRCLHFVNNSSISAGQSKAQISLQKIKPVFDFLVNKFSTVYTPNRNIAVDESLM---------LFKGPLAMKQYLPTKRVRFG
Uribo1 89 SAVMPRNRYQILLRFLHFNDNAAAVPPNEPGHDRLYKLRPLIDSLSQRFAEVYTPSQNICVDESLL---------LFKGRLRFRQYIPSKRARYG
Uribo2 89 SATMSRNRYQLLLRFLHFNNNATAVPPDQPGHDRLHKLRPLIDSLSERFAAVYTPCQNICIDESLL---------LFKGRLQFRQYIPSKRARYG
Pxu1-pble 89 RKTMVRDRYLLLMQMLHFND------NNIHCEDPLIKIRPIVDDLKKSFAESFYPGKKLCIDESLM---------LFKGRVYFKQYIPSKRSRFG
Xtr-PGBD6 89 SAIMPINRYQILLRFLHFNDNATAAPPNEPSHDRLHKLRPLIESLCKRFAEVYTPTQNICVNESPV---------FFKERLMFRQNNPKKRSCYG


β2
β3
α1
β4
α2
Tni-pble 167 IKILMMCDSGTKYMINGMPYLGRGTQT--------NGV-PLGEYYVKELSKPVHGSC----RNITCDNWFT-SIPLAKNLLQEPYKLTIVGTVRS
Hmel2-pble 156 FKAWVGATRL-GYVLWMEPYQGATTMCN---PI--YKELGLGASVVLTFCDVLISRGFDLPYHVVF-NFFT-GTPLLEEITKK--GLRCTGTVRE
Csu-pble 153 FKAWVGATRL-GYVLWMEPYQGATTM-----CNPIYKELGLGASVVLTFCDVLISRGFDLPYHVVFDNFFT-GTPLLEEITKK--GLRCTGTVRE
Smi7-pble 158 YKFWCAATRL-GYICWFEPYQGKNL--A---NK--YEEYGVGASVVLRFTDALTEQH-PGQYHFVFDNFFT-SIALLDKLGSM--GHQATGTVRK
Lfu-pble 157 YKFWCGATRL-GYICWFQPYQGKNP--N---TN--YEGYGVGASVVLQFTEALTKEH-PGQYHFVFDNFFT-SIVLLDKLSSM--GHQATGTVRK
Hvu-pble 158 FKFWCGATPL-GCICWFQPYQGKNP--N---TQ--FQEYGVGASVVLQFVEALTEEH-PGEYHFIFDNFFT-SISLLDKLTSM--GHHVTGTVRK
Ppa-PGBD3 158 YKFWCGATCL-GYICWFQPYQGKNP--N---TK--HEEYGVGASLVLQFSEALTEAH-PGQYHFVFNNFFT-SIALLDKLSSM--GHQATGTVRK
Hsa-PGBD3 158 YKFWCGATCL-GYICWFQPYQGKNP--N---TK--HEEYGVGASLVLQFSEALTEAH-PGQYHFVFNNFFT-SIALLDKLSSM--GHQATGTVRK
Ota-pble 158 FKIWCGATRL-GYVVNVDPYQGKNGY-----KEKDDLNLGLGGQLVYKFAKILKSQN-PNPFHLFCDNFFT-SIKLISELQKI--GIRATGTIKE
Hsa-PGBD2 158 YKIWCGTTSR-GYLVWFEPSQGTLF--T---KP--DRSLDLGGSMVIKFVDALQERGF-LPYHIFFDKVFT-SVKLMSILRKK--GVKATGTVRE
Ppa-PGBD2 158 YKIWCGTTSR-GYLVWFEPSQGTLF--T---KP--DRSLDLGGSMVIKFVDALQERGF-LPYHIFFDKVFT-SVKLMSILRKK--GVKATGTVRE
Tma-PGBD2 158 YKIWCGTTSR-GYLVWFEPSQGTLF--T---KP--DKGLDLGGSMVIKFVDALQERGC-LPYHIFFDKVFT-SVKLMSILRKK--GVKATGTIRE
Tma-PGBD1 154 YKIWCGTTTQ-GYLVWFEPYQEQSTVTA---DK--DLDLGIGGNLVMNFADVLLERG-QYPYHLCFDSFFT-SVKLMSALKKK--GLRATGAIRE
Hsa-PGBD1 154 YKIWCGTTTQ-GYLVWFEPYQEESTMKV---DE--DPDLGLGGNLVMNFADVLLERG-QYPYHLCFDSFFT-SVKLLSALKKK--GVRATGTIRE
Oaf-PGBD1 154 YKIWCGTTTQ-GYLVWFEPYQE-STMKA---DK--DLDLGLGGNLVMNFADILLERG-QYPYHLCFDGFFT-SVKLVSALKKK--GLKATGTIRE
Nvi-pble 165 FKVWCLADSSTGYVVKSEVYTGKARD-GS--------KNPLGERVVLNLVSNILGVG----YLVAFDNFFT-SVNLVEELLDK--GIYSVGTVRN
Smi2-pble 177 YKVWCLADSKTGYIQKFEIYCGKETSDN---K---AKKLNLGETVVTKLTEDVKHTG----RLIAHDNYFT-SFRLAKKLLAS--GLFSVGTVKP
Oaf-PGBD7 177 YKVWCLADSQTGFVSQFDICCGKKGV-----DN--NSSLSFGESVVLGLCHSWYHSH----RLISFDSYFT-SYHLMKSMKER--GLYAVGTVRA
Cas-PGBD7 176 YKVWCLADSQTGFVSQFDICCGKKGI-----DN--DSTFSLGESVVLGLCHSWYHSH----RLIAFDNYFT-SYHLMKSMNER--GLYAVGTVRA
Laf-PGBD7 177 YKVWCLADSQTGFVTQFDIYCGKKG-VD---H---DSSLSLGESVVLGLCHSWYHSH----RLISFNNFFT-SYHLMKSMYER--GLYAVGTVRA
Cgi-pble 169 IKVWVRADPHNGFMNDFQVYTGKEGN---------VPEAN-GGRVVLDLMDSVLNLG----HHVYCDSFFS-SPDLFLKLWNE--GTVACGTVKP
Bgl-pble 171 IKIWCCAEASSGYVLNFRVYTGKSNY---------SSPNGLGYNVVMSMTEPYLDKY----HEVYFDNFFS-SLKLAEDLLER--KTYSCATIRA
Pfl-pble 170 FKVWMLACPRTGYCLSLDPYTGKKGGQ--------VAQKGLGYSVVWNLSRMYHNMW----HILFYDNFFSSVLSRALDLQKQ--GTLCCSTVRT
Ola-PGBD9 171 LTVWMLCDSRSGYCHRTKIHVGRPS------ED--EGAASLGHRVVTFLVRGLERQY----HHLFMDSFFT-SVPLLQRLLRD--GLYACGPTQP
Hbu-PGBD9 171 LTVWMLCDSRSGYCHRAKIYVGRPS------ED--EAAACLGYRVVTSLVRGLEGQY----HHLFMDSFFT-SVPLLQRLLRD--GQYACGPTQP
Pny-PGBD9 171 LTVWMLCDSRSGYCHRAKIYVGRPS------ED--EAAACLGYRVVTSLVRGLEGQY----HHLFMDSFFT-SVPLLQRLLRD--GQYACGPTQP
Tcr-pble 164 YKVWVRADES-GFVCQFQIYTGK---V--EG----NIEKLLGARVVKDLTLELVGLG----HHVYFDNYFT-SVDLMISLKNQ--AIYACGTVRN
Bbe-PGBD5 168 FNLFVRQCTATGFISQIYIHVK-------DKND-NQTLHEKGYSIARALCSDLINSG--RNHIIFCEPLHT-SLRLFRELESR--NIYGCGVITD
Pma-PGBD5 174 FGLWVRQCSQTGFICQIHVHLKESGSSGSDCLDAFRSKSQLHGLVARRLCHNLPPSG----YVIFCGPSIT-SLPLFQEFQEK--GIHCCGLLSP
Hsa-PGBD5 166 FSLWVRQCSSTGFIIQIYVHLKEGGGP--DGLDALKNKPQLHSMVARSLCRN--AAG--KNYIIFTGPSIT-SLTLFEEFEKQ--GIYCCGLLRA
Oaf-PGBD5 166 FSLWVRQCSSTGFIIQIYVHLKEGGGP--DGLDALKNKPQLHSLVAKSLCRNAAGKN----YIIFTGPSIT-SLTLFEEFEKQ--EIYCCGLLSS
Api1-pble 169 VKV-FKLCVSPCYTLKFKIYSGKESVV--------DKNDSVSSRIVMNLVDEYLDLG----RTLYVDNWYT-SVSLAHQLLER--KTHLVGTLRS
Api3-pble 170 VKL-FKLCGTNGYTYNVQVYGGK--SQ--------VDGKGLGCRVVLDLSRRYLNMG----RTMVTDNFYT-SITLANELLSY--DTHLVGTLRS
Atr1-pble 170 IKL-FKLCTTPGYTTKVEVYGGK--TL--------DSQDNTPTNVVLNLARNYFGKG----HTMFVDNWYT-SLDLAEKLVQH--DTHLVGTLRK
Psi-PGBD8 169 IKLFKLCLKD-GYTWNVKIYCGK--------EA--RSGIPVSSSVVMKLIEGLLDSG----HVLYTDNYYT-SIHLAHQLLER--STYLVGTLRS
Mmu-PGBD8 169 IKLFKLCLKD-GYTWNVRIYCGK--------EA--RSGIPVSSSVVMKLIEGLLDSG----HVLYTDNYYT-SIHLAHQLLER--STYLVGTLRS
Dma-PGBD8 169 IKLFKLCLKD-GYTWNVKIYCGK--------EA--RSGIPVSSSVVMKLIEGLLDSG----HVLYTDNYYT-SIHLAHQLLER--STYLVGTLRS
Pva-pble 176 LKLYVLCESQSGYVWNALVHTGPSMNLK---NS--ADGL-KSSRIVLTLVSDLLGQG----YCVFIGNFYT-SPMLFRELHQN--RTDAVGTACL
Oab1-pble 173 IQFFVLCESESGYIWNFFIYTGKETYCD---SQ--YSEFNISARIVLQLCDELFERG----YRLYLDNWYT-GVPFIEKLCAH--KTDVVGTIRK
Hsa-PGBD4 176 LKLYVLCESQSGYVWNALVHTGPGMNLK---DS--ADGL-KSSRIVLTLVNDLLGQG----YCVFLDNFNI-SPMLFRELHQN--RTDAVGTARL
Uribo1 175 MKFYKLCESSTGYTSFFMFYEGKDSNLD---PPGCPLDLTASGKIVWELITPLLGRG----YHLYVDNFYT-SILLFRTLYCL--DTLACGTVRQ
Uribo2 175 IKFYKLCESSSGYTSYFLIYEGKDSKLD---PPGCPPDLTVSGKIVWELISPLLGQG----FHLYVDNFYS-SIPLFTALYCL--DTPACGTINR
Pxu1-pble 169 IKSFVLCDCKTNYVLNYIIYTGSKTECD---ST--FSNVGKSGDVVMSLIKPYLDKG----HTLITDNWYS-SPNLFDLLHKM--KTNAFGTVRK
Xtr-PGBD6 175 MKFYKLCESKTGYSGNFMIYEGTDSFLD---PPGCPLDFTFSDKIVWELLTPLLGRG----YHLYVDDFYT-SIPLFRALHSL--DTPACGIVSH


α3
Tni-pble 248 NK----REIPEVLKNSRSRPV--GTSMFCFDGP--LTLVSYKPKPAKMVYLLS-SCDEDASINE---ST---GK----------PQMV
Hmel2-pble 241 NR----TSSCPLITSKLL-KKERGAVDYRTTHDNTFIIAKWHDNNIFSIASNA-VGI-PKQSAKRFSQSEKRNIV----------IEEPHMV
Csu-pble 239 NR----TSSCPLITSKLLKKKERGAVDYRTTHDNTFIIAKWHDNNIFSIAS-NAVGINPKQSAKRFSQSEKRNIV----------IEEPHMV
Smi7-pble 241 GR----IDKPPLQSDVALKKMERGSFDYRVDGKG-NIVCRWHDNSVVTVASSG-AGVNPLCLVNRYSQEQKKKIQ----------VLQPNMI
Lfu-pble 240 DR----IDKPPLESEVALKKKERGAFDYRIDGKG-NIVCRWNDNSVVTVASSG-AGVNPLCLVKRYSQKQKKKIQ----------VQQPNMI
Hvu-pble 241 DR----IEKPPLDSDEVMKNKERGAFDYRSDDKG-NIICKWHDNSVVTVALSA-AGINSIGLVNRYSQKEKKRIQ----------IQQSRMI
Ppa-PGBD3 241 DH----IDKVPLESDVALKKKERGTFDYRIDGKG-NIVCRWNDNSVVTVASSG-AGIHPLCLVSRYSQKLKKKIQ----------VQQPNMI
Hsa-PGBD3 241 DH----IDRVPLESDVALKKKERGTFDYRIDGKG-NIVCRWNDNSVVTVASSG-AGIHPLCLVSRYSQKLKKKIQ----------VQQPNMI
Ota-pble 243 NR----SEHCPLLTKQNMKKESRGFYDYRVDESQELIVCKWNDNNVVSLCSNS-VGINPVHSTTRFSRAERKRIQ----------IAQPHHQ
Hsa-PGBD2 241 YR----TERCPLKDPKELKKMKRGSFDYKVDESEEIIVCRWHDSSVVNICSNA-VGIEPVRLTSRHSGAAKTRTQ----------VHQPSLV
Ppa-PGBD2 241 YR----TDRCPLKDPKELKKMKRGSFDYKVDESEEIIVCRWHDSSVVNICSNA-VGIEPVRLTSRHSGAAKTRTQ----------VHQPSLV
Tma-PGBD2 241 YR----TERCPLKDPRELKKMKRGSFDYKVDESEEIIVCRWHDSNVVNICSNA-VGIEPVRLTSRYLGAAKTRTQ----------VHQPSLV
Tma-PGBD1 239 NR----TEKCPLMDAEHMKKMKKGYFDFQVEENEEIMLCRWHGDGIISLCSNA-VGIEPVNEISCFTDD-EEILQ----------ISQPSIV
Hsa-PGBD1 239 NR----TEKCPLMNVEHMKKMKRGYFDFRIEENNEIILCRWYGDGIISLCSNA-VGIEPVNEVSCCDADNEEIPQ----------ISQPSIV
Oaf-PGBD1 238 NR----TEKCPLMNAEHMKKEKKGYFDFQVEENDEIMLCRWHGDGIISLCSSA-VGIEPVSEINYCFADDEESLQ----------ISQPSIV
Nvi-pble 244 NR----KGLPEMMKA--KEKMKRGDFNYRIKKG--ISAIKWMDRKPVTVL-TSAFNPTETVEITRKLKDESRLP----------IFCPITI
Smi2-pble 259 QR----KGLPQMLKTNDKLKR--GESKFLTKEG--VAAIKWMDNKAVTLLLTTA-HNPAFMTSVHRTKKDGTKAE----------VPCPKAV
Oaf-PGBD7 258 TR----KGLPDMLRRKDNLQR----GEFRFRTKGCVSAIKWQDNKPVTVLSTF-HNPK-DITVVKRKNKNGSSSQ----------IPCPRAV
Cas-PGBD7 257 TR----KGLPDMLRRKDNLQR----GESRFRTKGCVSAIKWQDNKPVTVLSTF-HNPK-DITIVKRKNKDGSSSQ----------IPCPRAV
Laf-PGBD7 258 TQ----KGLPDMLRRKDSLQR--GEFKFRTKGC--VSAVKWQDNKPVTVLSTF-HNPKDITVVKRKNKDGSSSQ----------IPCPRAV
Cgi-pble 247 NR----KG-PKDLDKI---KKNQGDHIIKQKAN--LVVSVWRDKRK-TILSTN-TNQSDQEVQRKQKD-GT-KN----------VTCPMSV
Bgl-pble 250 RR----KGWPLPKSKQ---KK--GDLKMRQKGL--IVATQWTDKRQVNILSTN-SDPVMTTVERRSKT-GM--LD----------VEIPLPV
Pfl-pble 251 NR----LGFPEDRNLKLRDNAS----EARQFGR--LVATSWQDKGTVNILSTS-SQPEITTIRRRGGDV----SC----------PKSV
Ola-PGBD9 251 GR----RGYPEVLRPRNVGKLCQG--EFYQCQHGNLVATVTRDVKMVNCLSTN-SAPGIVGISPGRQQHDKEEEGESDSMD--GCL-GVPRPLPL
Hbu-PGBD9 251 GR----KGYPEVLRPRNVGKLSQG--EFYQCQRGNLVATVTRDAKVVSCLSTN-SAPGIVGISPGRQQHEADRDGENDSIDSSGLSFGVPRPLPL
Pny-PGBD9 251 GR----KGYPEVLRPRNVGKLSQG--EFYQCQRGNLVATVTRDAKVVSCLSTN-SAPGIVGISPGRQQHEADRDGENDSIDSSGLSFGVPRPLPL
Tcr-pble 242 VR----TGLPKEQKIEKN--MKKGDSEFRTSFTG-VAWLKWKDNKCITFLS-NFHNPSLISEA-KRKNKDGTIQI----------VTCPQLV
Bbe-PGBD5 250 AKHSDCDGLPASVLAKPPASMRKYNSIVRTKGR--YSSMVWNSPRIPFRLLTNAFSPAAQGKLVVKKNRRPDEE----------LPCPEAV
Pma-PGBD5 262 RK-SDCTGLPRMALTGPKHPLARGQHSALIKGS--TSLLCWSNKSLFRFL-TNAYSPLKQGAIIKRKGGH-LME----------VACPQAV
Hsa-PGBD5 252 RK-SDCTGLPLSMLTNPATPPARGQYQIKMKGN--MSLICWYNKGHFRFL-TNAYSPVQQGVIIKRKSGE----------IPCPLAV
Oaf-PGBD5 252 RK-SDCTGLPLSMLANPATPQSRGQYRIKMKGN--MSLICWYNKGHFRFL-TNAYSPVQQGVIIKRKSGE----------IPCPLAV
Api1-pble 248 NR----KYNPDYVIKKKLKKG--QVIAQKSSSK--VMVLKFKDKRDLHMLSTK-HTDNMVMVPRRENKS----K----------PEVV
Api3-pble 247 NR----VKLPEITKKK-LSPG--EIIGKENSDG--IIVAKWRDKRDVTMLSTK-HNIDMIDTGKKNKKKESIVK----------PKII
Atr1-pble 247 NR----KNLPKDVMTGNLKPG--EFMAKENNNG--VTVMKWKDKRDVCLLSTK-HSIGFTKKKNKRGVET--IK----------PKIV
Psi-PGBD8 246 NR----KLNPPDVVKAKLNKN----EVIARESDTGVVVLKWKDKRDVLMLSTM-HGED-TKVVETRKG------V----------VEKPHMI
Mmu-PGBD8 246 NR----KLNPPDVVKAKLNKN----EVIARESDTGVVVLKWKDKRDVLMLSTM-HGED-TKVVETRKG------V----------VEKPHMI
Dma-PGBD8 246 NR----KLNPPDVVKAKLNKN----EVIARESDTGVVVLKWKDKRDVLMLSTM-HGED-TKAVETRKG------V----------VEKPHMI
Pva-pble 258 NT----KQMPIDLKKR-I-AK--GTTAARFYDE--LMALKWYDKKEVIMLSTF-HSDTVIEVDCRNG-----KK----------TKKPCVI
Oab1-pble 256 NR----IGICEEVRETRI-KK--GEYIARFKNK--IMLLKWRDKREVYLVSTV-HNDNIVEMQ-KRN-----VI----------KKVPEVV
Hsa-PGBD4 258 NR----KQIPNDLKKR-I-AK--GTTVARFCGE--LMALKWCDGKEVTMLSTF-HNDTVIEVNNRNG-----KK----------TKRPRVI
Uribo1 260 NR----KGLPKELVDKK-LKR--GEVHALRTDE--LLALKFADTKNVFMLTTI-HNESVVVQHRRGR-----PA----------KSKPLCC
Uribo2 260 NR----KGLPRALLDKK-LNR--GETYALRKNE--LLAIKFFDKKNVFMLTSI-HDESVIREQRVGR-----PP----------KNKPLCS
Pxu1-pble 252 NR----KDMPIMDEK---LEK--GLFDYKNTDN--LLALRWRDKKDVYMLTSA-HEPKMIVTTKTNYA--TGLN----------KSKPACV
Xtr-PGBD6 260 NR----KGLPRALINKI-LRH--GETYALRSND--LLAIKYCDTKNEYILTTI-HDESVIVEHKTTNG----PP----------KSKPLCS


α3
α4
Tni-pble 311 MYYNQTKGGVDTLDQMCSVMTCSRKT-NRWPMAL-LYGMINIACINSFIIYSHNVSSK-----GEKVQS-RKKFMRNLYMSLTSSFMRKRL
Hmel2-pble 316 SIYNKYMGGVDRSDENISHYRIGIRG-KKWY-MPLLTHMIDLAEHNAWQLYKIN--H--------GKLD-HLGFRRRVAIALIESNRKNAK
Csu-pble 316 SIYNKYMGGVDRSDENISHYRIGIRG-KKWY-MPLLTHMIDLAEHNAWQLYKIN-H---------GKLDH-LGFRRRVAIALIESNRKNAK
Smi7-pble 317 KVYNQFMGGVDRADENIDKYRASICG-KKWYSSPLLFC-FELALQNAWQLHKTY-DA--------KPMD-LLEFRRRVTCHYLETHGHPVE
Lfu-pble 316 KGYNQYMGGVDRADENIDKYRASIRG-KKWYSSPLLFC-FELALQNAWQLHKTY-DT--------KPMD-LLEFRRRVACHYLETHGYPVQ
Hvu-pble 317 KLYNQYMGGIDRADENIGQYQASICG-KKWYSSPLLFC-FEVILQNAWLLHRTI-EE--------KPLD-LLEFHRRIASNYLETYGQPPE
Ppa-PGBD3 317 KVYNQFMGGVGRADENIDKYRASIRG-KKWYSSPLLFC-FELVLQNAWQLHKTY-DE--------KPVD-FLEFRRRVVCHYLETHGHPPE
Hsa-PGBD3 317 KVYNQFMGGVDRADENIDKYRASIRG-KKWYSSPLLFC-FELVLQNAWQLHKTY-DE--------KPVD-FLEFRRRVVCHYLETHGHPPE
Ota-pble 320 TL---
Hsa-PGBD2 318 KLYQEKVGGVGRMDQNIAKYKVKIRG-MKWYSSFIGYV-IDAALNNAWQLHRICCQD--------AQVD-LLAFRRYIACVYLESNADTTS
Ppa-PGBD2 318 KLYQEKVGGVGRMDQNIAKYKVKIRG-MKWYSSFIGYV-IDAALNNAWQLHRICCQD--------AQVD-LLAFRRYIACVYLESNADTTS
Tma-PGBD2 318 RLYQEKVRGVGRMDQNIAKYKVKIRG-MKWYSSFIGYV-IDAALNNAWQLHRICCED--------AQVD-LLAFRRYVACVYLESNADTSS
Tma-PGBD1 315 KLYDECREGVAKMDQIISKYRVRIRS-KKWYSILVSYM-IDLAMNNAWQLHRACNPG--------ASLD-LLDFRRCVAHFYLEHNANLSD
Hsa-PGBD1 316 KVYDECKEGVAKMDQIISKYRVRIRS-KKWYSILVSYM-IDVAMNNAWQLHRACNPG--------ASLD-PLDFRRFVAHFYLEHNAHLSD
Oaf-PGBD1 315 KLYDECREGVAKMDQIISKYRVRIRS-KKWYSILVSYM-IDVAMNNAWQLHRSCYPG--------ASLD-LLDFRRCVAHFYLEHHANLSD
Nvi-pble 316 QAYNKIMGGVDRFDQLREVYMIERRS-KKWWHSIMYYY-IDMAIVNCYILYKCYKEG--------ECED-QLAFRIQLAKQLIGDYNGRKR
Smi2-pble 331 AVYNDVMGGVDRFDQRKERYQIGRRS-VKWWHRIFYFL-LDLGIINSFILWQVNKRN--------RNLD-QLTFRIALARQLIDGYS-SRK
Oaf-PGBD7 330 AEYNAIMGGVDHFDQSCERYAIGRRS-RKWWHRLLHFL-VDLAIVNSFTMWNCN-HG--------GRCN-QLSFRLALIRQLTVGRKRKRR
Cas-PGBD7 329 AEYNSIKGGVDHFDQRRERYAIGRRS-RKWWHRLLHFL-IDLAIVNSFIMWNCN-HG--------GRCN-QLSFRLALSRQLTVGRKRKRR
Laf-PGBD7 330 AEYNAIMGGVDRFDQRCERYAIGRRS-RKWWHRLLHFL-IDLAIVNSFIMWNCNH-G--------GRCN-QLSFRLALSRQLTVGQKRKRR
Cgi-pble 314 KLYNAYMNGVDHADQLRSTYNIARKS-LKWWKYLFL---YDVCIVNAFILMRESPNHGPRTKNNRVRTRTQLEFRKVAHQLILFS------
Bgl-pble 317 VNYNSGMFGVDLADQHRSYYPVGRVS-NKWWRYMLWYL-IDISMINSFLLMKKSR---PNDSTDSLSQNHLLFHKKVADQLLTAA-TGVTL
Pfl-pble 315 IDYNKFMGGVDLSDQLATYYPLGGKH-YKYWKYLFWFI-INRAICNAWIMYKIANT--PP---PRHRYR-QKDFRLDLSTQLVAGFTARKR
Ola-PGBD9 336 LLYQENMRGVDLCDQLRECYQVGRPC-KKWWRYFLWFY-VNLCIVNAYIVMRESRGGAPPAGFN-GKQFTQRHFRVRLAQQLI-GD-YQGA
Hbu-PGBD9 339 LLYQENMRGVDLCDQLRECYQVGRPC-KKWWRYFLWFY-VNLCIVNAYIILRESRGGAPPAGFN-GKQFTQRHFRVRLAQQLI-GD-YQGA
Pny-PGBD9 339 LLYQENMRGVDLCDQLRECYQVGRPC-KKWWRYFLWFY-VNLCIVNAYIILRESRGGAPPAGFN-GKQFTQRHFRVRLAQQLI-GD-YQGA
Tcr-pble 315 KDYNQHMGYVDKADMLKSLYEISRKC-RKWWHRIFWHF-VDVALTNGFVIYALQSEG--------THLS-LKKFRISVVDGLIGATQSSPR
Bbe-PGBD5 329 QAYQNHIKYIDKYNEKFSWYIISYKP-NKTWQQLFWTV-LSLAMNNSYIMYKMSPAY------AHSQLS-RNNFGADLLRAMAGPYYQEVP
Pma-PGBD5 338 EAFAAHLGYICKYDDKYSRYFISHKP-NKTWQQVFWLA-LSIAINNAYILYKLSPAY------QAQRYG-RAQFGERLVQQLLRIPEANGP
Hsa-PGBD5 325 EAFAAHLSYICRYDDKYSKYFISHKP-NKTWQQVFWFA-ISIAINNAYILYKMSDAY------HVKRYS-RAQFGERLVRELLGLEDASPT
Oaf-PGBD5 325 EAFAAHLSYICKYDDKYSKYFISHKP-NKTWQQVFWFA-ISIAINNAYILYKMSDAY------HVKRYS-RAQFGERLVRELLGLEDTSPS
Api1-pble 313 LDYNRCKSFIDLSDQMASYGSPLRKS-LKWYRKV-FELLLNTSVVNALCLFQKTTA---------TRIK-ITDFRSNLIKYLTFKPNMNQE
Api3-pble 315 VDYNSGKAGIDLSDQLSSYSTAVRKS-IRWYHKVATEILFGTAMVNALIVYNTINS--------DKKMK-ITQFRETLVDTILGLDNQGPQ
Atr1-pble 314 IDYNKAKAAVDLSDQMSAYSSPLRKT-LKWYKKLGIELLLNTAVVNARVLYESTTK---------KKIS-VVNFRRLLAKQLIEGSQENTG
Psi-PGBD8 312 VDYNNSKAFIDLSDQLKSYSCSLRRG-IKWFRKLAIELITGSALVNSLVLYRVV-NK--------KKIS-ITKFKEQVCLKLLQTDETSTA
Mmu-PGBD8 312 VDYNNSKAFVDLSDQLKSYSCSLRRG-IKWFRKLAIELITGSALVNSLVLYRVV-NK--------KKIS-ITKFKEQVCLKLLQTDETSTV
Dma-PGBD8 312 VDYNNSKAFIDLSDQLKSYSCSLRRG-IKWFRKLAIELITGSALVNSLVLYRVV-NK--------KKIS-ITKFKEQVCLKLLQTDETPTA
Pva-pble 323 VDYNKNMGAVDLADQMLASYPTECKRHKFWYKKFFRHL-LNITVLNSYILFKKDNPE--------HTIS-HVNFRLTLIERMLEKHHKPGQ
Oab1-pble 321 FDYNNKMGGVDMSDSIIIAYSTARKRLKKYYKKIFLHL-LDVICLNSYLIYKMN--G--------GKLS-RIHFLLEYIEDTIASYPIESR
Hsa-PGBD4 323 VDYNENMGAVDSADQMLTSYPSERKRHKVWYKKFFHHL-LHITVLNSYILFKKDNPE--------HTMS-HINFRLALIERMLEKHHKPGQ
Uribo1 326 KDYSKHMGGVDKTDQIQTYYDATRKT-RAWYKKAAIYM-IQMALYNSYVVYKAAVPG--------PKLS-YYNFLLQLLPALLFGD--VQE
Uribo2 326 KEYSKYMGGVDRTDQLQHYYNATRKT-RAWYKKVGIYL-IQMALRNSYIVYKAAVPG--------PKLS-YYKYQLQILPALLFGGVEEQT
Pxu1-pble 319 VDYNLNMGSVDMVDMVLSAINSCRKS-LKWYKKYFFHL-LDMSIYNSYILYQQVTN---------KKLK-YSKFHLILIKQILQKYPQENK
Xtr-PGBD6 327 KEYSRHIDGVNKTDQIQYY-NAARKT-KAWYKKAAFYM-IQMALNNSYVVYKATVPG--------PKLS-FYNYQVQLLPSLLFGD--VEE

Supplementary data 4 : Transposon donor sequences and primers used in the study.

a. Sequence of the *Ifp2*-TIR5'-NeoR-TIR3 donor (the cassette containing the promoter SV40 - NeoR ORF - sv40 terminator is highlighted in yellow)

TTAACCCTAGAAAGATAATCATATTGTGACGTACGTTAAAGATAATCATGCGTAAAATTGACGCAT
GTGTTTTATCGGTCTGTATATCGAGGTTTATTTATTAATTTGAATAGATATTAAGTTTTATTATAT
TTACACTTACATACTAATAATAAATTCAACAAACAATTTATTTATGTTTATTTATTTATTAAAAAA
AAACAAAAACTCAAAATTTCTTCTATAAAGTAACAAAACTTTTGGTATCGATGTCGACGAATTCTG
TGGAATGTGTGTCAGTTAGGGTGTGGAAAGTCCCCAGCTCCCCAGGCAGGCAGAAGTATGCAAAGC
ATGCATCTCAATTAGTCAGCAACCAGGTGTGGAAAGTCCCCAGGCTCCCCAGCAGGCAGAAGTATG
CAAAGCATGCATCTCAATTAGTCAGCAACCATAGTCCCGCCCCTAACTCCGCCCATCCCGCCCCTA
ACTCCGCCCAGTTCCGCCCATTCTCCGCCCCATGGCTGACTAATTTTTTTTATTTATGCAGAGGCC
GAGGCCGCCTCGGCCTCTGAGCTATTCCAGAAGTAGTGAGGAGGCTTTTTTGGAGGCCTAGGCTTT
TGCAAAAAGCTCCCGGGAGCTTGGATATCCATTTTCGGATCTGATCAAGAGACAGGATGAGGATCG
TTTCGCATGATTGAACAAGATGGATTGCACGCAGGTTCTCCGGCCGCTTGGGTGGAGAGGCTATTC
GGCTATGACTGGGCACAACAGACAATCGGCTGCTCTGATGCCGCCGTGTTCCGGCTGTCAGCGCAG
GGGCGCCCGGTTCTTTTTGTCAAGACCGACCTGTCCGGTGCCCTGAATGAACTGCAGGACGAGGCA
GCGCGGCTATCGTGGCTGGCCACGACGGGCGTTCCTTGCGCAGCTGTGCTCGACGTTGTCACTGAA
GCGGGAAGGGACTGGCTGCTATTGGGCGAAGTGCCGGGGCAGGATCTCCTGTCATCTCACCTTGCT
CCTGCCGAGAAAGTATCCATCATGGCTGATGCAATGCGGCGGCTGCATACGCTTGATCCGGCTACC
TGCCCATTCGACCACCAAGCGAAACATCGCATCGAGCGAGCACGTACTCGGATGGAAGCCGGTCTT
GTCGATCAGGATGATCTGGACGAAGAGCATCAGGGGCTCGCGCCAGCCGAACTGTTCGCCAGGCTC
AAGGCGCGCATGCCCGACGGCGAGGATCTCGTCGTGACCCATGGCGATGCCTGCTTGCCGAATATC
ATGGTGGAAAATGGCCGCTTTTCTGGATTCATCGACTGTGGCCGGCTGGGTGTGGCGGACCGCTAT
CAGGACATAGCGTTGGCTACCCGTGATATTGCTGAAGAGCTTGGCGGCGAATGGGCTGACCGCTTC
CTCGTGCTTTACGGTATCGCCGCTCCCGATTCGCAGCGCATCGCCTTCTATCGCCTTCTTGACGAG
TTCTTCTGAGCGGGACTCTGGGGTTCGAAATGACCGACCAAGCGACGCCCAACCTGCCATCACGAG
ATTTCGATTCCACCGCCGCCTTCTATGAAAGGTTGGGCTTCGGAATCGTTTTCCGGGACGCCGGCT
GGATGATCCTCCAGCGCGGGGATCTCATGCTGGAGTTCTTCGCCCACCCCAACTTGTTTATTGCAG
CTTATAATGGTTACAAATAAAGCAATAGCATCACAAATTTCACAAATAAAGCATTTTTTTCACTGC
ATTCTAGTTGTGGTTTGTCCAAACTCATCAATGTATCTTATCATGTCTGGATCCCCAgctagcTGG
CTCGAGCAATTGGCAGATCTAAGCTTGATATCTATAACAAGAAAATATATATATAATAAGTTATCA
CGTAAGTAGAACATGAAATAACAATATAATTATCGTATGAGTTAAATCTTAAAAGTCACGTAAAAG
ATAATCATGCGTCATTTTGACTCACGCGGTCGTTATAGTTCAAAATCAGTGACACTTACCGCATTG
ACAAGCACGCCTCACGGGAGCTCCAAGCGGCGACTGAGATGTCCTAAATGCACAGCGACGGATTCG
CGCTATTTAGAAAGAGAGAGCAATATTTCAAGAATGCATGCGTCAATTTTACGCAGACTATCTTTC
TAGGGTTAACCGCGGTGGAGCTCCAGCTTTTGTTCCCTTTAGTGAGGGTTAA

b. *Tcr-pble*-TIR5'-NeoR-TIR3 donor (the cassette containing the promoter SV40 - NeoR ORF - sv40 terminator is highlighted in yellow)

TTAACCCTTTGAAGCACACATTAAGGATTTAATTTATTTTTAATTTTTTACAATGAAAACTTATTT
CTAAATACATTTAAAATTTAATTTGAGCAAAAATAAAATGTTTTCATTAATATTGAGTGTTGCTAA
AAAATGGTGGTCATTCAGTGAAACCACTATGCAACTAAGCTTCACTATATTTTTACATGGTGTGGT
ATGCTACAAAGCAATTTTTAGATTCATTGAGACACAATGCCACTTCGCATGTCTTACAAGACCATT
TACTTCGGTGTGGCTGTAACTTCTGTGCTGCAGTGGGTTAACATGAATTCTGTGGAATGTGTGTCA
GTTAGGGTGTGGAAAGTCCCCAGCTCCCCAGGCAGGCAGAAGTATGCAAAGCATGCATCTCAATTA
GTCAGCAACCAGGTGTGGAAAGTCCCCAGGCTCCCCAGCAGGCAGAAGTATGCAAAGCATGCATCT
CAATTAGTCAGCAACCATAGTCCCGCCCCTAACTCCGCCCATCCCGCCCCTAACTCCGCCCAGTTC
CGCCCATTCTCCGCCCCATGGCTGACTAATTTTTTTTATTTATGCAGAGGCCGAGGCCGCCTCGGC
CTCTGAGCTATTCCAGAAGTAGTGAGGAGGCTTTTTTGGAGGCCTAGGCTTTTGCAAAAAGCTCCC
GGGAGCTTGGATATCCATTTTCGGATCTGATCAAGAGACAGGATGAGGATCGTTTCGCATGATTGA
ACAAGATGGATTGCACGCAGGTTCTCCGGCCGCTTGGGTGGAGAGGCTATTCGGCTATGACTGGGC

ACAACAGACAATCGGCTGCTCTGATGCCGCCGTGTTCCGGCTGTCAGCGCAGGGGCGCCCGGTTCT
TTTTGTCAAGACCGACCTGTCCGGTGCCCTGAATGAACTGCAGGACGAGGCAGCGCGGCTATCGTG
GCTGGCCACGACGGGCGTTCCTTGCGCAGCTGTGCTCGACGTTGTCACTGAAGCGGGAAGGGACTG
GCTGCTATTGGGCGAAGTGCCGGGGCAGGATCTCCTGTCATCTCACCTTGCTCCTGCCGAGAAAGT
ATCCATCATGGCTGATGCAATGCGGCGGCTGCATACGCTTGATCCGGCTACCTGCCCATTCGACCA
CCAAGCGAAACATCGCATCGAGCGAGCACGTACTCGGATGGAAGCCGGTCTTGTCGATCAGGATGA
TCTGGACGAAGAGCATCAGGGGCTCGCGCCAGCCGAACTGTTCGCCAGGCTCAAGGCGCGCATGCC
CGACGGCGAGGATCTCGTCGTGACCCATGGCGATGCCTGCTTGCCGAATATCATGGTGGAAAATGG
CCGCTTTTCTGGATTCATCGACTGTGGCCGGCTGGGTGTGGCGGACCGCTATCAGGACATAGCGTT
GGCTACCCGTGATATTGCTGAAGAGCTTGGCGGCGAATGGGCTGACCGCTTCCTCGTGCTTTACGG
TATCGCCGCTCCCGATTCGCAGCGCATCGCCTTCTATCGCCTTCTTGACGAGTTCTTCTGAGCGGG
ACTCTGGGGTTCGAAATGACCGACCAAGCGACGCCCAACCTGCCATCACGAGATTTCGATTCCACC
GCCGCCTTCTATGAAAGGTTGGGCTTCGGAATCGTTTTCCGGGACGCCGGCTGGATGATCCTCCAG
CGCGGGGATCTCATGCTGGAGTTCTTCGCCCACCCCAACTTGTTTATTGCAGCTTATAATGGTTAC
AAATAAAGCAATAGCATCACAAATTTCACAAATAAAGCATTTTTTTCACTGCATTCTAGTTGTGGT
TTGTCCAAACTCATCAATGTATCTTATCATGTCTGGATCCCACGTGACAATAGTATTAAATAATGA
TTTGAATACTTAGGGCCTACATAGTTATATATTTATTCCTTGAAAAATATAACAAAACCAATTAGA
AGAAACGTGTCACTTCAATAACAAAAACAAAATACATACTATATTGGAGGCATGAATACCAACAAA
AATTACTAATAATAAATTTTTACAAGAATAAAACTAGCAAATAAATTCAGATAGATGATTCATAAG
TGCTGCCATACAGTTAAATGTATCTCAAATAATTAAATAAAAGGTGGTTACCAAGTGAAACCACTA
TGCTTCAAAGGGTTAA

c. Sequences used as baits to identify transposon breakpoints in analyses done with lumpy.

Sequence of the *Ifp2*-TIR5'-NeoR-TIR3 sequence flanked upstream and downstream by 100-bp plasmid regions. The cassette containing the promoter SV40 - NeoR ORF - sv40 terminator was highlighted in yellow. TTAA TSD were bolded while 100-bp plasmid backbone sequences were typed in blue.

ATTAAGTTGGGTAACGCCAGGGTTTTCCCAGTCACGACGTTGTAAAACGACGGCCAGTGAGCGCGC
GTAATACGACTCACTATAGGGCGAATTGGGTACC**TTAA**CCCTAGAAAGATAATCATATTGTGACGT
ACGTTAAAGATAATCATGCGTAAAATTGACGCATGTGTTTTATCGGTCTGTATATCGAGGTTTATT
TATTAATTTGAATAGATATTAAGTTTTATTATATTTACACTTACATACTAATAATAAATTCAACAA
ACAATTTATTTATGTTTATTTATTTATTAAAAAAAAACAAAAACTCAAAATTTCTTCTATAAAGTA
ACAAAACTTTTGGTATCGATGTCGACGAATTCTGTGGAATGTGTGTCAGTTAGGGTGTGGAAAGTC
CCCAGCTCCCCAGGCAGGCAGAAGTATGCAAAGCATGCATCTCAATTAGTCAGCAACCAGGTGTGG
AAAGTCCCCAGGCTCCCCAGCAGGCAGAAGTATGCAAAGCATGCATCTCAATTAGTCAGCAACCAT
AGTCCCGCCCCTAACTCCGCCCATCCCGCCCCTAACTCCGCCCAGTTCCGCCCATTCTCCGCCCCA
TGGCTGACTAATTTTTTTTATTTATGCAGAGGCCGAGGCCGCCTCGGCCTCTGAGCTATTCCAGAA
GTAGTGAGGAGGCTTTTTTGGAGGCCTAGGCTTTTGCAAAAAGCTCCCGGGAGCTTGGATATCCAT
TTTCGGATCTGATCAAGAGACAGGATGAGGATCGTTTCGCATGATTGAACAAGATGGATTGCACGC
AGGTTCTCCGGCCGCTTGGGTGGAGAGGCTATTCGGCTATGACTGGGCACAACAGACAATCGGCTG
CTCTGATGCCGCCGTGTTCCGGCTGTCAGCGCAGGGGCGCCCGGTTCTTTTTGTCAAGACCGACCT
GTCCGGTGCCCTGAATGAACTGCAGGACGAGGCAGCGCGGCTATCGTGGCTGGCCACGACGGGCGT
TCCTTGCGCAGCTGTGCTCGACGTTGTCACTGAAGCGGGAAGGGACTGGCTGCTATTGGGCGAAGT
GCCGGGGCAGGATCTCCTGTCATCTCACCTTGCTCCTGCCGAGAAAGTATCCATCATGGCTGATGC
AATGCGGCGGCTGCATACGCTTGATCCGGCTACCTGCCCATTCGACCACCAAGCGAAACATCGCAT
CGAGCGAGCACGTACTCGGATGGAAGCCGGTCTTGTCGATCAGGATGATCTGGACGAAGAGCATCA
GGGGCTCGCGCCAGCCGAACTGTTCGCCAGGCTCAAGGCGCGCATGCCCGACGGCGAGGATCTCGT
CGTGACCCATGGCGATGCCTGCTTGCCGAATATCATGGTGGAAAATGGCCGCTTTTCTGGATTCAT
CGACTGTGGCCGGCTGGGTGTGGCGGACCGCTATCAGGACATAGCGTTGGCTACCCGTGATATTGC
TGAAGAGCTTGGCGGCGAATGGGCTGACCGCTTCCTCGTGCTTTACGGTATCGCCGCTCCCGATTC
GCAGCGCATCGCCTTCTATCGCCTTCTTGACGAGTTCTTCTGAGCGGGACTCTGGGGTTCGAAATG
ACCGACCAAGCGACGCCCAACCTGCCATCACGAGATTTCGATTCCACCGCCGCCTTCTATGAAAGG

TTGGGCTTCGGAATCGTTTTCCGGGACGCCGGCTGGATGATCCTCCAGCGCGGGGATCTCATGCTG
GAGTTCTTCGCCCACCCCAACTTGTTTATTGCAGCTTATAATGGTTACAAATAAAGCAATAGCATC
ACAAATTTCACAAATAAAGCATTTTTTTCACTGCATTCTAGTTGTGGTTTGTCCAAACTCATCAAT
GTATCTTATCATGTCTGGATCCCCAgctagcTGGCTCGAGCAATTGGCAGATCTAAGCTTGATATC
TATAACAAGAAAATATATATATAATAAGTTATCACGTAAGTAGAACATGAAATAACAATATAATTA
TCGTATGAGTTAAATCTTAAAAGTCACGTAAAAGATAATCATGCGTCATTTTGACTCACGCGGTCG
TTATAGTTCAAAATCAGTGACACTTACCGCATTGACAAGCACGCCTCACGGGAGCTCCAAGCGGCG
ACTGAGATGTCCTAAATGCACAGCGACGGATTCGCGCTATTTAGAAAGAGAGAGCAATATTTCAAG
AATGCATGCGTCAATTTTACGCAGACTATCTTTCTAGGG**TTAA**CCGCGGTGGAGCTCCAGCTTTTG
TTCCCTTTAGTGAGGGTTAATTGCGCGCTTGGCGTAATCATGGTCATAGCTGTTTCCTGTGTGAAA
TTGTTATCCGCT

Sequence of the *Tcr-pble*-TIR5'--TIR3 sequence flanked upstream and downstream by 100-bp plasmid regions. TTAA TSD were bolded while 100-bp plasmid backbone sequences were typed in blue.

CGTTGTAAAACGACGGCCAGTGAGCGCGCGTAATACGACTCACTATAGGGCGAATTGGGTACGGCC
GTCAAGGCCAAGCTTCCCGGGATACTTACCATGG**TTAA**CCCTTTGAAGCACACATTAAGGATTTAA
TTTATTTTTAATTTTTTACAATGAAAACTTATTTCTAAATACATTTAAAATTTAATTTGAGCAAAA
ATAAAATGTTTTCATTAATATTGAGTGTTGCTAAAAAATGGTGGTCATTCAGTGAAACCACTATGC
AACTAAGCTTCACTATATTTTTACATGGTGTGGTATGCTACAAAGCAATTTTTAGATTCATTGAGA
CACAATGCCACTTCGCATGTCTTACAAGACCATTTACTTCGGTGTGGCTGTAACTTCTGTGCTGCA
GTGGGTTAACATGAATTCGAGCTCGGATCCCACGTGACAATAGTATTAAATAATGATTTGAATACT
TAGGGCCTACATAGTTATATATTTATTCCTTGAAAAATATAACAAAACCAATTAGAAGAAACGTGT
CACTTCAATAACAAAAACAAAATACATACTATATTGGAGGCATGAATACCAACAAAAATTACTAAT
AATAAATTTTTACAAGAATAAAACTAGCAAATAAATTCAGATAGATGATTCATAAGTGCTGCCATA
CAGTTAAATGTATCTCAAATAATTAAATAAAAGGTGGTTACCAAGTGAAACCACTATGCTTCAAAG
GG**TTAA**CTCGAGTAAGTATCACGTGAAGCTTTGCAAGCTCCAGCTTTTGTTCCCTTTAGTGAGGGT
TAATTGCGCGCTTGGCGTAATCATGGTCATAGCTGTTTCC

Sequence of the *Ifp2*-TIR5'-NeoR-TIR3 sequence flanked upstream and downstream by 100-bp plasmid regions. This sequence was that contained in the *Ifp2* plasmid construct used in Henssen et al, 2015 to monitor integration assays in HEK293 cells. The cassette containing the promoter EF1 - NeoR ORF - sv40 terminator was highlighted in yellow. TTAA TSD were bolded while 100-bp plasmid backbone sequences were typed in blue.

TTGAACCCGTGGAGGACGGGCAGACTCGCGGTGCAAATGTGTTTTACAGCGTGATGGAGCAGATGA
AGATGCTCGACACGCTGCAGAACACGCAGCTAGA**TTAA**CCCTAGAAAGATAATCATATTGTGACGT
ACGTTAAAGATAATCATGTGTAAAATTGACGCATGTGTTTTATCGGTCTGTATATCGAGGTTTATT
TATTAATTTGAATAGATATTAAGTTTTATTATATTTACACTTACATACTAATAATAAATTCAACAA
ACAATTTATTTATGTTTATTTATTTATTAAAAAAAACAAAAACTCAAAATTTCTTCTATAAAGTAA
CAAAACTTTTATGAGGGACAGCCCCCCCCCAAAGCCCCCAGGGATGTAATTACGTCCCTCCCCCGC
TAGGGGGCAGCAGCGAGCCGCCCGGGGCTCCGCTCCGGTCCGGCGCTCCCCCCGCATCCCCGAGCC
GGCAGCGTGCGGGGACAGCCCGGGCACGGGGAAGGTGGCACGGGATCGCTTTCCTCTGAACGCTTC
TCGCTGCTCTTTGAGCCTGCAGACACCTGGGGGGATACGGGGAAAAGGCCTCCACgGCCaaggatc
tgcgatcgctccggtgcccgtcagtgggcagagcgcacatcgcccacagtccccgagaagttgggg
ggaggggtcggcaattgaacgggtgcctagagaaggtggcgcggggtaaactgggaaagtgatgtc
gtgtactggctccgccttttttcccgagggtgggggagaaccgtatataagtgcagtagtcgccgtg
aacgttcttttcgcaacgggtttgccgccagaacacagctgaagcttcgaggggctcgcatctct
ccttcacgcgcccgccgccctacctgaggccgccatccacgccggttgagtcgcgttctgccgcct
cccgcctgtggtgcctcctgaactgcgtccgccgtctaggtaagtttaaagctcaggtcgagaccg
ggcctttgtccggcgctcccttggagcctacctagactcagccggctctccacgctttgcctgacc
ctgcttgctcaactctacgtctttgtttcgttttctgttctgcgccgttacagatccaagctgtga
ccggcgcctactctagagctagcgaattcgaatttaaatcggatccGCGGCCGCgcccctctccct

ccccccccccctaacgttactggccgaagccgcttggaataaggccggtgtgcgtttgtctatatgt
tattttccaccatattgccgtcttttggcaatgtgagggcccggaaacctggccctgtcttcttga
cgagcattcctaggggtctttcccCtctcgccaaaggaatgcaaggtctgttgaatgtcgtgaagg
aagcagttcctctggaagcttcttgaagacaaacaacgtctgTagcgacccttgcaggcagcgga
acccccacctggcgacaggtgcctctgcggccaaaagccacgtgtataagatacacctgCaaagg
cggcacaaccccagtgccacgttgtgagttggatagttgtggaaagagtcaaatggctctcctcaa
gcgtattcaacaAggggctgaaggatgcccagaaggtaccccattgtatgggatctgatctggggc
ctcggtgcacatgctttacatgtgtttagtCgaggttaaaaaaacgtctaggccccccgaaccacg
gggacgtggttttcctttgaaaaaCACGATGATAATATGGCCACAACCATGgcgtccggaATGATT
GAACAAGATGGATTGCACGCAGGTTCTCCGGCCGCTTGGGTGGAGAGGCTATTCGGCTATGACTGG
GCACAACAGACAATCGGCTGCTCTGATGCCGCCGTGTTCCGGCTGTCAGCGCAGGGGCGCCCGGTT
CTTTTTGTCAAGACCGACCTGTCCGGTGCCCTGAATGAACTGCAGGACGAGGCAGCGCGGCTATCG
TGGCTGGCCgCGACGGGCGTTCCTTGCGCAGCTGTGCTCGACGTTGTCACTGAAGCGGGAAGGGAC
TGGCTGCTATTGGGCGAAGTGCCGGGGCAGGATCTCCTGTCATCTCACCTTGCTCCTGCCGAGAAA
GTATCCATCATGGCTGATGCAATGCGGCGGCTGCATACGCTTGATCCGGCTACCTGCCCATTCGAC
CACCAAGCGAAACATCGCATCGAGCGAGCACGTACTCGGATGGAAGCCGGTCTTGTCGATCAGGAT
GATCTGGACGAAGAGCATCAGGGGCTCGCGCCAGCCGAACTGTTCGCCAGGCTCAAGGCGCGCATG
CCCGACGGCGAGGATCTCGTCGTGACCCATGGCGATGCCTGCTTGCCGAATATCATGGTGGAAAAT
GGCCGCTTTTCTGGATTCATCGACTGTGGCCGGCTGGGTGTGGCGGACCGCTATCAGGACATAGCG
TTGGCTACCCGTGATATTGCTGAAGAGCTTGGCGGCGAATGGGCTGACCGCTTCCTCGTGCTTTAC
GGTATCGCCGCTCCCGATTCGCAGCGCATCGCCTTCTATCGCCTTCTTGACGAGTTCTTCTGAgtc
gacaatcaacctctggattacaaaatttgtgaaagattgactggtattcttaactatgttgctcct
tttacgctatgtggatacgctgctttaatgcctttgtatcatgcGTTAACTaaactTGTTTATTGC
AGCTTATAATGGTTACAAATAAAGCAATAGCATCACAAATTTCACAAATAAAGCATTTTTTTCACT
GCATTCTAGTTGTGGTTTGTCCAAACTCATCAATGTATCTTATCATGTCTGGAATTGACTCAAATG
ATGTCAATTAGTCTATCAGAAGCTCATCTGGTCTCCCTTCCGGGGGACAAGACATCCCTGTTTAAT
ATTTAAACAGCAGTGTTCCCAAACTGGGTTCTTATATCCCTTGCTCTGGTCAACCAGGTTGCAGGG
TTTCCTGTCCTCACAGGAACGAAGTCCCTAAAGAAACAGTGGCAGCCAGGTTTAGCCCCGGAATTG
ACTGGATTCCTTTTTTAGGGCCCATTGGTATGGCTTTTTCCCCGTATCCCCCCAGGTGTCTGCAGG
CTCAAAGAGCAGCGAGAAGCGTTCAGAGGAAAGCGATCCCGTGCCACCTTCCCCGTGCCCGGGCTG
TCCCCGCACGCTGCCGGCTCGGGGATGCGGGGGGAGCGCCGGACCGGAGCGGAGCCCCGGGCGGCT
CGCTGCTGCCCCCTAGCGGGGGAGGGACGTAATTACATCCCTGGGGGCTTTGGGGGGGGGCTGTCC
CTGATATCTATAACAAGAAAATATATATATAATAAGTTATCACGTAAGTAGAACATGAAATAACAA
TATAATTATCGTATGAGTTAAATCTTAAAAGTCACGTAAAAGATAATCATGCGTCATTTTGACTCA
CGCGGTCGTTATAGTTCAAAATCAGTGACACTTACCGCATTGACAAGCACGCCTCACGGGAGCTCC
AAGCGGCGACTGAGATGTCCTAAATGCACAGCGACGGATTCGCGCTATTTAGAAAGAGAGAGCAAT
ATTTCAAGAATGCATGCGTCAATTTTACGCAGACTATCTTTCTAGGG**TTAA**TCTAGCTGCATCAGG
ATCATATCGTCGGGTCTTTTTTCCGGCTCAGTCATCGCCCAAGCTGGCGCTATCTGGGCATCGGGG
AGGAAGAAGCCCGTGCCTT

d. Primers used to amplify the Mm524 cDNA by RT-PCR.
Mn524 (Accession N°XM_006530804.1)
Forward primer (flanked with *Eco*RI) : 5'-CCGAAGAATTCATGGCCGAGGGCGGTGG-3'
Reverse primer (flanked with *Xba*I) : 5'-CCGTTTCTAGATCAGTGGGCTGGCGATGAG-3'

e. Primers to amplify loci containing *Ifp2* or *Tcr-pble* insertions
First linear amplification (B = biotin)
(B)-NeoR 5' : (B)-TTCTTCGCCCACCCCAACTTGTTTATTGC-3'
(B)-NeoR 3' : (B)-GACTTTCCACACCCTAACTGACA-3'

Linkers
Oligo I : 5'-GACCCGGGAGATCTGAATTCAGTGGCACAGCAGTTAGG-3'
Oligo II DpnII : 5'-GATCCCTAACTGCTGTGCCACTGAATTCAGATC-3'
Oligo II PNB : 5'-CATGCCTAACTGCTGTGCCACTGAATTCAGATC-3'

First round of nested PCR
Primer anchored in linker
LC1 : 5'-GACCCGGGAGATCTGAATTC-3'

Primers anchored in the 5' or 3' end of the *IFP2* vector
(B)-TIR-UTR 5' PB III : (B)-CGATAAAACACATGCGTCAAT -3'
(B)-TIR-UTR 3' PB III : (B)-GCTATTTAGAAAGAGAGAGCA -3'

Primers anchored in the 5' or 3' end of the *Tcr-pble* vector
(B)-TIR-UTR 5'Tcr-pbleII : (B)-GTTTTCATTAATATTGAGTGTTGC-3'
(B)-TIR-UTR 3'Tcr-pbleII : (B)-GCAGCACTTATGAATCATCTATC-3'

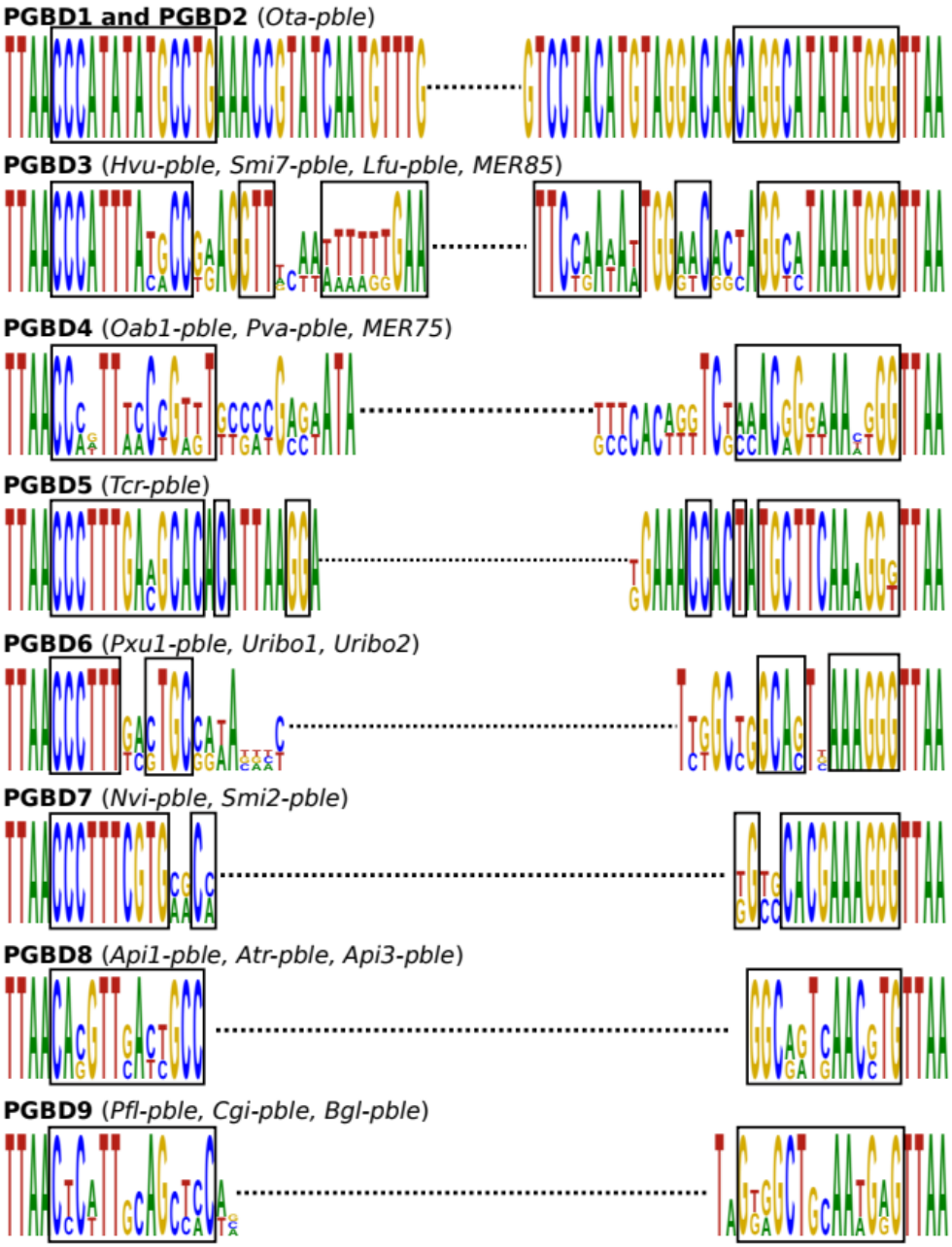
Second round of nested PCR
Primer anchored in linker
LC2 : 5'-GATCTGAATTCAGTGGCACAG-3'

Primers anchored in the 5' or 3' end of the *IFP2* vector
TIR-UTR PB 5'III : 5'-CGATAAAACACATGCGTCAAT -3'
TIR-UTR PB 3'III : 5'-GCTATTTAGAAAGAGAGAGCA-3'

Primers anchored in the 5' or 3' end of the *Tcr-pble* vector
TIR-UTR 3' Tcr-pble III : 5'-CACCTTTTATTTAATTATTTGAGATAC-3'
TIR-UTR 5'Trc-pble III : 5'-CAATGAAAACTTATTTCTAAATAC-3'

Supplementary data 5. Nucleotidic motifs conserved at the ends of *pbles*.

Supplementary data 5a: Gapped local alignment of the 5' and 3' *pble* extremities. The alignment was performed by GLAM2, and the logo was generated by Skyline. The black rectangles indicate similar nucleotides between the two ends. Sequences of *pble* were supplied by PGBD clades described in Figure 3.



Because motif identity could only be observed when a pool of sequences was analyzed, we sought to investigate the conservation of the 24 *pbles* used in Figure 3. For each PGBD clade, the conservation of sequence motifs within the sequence termini of their closest *pble* relatives was analyzed with GLAM2 [47]. This software was preferred to MEME or RSAT [a-c], which only allowed identification of conserved motif tracks that did not contain indels of several nucleotides. GLAM2 allowed identification of conserved motif tracks containing indels of 1 to 10 nucleotides in size in windows as large as 100-bp.

For the PGBD1 and PGBD2 relatives, conserved motifs were found within the termini of their relatives. They displayed a 14-bp conserved motif juxtaposing the target site duplication (TSD), TTAA (motifs boxed with a black line).

To analyze PGBD3 relatives, the human *pble MER85* was added to *Smi7-pble*, *Hbu-pble* and *Lfu-pble* because its TIRs flanked *pgbd3* at both ends (supplementary data 2). It has previously been shown that PGBD3 binds to these *pbles* [d]. The"CCCATTTA" motif was perfectly conserved between the 4 *pbles* relatives juxtaposed with a TTAA TSD. TIRs at the ends of these *pbles* were the longest (~32 pb) among *pbles* and displayed patchy conservation in their inner regions.

For the PGBD4 relatives, human *pble* TIRs of *MER75* flanking both ends of the *pgbd4* gene (supplementary data 2) were aligned with sequence ends of *Pxu1-pble*, *Oab1-pble*, and *Pva-pble*. These were found to display an outer motif corresponding to CCNNTTN that was less conserved than those of *pble* relatives of other *pgbds* since only 7/14 residues were conserved (motifs boxed with a black line).

For the PGBD5 relative, TIRs of *Tcr-pble* are shown in Fig. 7. Their 22-bp TIRs displayed patchy conservation in their inner regions.

For the PGBD6 relatives, the ends of three *pble* relatives were used: *PxuI-pble*, *Uribo1* and *Uribo2*. Their outer ends displayed limited conservation of only 6 residues. These were the smallest ends (12 bp) among the *pbles* investigated here.

For the PGBD7 relatives, conservation between the ends of *Nvi-pble* and *Smi2-pble* was strong (the conserved motif was CCCTTTCGTG), and inner regions displayed patchy sequence conservation (Fig. 3, motifs boxed with a black line).

For the PGBD8 and PGBD9 relatives, inverted ends were 13 and 14 bp in length respectively and their sequences displayed few conserved positions within each set of *pble* relatives.

In conclusion, beyond the conservation of the outer CNNNTN motif at the outer ends, we did not find strongly conserved motifs between the TIRs of the nine sets of *pbles* we analyzed. This sequence conservation between TIRs was in agreement with that observed between all pairs of complete *pble* sequences (supplementary Table 2).

Supplementary references

a. Bailey,T.L., Johnson,J., Grant,C.E., Noble,W.S. (2015) The MEME Suite. Nucl. Acids. Res. 43, W39-W49.

b. Thomas-Chollier, M., Herrmann, C., Defrance, M., Sand, O., Thieffry, D., van Helden, J. (2011). RSAT peak-motifs: motif analysis in full-size ChIP-seq datasets Nucl. Acids. Res. 40, e31.

c. Thomas-Chollier, M., Darbo, E., Herrmann, C., Defrance, M.,Thieffry, D., van Helden, J. (2012). A complete workflow for the analysis of full-size ChIP-seq (and similar) data sets using peak-motifs. Nat. Protoc. 7, 1551-1568.

d. Gray, L.T., Fong, K.K., Pavelitz, T., Weinern A.M. (2012) Tethering of the conserved piggyBac transposase fusion protein CSB-PGBD3 to chromosomal AP-1 proteins regulates expression of nearby genes in humans. PLoS. Genet. 8, e1002972.

Supplementary data 5b: Location of TIR and IR in nucleic acid sequences of *pble*s. Sequences of *pble* were supplied by PGBD clades described in Figure 3. For each *pble* sequence, the TTAA TSD motifs are in bold and the TIRs are in red. Regions containing the internal IR (IIR) that were not conserved in sequence between both ends, and the subterminal IR (STIR) that are conserved in sequence between both ends were underlined. IR in IIR and STIR were typed in blue and highlighted yellow. The potential ability of the highlighted IIR and STIR to adopt SII/SIII was verified using publicly available tools (https://nonb-abcc.ncifcrf.gov/apps/nBMST/default/).

Clade PGBD1 et PGBD2

>Ota-pble (TIR + IIR)

TTAACCCATATATGCCTGAAACCGTATCAATGTTTGTTTTTGCGCAAAAATGTTTTTTAATAATCATAATGGTCAAGAAG
AACACGCTAGTAACATTTTTGCAAAACAATATATTTTCCAACTTAGTGTTACAATGTGTACTACATATAGGACACTAGGC
ACTTGTATGAACATTTTATATACTATTTTTATTTATATTGTATTACATTTTGTATGGTAACTGATAAAACACTTAACATG
AACACCTTTGTTACACTTGACGCAGCGTGTGGTACATTTGGTATGACATAAGCCACAACGTGTTTGATTTTCTTGCGGAC
AAACAAAGTGATATAAGCCATCGTACCGAGTTTCCGTGGTTACTTGACAAGGTAAATTTCGTTTTTTCGAAGTTTTTGAG
TAATGCTCCATATAAAATCGAGCAATGCGACGACGGAAGGCAAGCAGATCAAGGTCATTATATTTTTGTCGATATAGCTG
GAATGCATTATTTACTGCCATATCAATGAAGAATGTTATTAGAGACCAATACCACTTTTTTCCATGTATAGCGATCCTAT
ATTTACTGACATTTTGATCCATCCTATCTACTCCACCCATATTTTCATTATAAAGTTTGATGATGTGGCTGGGCAATCTG
AATGCGTTTTCGTTCTGCTCGGGAGAAACGAGTAGTAGAATGTACTGGATTTATTCCAACTGAATTAGAGCACAAACTGA
CCACATTATTGTCATTCCATTTGCATACTATTAGCTCTTGTGATTCATCCACTCTGTAGTCGTAGAAACCCCTGCTTTCC
TTTTTCATATTCTGTTTTGTTAGTAAAGGACAATGCTCGGACCTGTTCTCTTTTATCGTTCCTGTAGCTCTTATTCCTAT
TTTTTGCAACTCTGAAATCAATTTTATGCTAGTGAAAAAGTTGTCACAAAATAAATGAAATGGGTTCGGATTTTGAGATT
TTAAAATTTTAGCAAATTTATAAACAAGCTGCCCTCCCAAACCCAGATTTAGATCGTCCTTTTCTTTATATCCGTTTTTG
CCTTGATACGGATCAACGTTAACAACATAGCCTAATCGTGTTGCACCACACCAAATCTTGAACCCAAATCGTATCGGTTT
GCCGCGTATAAATTGCTTGAGTCCGTGGCGTCCAAAATATTTGCACATGGATTCGTCAAAACTGTAGTACTCTTCAATAG
GCGCATAACATAAAAATTTCTCATTTATCATTGTCACTAATGGACGTATTTTAGCATATTTATCTTCCGCATCAAGTTCA
TAGTTGTTGGCTAAATAAAAATGGGATAAGATAGCTTCAAATCGATCACGTTTCATGCCATTTATTACAATTTCGTGATG
TGCATCACTAGCATTTTCCCAGAACATCCTTCTTCTGGGAACCTGGGTATAGCCACTATAGAGCAAAATACCAATAAAAC
ATTTGATTTCTTCAGTGCAAACGTTCAAAGCTCGTTTTTTTTGCAGCGAATATCGATTCGTTTCCAGAACGATTTTATCA
ACTAGTTCTGCATCAAAAAATAGTTCAAAAAAATCTATTGCAGTGCTTGCTTCATCTGTAGTTTCTGCATTATTATTGGT
GGCTTTCCAGTCGGGAAACTCTGCATTGATGTCTCTTTTACGCCACTTGCGCTTGCCAATCTGTTGCACTCTCAATCTCT
TCTGTAGTTCTGCTAATGATATATTTTCTTCCTCATCAGTGTCGCTATCACTTAAATTTTCCCCAACTTCTTCTACTTCT
GCTTCTTGTCTCAAAATGCTACCAGGTAAATTATTGATTGATCATTCATCTTCTTCCGCTGAGTCTTCATCGGTTATATA
TTCCGTTGCATTTACGGGAGGTAGTATATATACATCCGCTGCTTTTGTAGAATTTTCTTCAATTCGTGACAACAAATCGC
AAAGTCTGCCCCTGTTTATATTTTTTCGTGAATAAAATATATTGGGTGTAATTGGATCCGGTTCCCTGAAAGTGGGGAAA
AGTTAACCTATATGCCTGCTGTCCTACATGTAGTACACCGAGTTTCAATCTACTGCAAATACGCCATCACGGATGCGTAC
AGTTTGAGGTTAAGTATACACATTTGTATGTTCATTGGTTTGTAATTTGTAAAAGAAAAAGTACCCAAAAAAAGAGCATC
TAGAAGAACACTTACCTTGATGGTAGTGAACTGCTCGAACTCGCCATTTTTTAGAAACCGTGGTAATTACTCACTTATTC
AAAACATTTAAATGCAAATGACACCTTGGTAAATTATATATAGTAGATGCCTACACATTCAAGAAACGATTATAAATTGC
GTCACCTGTTGTTTTCACTGAAAAACTTTAGTAGGGCTCTATAAATGCCTGTCCTACATGTAGGACAGCAGGCATATATG
GG**TTAA**

Clade PGBD3

>Lfu-pble (TIR + STIR)

TTAACCCATTTATGCCGGAGGTTTAAATTTTTGAATTTGTTTATGGTATTGATATATATTAGCTTAAAAAATATAAAAAG
CAAGAATTTAAAGGAAAATTAAATTTATGTCATCCAGTATGTGAGTTATATGGTGTTCCAAAAATGGAACGCTAGGCAAA
AACAGTACGTAACATACAAAATTCCACTTAATGTCTCAAAATCCACTATTGGGTATGATATTGAAGTGAACACTTCACAT
GAAGGGCAACATTACATTTTTCACAACAAAAAGTGGTGTTTTTGTGGCATTCAGCGCAACGAGTTTGCTTTGACTGCTTT
GCGATGACATGATTTAGGCCATCATAATGGGAGTCAATGTTGCGCTACTGAGGTCTTCCTTTCTTCCCAGGTTGGACAGG
ATAACCATGGGTCTCCAGATAATGACATGCTACACGGCGACGAAACTCCAAAAGATCCATTGGCTTTGTATCATAAGTTT
TATGCAGCTGCCAAGCATTTTGTAGGGCTAATTCGAAGCAGAACAGCAGAGGGCTTGAGTACCATTTCTTGCCACGAATT
GATGCTCTATACTTGTCAATGTTTTCATCAGCTCTGTCTACACCTCCCATGTACTGGTTATACCCTTTGATCATATTTGG
CTGCTGAACTTGGATCTTCTTTTTCTGTTTTTGAGAATAACGCTTGACAAGACACAGTGGATTGACGCCAGCTCCAGATG
ATGCAACTGTCACAACACTGTTATCATTCCATCTGCAGACGATATTGCCTTTGCCATCAATTCTGTAATCAAATGCACCT
CTTTCTTTCTTCTTTAGAGCAACTTCTGATTCTAATGGGGGTTTGTCAATTCTATCTTTTCTCACTGTGCCGGTTGCTTG
ATGACCCATTGAACTGAGCTTATCAAGAAGTACAATACTAGTGAAAAAGTTATCAAATACAAAATGATATTGTCCAGGGT
GCTCCTTTGTGAGTGCCTCAGTAAACTGTAGGACAACGGATGCACCGACACCATATCCCTCATAATTGGTGTTTGGGTTT
TTGCCTTGATATGGTTGAAACCAACAAATGTAACCTAGGCGTGTGGCACCACACCAAAACTTGTACCCAAACCGAATAGG
TTTTCCACGGATGAACTGTTTGCAGCCATGACGACCAAAGTAAGGAATCATGGACTCATCAAAGCTGAAGCATGTTTCAT
TTGGTACAAATAGCATGCATCTCTCATTGAGTTTTATAATAAGAGGTCTTAATTTTGCAAACTTGTCTGTTGGGTCCAAA
TTTGCATTATCTGCAACATGCAAGTGAGAAAATATGGTTTCGAAGCGGTCACGTCTCATGGCACCACTGACAAGGCCATT
GTGTGCATCCGCTCTTTGCTCCCAGAACATACGTCGTCGAGGAACTGAGACATAACCGCTCAGAAAGATAATACCAAGAA

AACACTTCAATTCTGAGTTAGACAATCCAAGGTTAACACCCTTACTTGCAGCATATAAGTTCGAGTATCATACAATGAGT
TCAATGACCTCGTCATTCAAAAAACGTTCGAGAAGTTCTGTAGGAGTTCCCACTTCAGGCATTTGTAGATTTGTAACTCT
ACCTGTAACAGGCCTTGCCTGTAGGTTAGCTTTTTTCCATTCACGCACAATTTTTGAAAGCTTCTTTCGCTTTGTTGGAA
ATGACTAATGAGCAATAGAGGTAGTAGCAACAGCATCTGAGGGAGAATCATTTTCAGGGACGAATGAAGAATCATCTGGA
TCAGAACGATCTAGTACAAGATACGCAGGGGCACGCAACAAAGAACCAGGCAGATTATTTATTGTCCCACCTTCTTCATC
ACCAGAGTCCTCGTCAGAAACAGGAGCTGTAGCATTTTCAGGAGGTTCTATGACAATTGCACTTGCTTCTATGTCATCAT
CTGTCTCTAACAGGTCAGCTATTTCATGTACAGTTAACGTTCTGCGCATCCTGAAACTTCAAAGTGGTGAAAACAAAGAT
AAAGGCCATATGTGTAAATAAAGAAATATACGGCAAAGTCTATGTAGTGGTTTTGCCTAGTGTTCCAAAAATGGAACACG
CAAACTTTCTAACCTCTAACTTATATAAGAAACAAAATTTCTGTATTTTCTGATTGGAACATGAAACAGTGAATAGTCTA
GTATTAACAGAAAAAAAATTACAGCTATAGATTTTAAATCACTTAAGTTATAATTACCGTTTGTTAAGTGGGCGTTGAAA
AATCGACTCAGAAAGACAGAGCAAAGTTTGACAACGTCCATGCAGCTCTGTGTTCACAGCAGTGACTAACTCTGCATGTT
CGGTAATGCTATATACCCATACAAAGAAGGCAGGATGTTGTGACTTACAACAAAAAGTTCTACATAAGTGGTAGGCTATA
AATAACTTTGTGTGACTGCCGTTCCAATATTGGAACGCCAGGCATAAATGGG**TTAA**

>Smi7-pble (TIR + STIR)
TTAACCCATTTATGCCGAAGGTTAAAAAATTGGAATTTTTGCATGAAATTTCTATATGTTAGGTAAAAAAATATAATAAA
CAATGATTTGAAGAGGAAATTTTTTTATTTTAATTAGTTTATAAGTTATATGGTGTTCCAAAAATGGAACACTAGGCAAA
ACCATTGCTATAATGTTACAATGTATAGAATCATGGGTTTCTCTAAAAATTTGGTTATTCGGATTATTCGGTATTGTAAT
TGGAAGAACACTTCACATGTAAAGCCACATTACATTTTCCACATCGAAAAGTGGTGTTTTTATGACATTGAACACAGCGA
GTTTGCTTTGCTTGCTTTACGATCGTATGATTTATGCCATCATAGCATGAATCAATGTTTCGTTTCTGAGAGGGTTTTCC
TTTCCTGCCAGGTTCTACAGGATGACCATGAGTCTCAAGATAATGGCATGTTACACGACGACGAAATTCTAAAAGATCCA
TTGGTTTTGCATCATAAGTTTTGTGTAATTGCCAAGCATTTTGTAAGGCCAATTCGAAACAGAACAACAATGGACTTGAA
TACCATTTCTTTCCACAAATTGATGCTCTGTATTTGTCAATGTTTTCGTCCGCTCTATCCACACCTCCCATAAATTGGTT
ATATACTTTTATCATGTTTGGCTGCAGAACTTGAATTTTCTTTTTTTGTTCTTGTGAATAACGATTTACAAGACACAGTG
GATTTACACCAGCACCAGATGATGCAACAGTCACAACACTGTTGTCATGCCATCTACAGACAATATTGCCTTTGCCATCA
ACTCGATAATCGAATGAGCCTCTTTCCATTTTCTTTAGGGCAACATCCGATTGCAATGGTGGTTTATCAATACGACCTTT
TCTAACTGTGCCTGTTGCCTGATGACCCATTGAACCGAGTTTATCAAGCAGGGCAATACTTGTGAAAAAATTATCAAAAA
CAAAATGATATTGCCCAGGGTGTTGCTCTGTTAGTGCATCCGTAAACCGTAAGACAACAGAAGCACCAACACCATACTCC
TCATATTTGTTTGCCAGATTTTTACCTTGATAGGGTTCGAACCAACAAATGTAACCTAGACGAGTTGCCGCACACCAAAA
CTTGTATCCAAAACGAATAGGCTTTCCTCTTATAAATTGTTTGCATCCGTGTCGACCAAAGTAAGGAACCATGGATTCAT
CAATGCTGAAATGTGTTTCATTCGGAACAAACTTCATACACAGCTCATTTAGTTTGTTTATAAGAGGTCGTAATTTAGCA
AACTTGTCAAGTGGGTCTAAATTTGCATTGTCTGCAACATGGAGATTTGAAAATATTGTTTCAAAGCGGTCACGTCTCAT
GGCACTACTAACAAGGGCATTTTGTGAATCGGTTCTTTGTTCCCAGAACATGCGCCGTCTAGGAATTGAGACATATCCAC
TCAGAAATATAATCCCCAGGAAACATTTCAATTCAGAGTTTGTCAAGCTTAGGTTTACACCCTTTGTTGAAGCATATAAA
TTTGAGTATGTTACAATGAGTTCCATGACATCGTCATTTAGAAAAAGTTCAAAATATTCTGTAGGAGTTCTTACTTTGGT
AATTGGTAGCTGAGTAATTCTACCTGATATGGGTTTGGCATTTAGATCTGCTTTTTCCCATTTTCGTACAACTTTTGCTA
GCTTTTTTTTCTTTGTTAGAGGTGGTTGTTGCAAAGTAGAAGAAATCGATGGAACCCCATCTGCAGGAGAATCAAATTTA
GGGAAATATGAGGGATCATCTGGGTCAGAGTCAGCATCAAAAGAACCATCCCGATCAAGATAAGCCGGTGCGCGCAGCAA
AGAACCAGGCAGATTATTTATTGTTCCCCCATCTTCCTCACCGGAGTCCTCATCAGAAACAAGCGCTGTAGCATTTTCAG
GTGGTTGTATGATTATTGCACTTGCTTCTATGTTGTCACACGTCTCTAAAAGATCAGTTAACTCATGTAAGGTTATTCCG
GACATCTTGTGTTTTGGAAAATTATTTCAACTAAAAAAATATGAAGAGAAAATATATGAGTTCCTGTGGTAGTTGTTTTG
CCTAGTGTTCCAAAAATGGAACACTGGAAATTTCAAGCCTATAACTAATTCAAAACTAATAGCTTTTCTATTTTCTTTTT
GCAGTAATAAAGAGTAACTAATTGAATAATAACACAATTTTTTTTACAATTTTAAAGTATAATGCAACAGAGTTATAGTG
ACCTATTTATAGATGAATCTACAAAAACCTTGACGAAAATGAAGAGCAAAATTTCAAAACATCACAGCAGCCAAATATCT
GCAGGAGTGACAAGCTCTGTGTTCTAAGCAAAACCATATGTCCTTACAAACTGGCTGAGGTCACAAATCCAACCATGTGT
GGTGTTCCAAAAATGGAACAGTAGGCATAAATGGG**TTAA**

>Hvu-pble (TIR +SIR or IIR)
TTAACCCATTTACACCGAAGGTTTTTTTTAGTGAATTTTTTCGTGATATTCATATGTGTTGCATAAAATGGTGTAAGAAC
AAAATTGTTTCAAGTAAATTTTTTTTTAATTTTTCTAGTTTTTAAGATACAAGAAGTTCCGAAATCAGAACACTAGGTTTA
TGGACTACATTCTCATGACATTAGAAGAATATATATAAATGTACTCTTTTTTCATAATTATTTGGTCACTTTTAGATACT
AAAAAAATCTTGTTATATTTATTACCATAAATATAATTACAAATATAAAATTACAAAAAAATACAATCACAAAAAAATAT
ATAGTTATTATGAGTATGATATTTGGTTGAACATTTAATATGCAAGGCAGCACTACACTTTTCACATTTGAATGTTGTGT
TTTTATGGCATTGTCGACAACGTGTTTGTTTATTTTGATTCACTATATTATGATTTAGACCATCATATCTTGAATCAATG
TTGCACTTCTGTGATGCTCTTCCCTTTCTTCCAGTTTCTGGAGGTTGACCATAGGTTTCCAGATAATTGGATGCTATACG
TCGATGAAACTCCAGAAGATCTAAAGGCTTTTCTTCTATTGTTCTGTGTAGTAGCCAAGCATTTTGTAAAATTACTTCAA
AACAGAATAACAGAGGGCTTGAGTACCATTTTTTTCCACATATGGATGCTTGATATTGCCCTATATTTTCATCAGCTCTG
TCGATTCCTCCCATATATTGGTTGTATAATTTTATCATGCGAGACTGTTGAATCTGAATTCTCTTTTTTTCTTTTTGTGA
ATAACGATTCACTAGGCCAATAGAATTGATTCCAGCTGCTGATAAAGCAACAGTGACAACACTGTTGTCATGCCATTTGC
AAATAATATTGCCTTTATCATCACTTCTGTAGTCAAACGCACCTCTTTCTTTATTCTTCATAACTTCATCTGAATCAAGT
GGAGGCTTTTCAATTCGATCTTTTCTCACAGTTCCTGTCACATGATGGCCCATTGATGTGAGCTTATCAAGTAAAGAAAT
GCTGGTAAAAAAATTATCAAATATAAAATGATATTCTCCAGGATGCTCCTCTGTTAGTGCCTCAACAAATTGAAGTACCA
CCGATGCACCGACACCATATTCTTGAAATTGTGTATTTGGATTCTTACCCTGATATGGTTGGAACCAACAGATGCATCCT
AGTGGTGTTGCACCACACCAAAATTTGAAACCAAATCGAATAGGTTTGCCTCGAATAAATTGTTTACACCCATGACGACT

AAAGTAGGGAACCATAGACTCATCAAAACTGAAGTTTGTTTCCATTGGAGTAAATTTCAAACATCTATTATTTAGTTCAG
AAATTAGAGGTCGTATTTTTGCAAATTTGTCCTCATGGTCTAAGTTTGCATTATTAGAAACATGCAAGTTTGTAAAAATT
GATGAAAATTGATCACGCCTCATGGCACTACTGACCATTTTATTATGGGTATCAGGTCTTTGCTCCCAAAACATTCTTCG
TCTTAGTAGTGGGACATAGCCACTCAAAATAATTATGCCTAAAAAAAAATTTGAACTCTGAATTACTCAAGCCAAGAGTTA
ATCCATTTTCAGAAGCGTACAAATTTGTATATGTCACAATTATTTCCACAACTTTGTCATCTAAAAAAAGTTCAAAAATT
TCTGTGGGAGTTTTTGCTGTTGTCTGTTGAGGTGGTGTTATTCTTCCTTTTACGGGTCGGATGTTCAAGTATTTGTTTTT
CCATTTGCATGTGACTTTAGAGAGTTTTATTTTTTTCACTTTAGGAAGTTCATGCACAGTAGAAATAGAAACATCTTCTT
CGTCTAAGTTGGAAATAAAATTACTATCATTGACTAAATGAGCAGGTGCTTGGAGTAAAGAACCTGGTAAATTATTTATA
GTTCCACCCTCTTCATCTCCAGAATCTTCATCTGAATTGGTCGCATTTGCGTTTACAGGTGGCTCAATGATTATAGCACT
TGTTTCAATATTATCTTCTTCTTCTAGAAGTTCCATGATTTCATTAAACGATAAGGTCTTGGGCATCTTGGGGAAGATCT
GTAAGAAATAAATAAAAACAAAATAAATAAAAAAAGAGTTTTTTTAAAAATTCAACTTATTGAAAAAAATTTTTGGTTCT
ATGAAATTCCATATTTAAAATTACAATTAGAATAAATATAAATTTATTTTATTTTTTAAATTATAATTGTTGTTTTTGCA
ACTAAACAAAGTAAATTGTGTACTTTGTAAGGCTGATAGGGGGGGCAAAGAATACTGAGGGGGGGAGGGGGCATGGCCCT
TGGCTCCCCCCTGGCTATGTGTCTGTTTCATATCATTCCAATAAGAAACAAAAAAGTTATTATAAAATTAGTTTTTCAAT
TATGCATATAATGTGTGAATGATTATGTATATATATATTTTACTGACAGAGCTATGGCGGGTGGGGGTAGTTTCCTCATA
CATACAACTACATACATACATACATAAAAAAATAAATAAATAAATAAAAATATACATAAATGCTTCATTCACCCTGTATG
TGCCTAGTATTCCGAAATCGGAACACCTGATTTTAGGAACCCATCATTATGACAAAAAAAAAATTCTGTGTACTAATTCT
AAAACACGCTAGTGTATATACAGAAGATAAGTAATGACCAAAAAATATGTTTATAGAACCAAAAAAGTAAAAGTAATACT
TGCATGAGTTGAAAGAAACGTTGCAAAAACTGTCAAATTTTATACTTCAATGTTTTTAACTGTACAGCAACAGAATTTCA
CGTCTATGCACATTCAAATACTGACCACAAGAAAGCAACAGCATTGCTCTCACTGACAAAGTTATCAGAATGCACCCTGG
TGTCAAAATTAAGAAATAAAAAAATAAAATAAAAAATTGAGTTGAGTGTTCTGAAATTGGGTCACTAGGTCTAAATGGG**T**
TAA

>MER85 (TIR + STIR)
TTAACCCATTTATGCCTGAGGTTGCAATTTTTTGAATTTTTGCATGAGTGAAAAATCAGACCTTGGCGATGACCTTGAGC
AGTAGGATATAAATAACTCCCACATGCTTAGCGTTCCAATAATGGAACACTAGGCATAAATGGG**TTAA**

Clade PGBD4
>Oab-pble1 (TIR + STIR)
TTAACCAGTTAACCGTGTTTGACGACTATATCCGTCATGGAGATTTGGCAGAATTTTGTGCCATGACGACTTTATCCGTC
ATGTGTAAAATTTTGTTACAATTCGATATTTAAATTGTAGTTCTGGAGAAAATTAAATTTTATTCCTAAATTTTAATATG
TTTAAAATGTTCAGAGGTCACAACGAAATGAAACAAGCAAATAAAAATTATAAATAATTTAAGTTGATTCATAACGGTCT
GAGAGCTAACTCGTGAACAGTGCAGTTATATCTGGATATTTGATTATCCCGATTTTCATCTTGATGTACCAAAACCAAAA
AAAAGTTTATTAGATTCTTAATTTAGATTCTGTATTCTTTTTTATACTAGATTCTCTATTTGCGGTGATACCTGGAAACA
GCTCATGTGCGCTATCTCACTCCTTGTTTGTTTTGGATCAGGCAATTTTTGTTGCTATTCAGCGCAGTTCTAGGACTTAT
CGGGGATGGAGATGAAACGGTTGATTTTTATCAGTTGTCCGAGAATTCGTAAACGGGGGCCACCCGGATATATTGTTAGC
CTCACATCTTGTGCAAACTTTAGCGCAGGCCGCGATTAGATATCCCCACTTAGGTTTATTCACTCTTCCGCGGTTTACTA
GACTCTAACCCACGTTTGTCATTTCAATTTTTGAACGTGAACGATGCAGTGCATTTCAGTGTGTCTGAAATAGTCAGTGA
ATCGTGCCAAGTCATATAAAGTGCAATGTAAGTGTTTCAAGTGTCTAATTATTATAGTTATCCAGAAAACTTTCGTAACC
CACTATTTTTTTATTCTTAGCTTATTTTATCCCCAAAATGAGTGACTCCAGTGACGATGACATTGTTTCACCTGGAATGA
AGCGAAAACGAGTGAATATATTTCCATCGTCTTCTGAGGAAAACGATGAAAGCATCGAATCCGACAGTGAAACCGACCAA
TATGACTCTGATGACATTTATTCCGAGGACAATAAAAACAACAAAGAAAATCAATATACAAATGTTAAGTGGAAGAGTAC
AAGTGGCAACAGAAAACCATTTGATTTTATTGCCAACAATGGTCAGCAGGAAATTGTACCAGGAAGATTCCGATCCAAAT
GTTCATCTTACGTAGAAAAATATTTAGATGATCATCTTATTAGCATTATAGTAAAAGAAACAAATTTATATGCGGATCAA
TTTTTGCAAAGTCACCCGAACTTGAAGCCGCGATCAAGAATGAGAAAATGGTATTCAACAACTAATAATGAAGTTAGATG
TTTTATCGCGATTTTGATTTTGCAAGGTATTGTAAAGAAGCCAGCACTTGATATGTATTTTTTGAAAAGGGAAATCATTT
GTTCACCGTTCTTTGGCAAAATTTTCTCTGCGGATCGATTTTTACTATTATGCAAATTTTTATATTTTGAAAATAACGCA
CCCCACAATGATATGCCATCGAAAAAGCTTTGTAAAATTAAAACTGTTTTGGAATATGTGATAAATAAGTGTAAATCTTT
GTATACTCCAAAAATGGATATCTGTATTGATGAATCGTTGCTAATGTGGAAAGGGCGGTTATCATGGAGGCAGTATATAC
CTTCAAAAAGAAGTCGGTTTGGGATACAATTTTTCGTCCTATGCGAGAGTGAATCGGGTTATATTTGGAATTTTTTTATA
TATACGGGAAAAGAAACCTATTGTGATTCACAATATTCAGAATTTAATATTTCAGCTCGTATTGTCCTTCAATTATGTGA
TGAACTTTTCGAACGTGGATATAGACTTTATTTAGATAATTGGTACACTGGTGTTCCATTCATAGAGAAACTGTGTGCAC
ATAAAACGGACGTTGTTGGCACAATTCGGAAAAACAGAATTGGCATTTGTGAAGAAGTTCGAGAAACGCGTATCAAAAAA
GGCGAATATATTGCAAGATTCAAAAATAAAATTATGTTGTTGAAATGGCGTGATAAGAGGGAAGTTTATCTTGTAAGCAC
CGTCCATAACGACAACATCGTAGAGATGCAAAAAAGAAATGTAATAAAAAAAGTCCCCGAAGTAGTTTTTGATTATAATA
ATAAAATGGGTGGAGTAGATATGAGTGACAGCATAATTATTGCATACTCAACTGCAAGAAAAAGGTTAAAGAAATATTAT
AAAAAAATATTTCTCCACCTTTTAGATGTCATATGTTTAAATTCATATCTAATATACAAAATGAATGGTGGAAAATTAAG
CCGAATACACTTTTTATTAGAATATATTGAGGATACAATTGCATCGTATCCCATCGAATCGCGAAACTTACCAAAATCTC
GATCAACGGCACCTAATGTTTCGCGCTTGATCGAAAAACATTTCCCGGATTACATACCAGGCAGTTCAAGAACAACAAAT
CCTTCGCGAAGATGTGCTGTTTGTTATAAAAAAAAGATTAGAAAAGAATCGCGGTATTGGTGTCCGACGTGCCAAGTGCC
ACTATGCGTTGTTCCTTGTTTTCGTGAATATCACACGACAGAAACAATTTAAAATTTTGTTCAACTTTTTAAAGTTTTCC
ATTCATTATGTATAAATATATGTTCACATTAAGCGACTGTTACAAGTATTACGAATAAAATTGTTCATTTTCTAAAATCAA
GTCGTCATTTTGACCCAATATTTCAACAGTGCCACTTACTCTGTGTTTGACGACTATACCCGTCATCGCTTGATTATCGC
GACACACACTGGTGCGCGCTTTGCCCACAGGTCGCCACGGTTAACTGG**TTAA**

>Pva-pble (TIR + IIR)

TTAACCCATTTCCTGATTGCCCTGCGAATACTAGCTGGCGGTTGCAACATTTACCTCAAGGTAAACTTGTGGATTACAAT
CCTAACCTTACCCCCAAAGTTTGTTATGAATATCTCGCTCTAGTATATGGATAGTCTTTGGAAACAAATGACCTCATTCA
CTTTATAGCATTCTGTTTTTAGTAGTGGTATTTCTATTTATAAAATATAGTAATTCTCAAACGCTGATAATGTCAAATCG
TAGAAAACGCAGCATCCCTACACGTAATGTTAATAACATCGTTCTCGAACAGTTGTTAGCCAAAGATTCATTTGATGAAT
CTGATTTTTCCGAAATAGACGATTCTGATGATTTTTCAGATAATGTTTCAGAAGACTGTACAGTCAGGCCTTCGTCCGAT
TTGGAATCTGATGGAAATATCCCTACATCAGCTAACTCATGTCACGCTCTGAAATGGTCAACTTGTGTTATGATTCCACG
TCAGAGGTATGACTTTACTGGCATACCTGGCAGAAAAGTTGATGTCAGTGATACTACTGACCCACTTCAGTATTTTGAAC
TGTTCTTTACTGAGGAATTAGTTTCAAAAATTACCAGCGAAACTAATGCCCAAGCTGCCTTGTTGGCTTCAAAGCCACCT
GGTCCAAGAGGATTTCCACGAATGGATAAATGGAAAGACACTGACAATGATGAACTAAAAGTCTTTTTTGCAGTAATGTT
ACTACAACGTATTGTACAGAAACCCGAGCTGGAGATGTTTTGGTCGACAAGGCCTCTTTTGGATACACCTTATCTTAGGC
AAATTATGACTGGTGAAAGATTTTTGCTTTTACTTCGGTGCCTGCATTTTGTCAACAATCATTCTATATCCGCTGGTCAA
TCAAAGGCCCAGGTTTCAATGCAGAAGATTAAACCTGTGTTTGACTTTCTTGTAAATAAGTTTTCAACTGTTTATACTCC
AAACAGAAACATTGCAGTCAATGAATCACTGATGCTATTCAAGAGGCAGTTAGCTCTGAGGCAGTACATCCCAATGAAAC
GGGCACAGTGTGGTCTCAAGCTCTACGTACTTTGTGAAAGTCAATCTGGTTACGTGTGGAATGCACTTGTTCACACAGGG
CCCAGTATGAATTTGAAAAATTCAGCTGATGGTCTGAAATCGTCACGCATTGTTCTTACCTTGGTCAGTGACCTTCTTGG
CCAAGGATATTGTGTCTTCATCGGTAACTTTTATACATCTCCGATGCTTTTCAGAGAATTACATCAAAACAGGACTGATG
CAGTTGGGACAGCTTGTTTGAACACAAAACAGATGCCAATTGATTTGAAAAAAAGGATTGCAAAGGGGACAACTGCAGCC
AGATTCTATGATGAACTTATGGCACTGAAATGGTATGACAAAAAGGAGGTGATAATGTTGTCAACATTCCACAGTGATAC
TGTGATTGAAGTAGACTGCAGAAATGGCAAGAAAACTAAGAAGCCTTGTGTCATTGTGGATTATAACAAGAATATGGGAG
CAGTAGACTTGGCTGATCAGATGCTTGCTTCTTATCCTACTGAGTGCAAAAGGCACAAGTTTTGGTATAAAAAATTCTTT
CGCCACCTTCTAAACATTACAGTGCTCAACTCCTACATCCTGTTCAAGAAGGACAATCCCGAACACACAATCAGTCATGT
AAACTTCAGACTGACGCTGATTGAAAGAATGCTGGAAAAGCATCACAAGCCAGGGCAGCAACATCTTCGAGGTCGCCCAT
GCTCTGATGATGTCACACCACTTCGCCTGTCTGGAAGACATTTCCCCAAGAGCATACCACCAACATCAGGGAAACAGAAT
CCAACTGGTCGCTGCAAAGTTTGCTGCTCGCACAACAAGGATGGCAAGAAGATCCGGAGAGAAACGCGGTATTTTTGTGA
GAAATGTGATGTTCCGCTTTGTGTTGTTCCTTGTTTTGAAATTTACCACACGAAAAAATATTATTAAATACTGATCATCA
TATACATTTTTTACATTAGGATTAGAGACAAGTTCTGTGCAGAAACAACTCCAAGAACAGTTATATTTTATTTTCACATT
GAAAACCACTTAGATTTTCTTTGGCCTCAAAGAGTTTATGTAAAATTAATGCTGGCAGTGAGCTGCACTTTTTTTCTAAA
CAGGAAATGGG**TTAA**

Clade PGBD5

>Tcr-pble (TIR + STIR)

TTAACCCTTTGAAGCATAGTGGTTTCACTTGGTAACCACCTTTTATTTAATTATTTGAGATACATTTAACTGTATGGCAG
CACTTATGAATCATCTATCTGAATTTATTTGCTAGTTTTATTCTTGTAAAAATTTATTATTAGTAATTTTTGTTGGTATT
CATGCCTCCAATATAGTATGTATTTTGTTTTTGTTATTGAAGTGACACGTTTCTTCTAATTGGTTTTGTTATATTTTTCA
AGGAATAAATATATAACTATGTAGGCCCTAAGTATTCAAATCATTATTTAATACTATTGTCACGTGGGATCCGAGCTCGA
ATTCATGTTAACCCACTGCAGCACAGAAGTTACAGCCACACCGAAGTAAATGGTCTTGTAAGACATGCGAAGTGGCATTG
TGTCTCAATGAATCTAAAAATTGCTTTGTAGCATACCACACCATGTAAAAATATAGTGAAGCTTAGTTGCATAGTGGTTT
CACTGAATGACCACCATTTTTTAGCAACACTCAATATTAATGAAAACATTTTATTTTTGCTCAAATTAAATTTTAAATGT
ATTTAGAAATAAGTTTTCATTGTAAAAAATTAAAAATAAATTAAATCCTTAATGTGTGCTTCAAAGGG**TTAA**

Clade PGBD6

>Pxu1-pble (TIR + 2 IIR)

TTAACCCTTTGACTGCGGAATTTTTTTGACCCATACGTTCCGTAGAAACGGAATAAAATTATGTAAATCATCGCCAGTGA
CGGAATTTTTCTAGAGCAGTATGGTTTAGATTGACAATAAATTATTATTAATAATACAAAATGCAATAAATTTACCGTTT
TATTTTCTTTTTATAATTAACAAAAGAATTTTTATTTACGAAACAAAAAATCTTGCTTCGGAAGATGTAATAATCTTTAA
ACAGTAGATATTCACTTGAATGGCATAAGAATTGCTTGAAAATTTTAAGTTTTACATTATAACTTCAAATAATTATGTTC
TCTAGTGTTATATTAAAACAGGTGAAATGTTAATAATTTATTTGGGTATGATGAATCTCAAAACATGGGTCAACACAAAG
TCCTACATCACATATATTGCATTGGTATGATGTATCTTTTCTTTTTTCCTTTCTTTTACAAACAACACAAATCCTCCTAG
CTACTTTTCCTTCTTGATTTAAAATTTTAGAAGGAAAATGTCTTTCTGTTAAACGACTTGGTAAATTTTCAGCAAGTCTA
GGTCTTCCGCGAGCTTGTCTTTTATTTTCCTGCGGATATTTTTGAAGGATTTGTTTAATTAATATTAAATGAAATTTGCT
ATACTTTAACTTTTTGTTGGTTACTTGTTGATAAAGGATATAGGAGTTGTAAATAGACATATCAAGTAAGTGAAAAAAGT
ATTTTTTATACCATTTCAATGATTTTCTGCATGAATTTATGGCACTTAAAACCATATCGACCATATCCACAGAACCCATG
TTTAGATTGTAATCGACTACGCATGCTGGTTTTGACTTATTAAGGCCTGTAGCATAATTTGTTTTAGTAGTTACGATCAT
TTTTGGTTCATGGGCACTTGTCAACATGTAAACATCCTTCTTGTCTCTCCATCTTAAGGCCAAGAGATTATCAGTGTTTT
TGTAATCGAACAATCCCTTTTCTAACTTTTCGTCCATAATCGGCATATCTTTTCTGTTTTTTCTAACCGTTCCAAACGCG
TTAGTTTTCATTTTATGTAAAAGATCGAATAAATTTGGACTCGAATACCAATTGTCCGTTATTAAAGTGTGACCTTTATC
AAGATAAGGTTTTATAAGAGACATTACTACATCTCCAGACTTTCCTACATTAGAAAAGGTTGAATCACACTCTGTCTTGC
TTCCAGTATAAATTATATAATTTAAAACATAATTTGTTTTACAATCACATAAAACGAACGATTTTATCCCAAACCTGCTT
CTTTTTGAAGGAATGTATTGTTTGAAATAAACTCTGCCTTTAAACAACATTAAACTTTCATCTATACATAATTTCTTACC
TGGATAAAAAGACTCTGCAAATGATTTTTTTAAATCATCAACTATAGGACGAATTTTAATTAGTGGATCCTCACAATGTA
TATTATTATCATTAAAATGTAGCATTTGCATTAGAAGCAAATATCTGTCACGGACCATGGTTTTTCTAAAAGTATTTGTT
TGCAAAAGCTCATCAGTAGACCAATATTCATTAATGGCCAATTTTTTCACGCGGGGCATTAACATCGTAAGTGCAAAAAA

AGCATACATCTCACTTATATCGGTATCCTTCCAACTGTTAATACGAGACTTAATTTTATTTGACGAGTTTGCTATAACAT
ACTTGTAATAATTGTTTGTCTCTCGAACGATATTTGAAACTAAATTTTCTGAAAAAAACAACTGAAATGCATCTAATGGA
GTAGAGTTAGGAGTTAAATTGGCCTTGATGCCGGAAAATTCTTGGTTAAATTCATGTAAAATTGGTTTAAAATTATTTTT
TGTCCAATTAAATGTCGAGGATACTTCGGATTCTAGAGGGTCAAATAAATTAGAGTTTGTATATGTTTCTTCTGTTGATA
TGCTGTTAAATTTTTGGCGTTTATATATGTACTCATTTTCTTGGGTTTCATCATAAAAAAGATTTTCTTTAACCTCAGAA
TCTGAAAAAGAAAACACATCTTTCATCGACATCGTAACATCATCTTCATGACTATCGCTCATTCCTAACAATTCATCCAT
CTCGAAATCACTACTATTTCCAACGTATATTTCTTTAAAATTATTACTCATCTTTACTGTTAATGAAAGGGGGAAAGGAG
AAAAAATTTTAAACGATAAAAAAAAAATAAAGAACAACATAACATGTTGGACAATTAGCCCTTTAAGTAAAAAATGGCTTA
CAAAACATAATGCTTAAATATTAAATGCAATAAAAAAATTTTTTTTGATAAAATATAAATTTTTATTAAAATTAGTTTTA
AAACAATAAAAATGCTTCTTTTGCATTAAAAATCATATTTATGTTTCATTTTACAGCTTTTTATCGATGTTTCTGCGCGT
TTAGAGGATTTATTTTCAAAACGCATATATGCGTTTGCCGTGTCTGGGGACAGGTTGTGTAAAACGCAAATATGCGTTTG
CCGTGTCTGGGGACAGGTTTTGTAAAACGCATATATGCGTTTGCCGCAGTCAAAGGG**TTAA**

>Uribo1 (TIR + 2 IIR)
TTAACCCTTTTAGTGCCATAGGACGTAGAATCTACGTCCTATGGCACTATCGGGTTTACAAGCGCTGCGCTCGCTTTTAA
AGCAGCGCAGCGCTTGTAAAGCCTGCACAACCCCCTAGGCAACGAGCAAGGAAGGACATACTTACGGATCCGGTCCCCCA
GCCGCACCGATCCGGTCCCTCAGCCAATGACAGCAGTGGACACGCATTGATGACGTGTCCACTGCTGTCCCTTTAAATAT
CGCCGCCCCCGCCGTCGCCTCTTCACTCCTGCTGCGCACATGTGACTGCTGGAGCCCCCCTGCTGCCTTCCTGCTGCCTT
CCTGCTGGATTGGTCGCCCTGATTGCCTGCCTGCATTGTGTGAGCTGAACTACAACTCTCTAACCACTGTTTATTTTGCT
TATTTCTATCTCAACTGTCTAAAAATTTTTTTTTTTTTTTCCATTTCTTCTTCTATCTTTGTCATTATTACACTTTTGC
ACACATACACACTTATTTACAACTTTCCCAAGCACACACATACACTTACAYTTTTTTTTASAATTTCTTTCTTTTCTTTT
CTTTCTTTCTTTGCTTTCTTTGGCAGTCTTTTTTTTTTTACTAAAACTTTATTTCTGATCTTGCTCTATAATTATCTGATT
TATTTGGTTGCATTTTAGTGTTTTTTTGGCATTTTCATAATTGATTTCTGTTTTTTAATTATTCTATTGCACTGTAACTT
TTTTATTGCTATTGGGTGCTGAATTTGCAGTGACGTTGGTGGATCCAGGGCTCTGGCCACTGATTTCATTGCAATAATTG
TTCTTCTGTTGGTTTAGGTGGTTTTATTGGATTTTACTGATTTTATGGTGTTTTATCTTTGCATTGTTCTTTATTGTGTA
TTTAGTGCCCCAAAAAAGGTTGCTTTTGCAGTGACATTGGTGGATCCAGGGCTCTGACCACCGATTTCATTGCAATTATT
GTTCTTCTGTTGGTTTAGGTGGTTTTATTGGATTTTACTGATTTTATGGTGTTTTATCTTTGCATTGTTCTTTATTGTGT
GTTTAGTGCCCCAAAAAAGGTTGCTTTTGCAGTGACATTGGTGGATCCAGGGCTCTGACCACCGATTTCATTGCAATTAT
TGTTCTTCTGTTGGTTTGGGTGGTTTTATTGGAATTTACTGATTTTATGGTGTTTTATCTTTGCATTGTTCTTTATTGTG
TATTTAGTGCCCCAAAAAGGTTGCTTTTGCAGTGACATTGGTGGATCCAGGGCTCTGACCACCGATTTCATTGCAATTAT
TGTTCTTTGATTGCTTTTAGTGGTTTTATTCCATTTGATTTCAGTTGTTCTAGAAAGGTCCAGAGGTGCTAAGTTCAGAT
TGTTCTGGTAGTTTCTATTGCCAAAGGTTAACTCCTGGTCATCTCTTGCTGGCTTGAGTTTATCAAAAGGTTACAGTGAG
TTTTTTTTTCTAGCTTCTCCTATTACAAGTGTTCTCCTATTACAAGTGTTGTCTTGCTTGCTTTTTTTTTTTTTTTGCT
TGCTTGTTTTGACTTTGCCTGCTGATTGCTAGATTTGCAGTTGTCGGTATATTTCTGGCACAATGGCCAAACGGTTTTAC
AGCACCGAGGAAGCTGCTAGGTATTGCATGGACTCCAGCTCAGAGGAATTTAGCGATTCTGGGTCTGAGTATGTGCCATC
TGACCACTCTACTGAGGAATCAGAGATAGAGGATAGTGACACGGTTAGCGCAGAGGTTAGTAGTGGTGGCACGCTTACTG
CTGAAGAGGCAGAGGCTGGTGGTAGTGCCCAAGATGTACCCATGGATGTAGAAGAGGGTGAGGGTAGCACTGATGCAGCA
GTTGGAGAGCCTGCGTGGGGGCCTCCCTGTAATTATGTCCCTGAAATTCCCCCTTTCACTGCAGTCCCTGGCCTCAGTGT
GGACACCACTAATTTTGAAATTATAGATTTTTTTAATCTCTATATTACAGAGGCCATCCTACAGGACATGGTCCGTTACA
CAAATTTGTATGCTGAGCAGTATCTTGCCAGCAACCCTTTACCGGGTTTTTCAAGAGCCCAGGCATGGTATCCCACAAAT
ATAAACGAAATAAAAAGATTCCTGGCTCTTACATTGGCAATGGGACTCGTAGAACGTAATTCTTTAGCTTCCTACTGGGA
TACCAATACAGTCCTTTCCATCCCTCTTTTTTCAGCCGTTATGCCAAGAAACCGTTATCAAATTTTATTGCGGTTTCTTC
ACTTTAACGACAACACAACGGCTGTTGCCCCTAATGAGCCCGGTTATGACAGGCTTTATAAATTGAGGCCCCTTATAGAT
AGCCTGTCTCAGCGCTTCGCAGAGGTTTACACCCCCTCCCAAAATGTCTGTGTAGATGAGTCCCTTTTGCTCTTCAAAGG
GCGCCTAAAATTTCGCCAATATATCCCTAGTAAGCGTTCTCGCTATGGGATGAAATTTTACAAACTCTGTGAAAGCAGCA
CGGGCTACACTAGTTATTTCATGCTGTATGAGGGAAAGGACACTAATTTGGACCCCCCCGGTTGTCCCACTGACTTGACT
ACCAGTGGTAAAATTGTTTGGGAGCTCATAACTCCACTGCTGGGCCGAGGTTACCACTTATACGTGGATAACTTTTACAC
AAGCATCCCTTTATTTAGAGCCCTAAATTCTTTGGACACCCCAGCCTGCGGCACAGTCAACCGTAACCGCAAAGGACTGC
CCAGGGAATTGCTGGATAAGAAACTGAAGCGTGGAGAAGTACATGCTCTAAGAAGCAATGAGCTCTTGGCCATAAAATAT
TCAGACACCAAAGTTGTTCTTATGCTTACTACTATCCACGATGAGACCATGATTGTCCAACACAGAAGTGGCGGTAGGCC
CTCAAAAACAAAGCCCCTGTGTAGTAAAGAATATAGTAAGCATATGGGTGGTGTTGATAAATCCGACCAAATACAACATT
ATTATAATGCCACCCGAAAAACCAGGGCCTGGTACAAAAAAGCAGCAATTTATATGATCCAAATGGCCCTTTATAATTCT
TATGTCGTTTACAAGGCGGCAGTACCAGGCCCCAGATTGTCTTATTACAATTATTTGCTGCAGTTGCTCCCTGCCTTGTT
GTTTGGTGATGTAGAGGAGGTTCCTGACATGCCAGGTAACGACAATGTTGCCCGAATGGTGGGTAAGCATTTTATAGCAC
AAATTCCACCAACACCTAATAAAAGATATGCCCAGAGAAAATGTAAGGTTTGCCGTTCCAAGGGTGTTAGGCGAGATGTC
CGGTATTATTGTCCCAAGTGTCCTAGCAAGCCTGCTTTGTGTTTCCACCCATGTTTTGAAGTGTACCATACTGTGGTACA
CTATGAGAGGACTTGAATTTTGTGTTTTTATAGGTAGGTTATGGTTCTCTGGGTTTGTTGGTGGAATAAGGATTTAGCAT
ATCAGAATAGTGGACTTGGCATGTTCTAGGTTTTTATGCACAAGGTTGCAGGCCTCCAAACTTAGCTGTGCTGGCAAAGT
GACGCTGTTTTTTATTTATTGAGTTTATCTATTCTGGCAGGTCCGTACTTTACTTATAACTAGGTTATAGATATATGTGA
TGAAAAGTTTGGTAAGGTGTGTCATATTATTGGCAAAAGTTTTTTTTTTTTAGTAAAGTAGGGATTTATGGTGTGTGATT
TTGTTTTTTGCACAGGTTGTCAGAAAAACAAAGTGTGCTTGTAGCTTGTGGTTGTTTAGACTTGTTCTTGGGCATTCAGA
TTTTTGTGTTTGAGCTATGGTTATTCAGGCTAGTTGCAGTGTTTCTTTGGTTATGGGGTAGGACTTATATTGTGCTGTTG
GTATAGCGAACTGAGGTGCTGGGTGACTGACAGGGCGCATGATAAAACTTGTGGTGTGTGATGAATCAATGAGATGTTTT
TATCTCTAAAGCTGATGCCACAAGTCCTTTTACTGGCACCTCAAAAGCTTGCCTGTCAGCGCAGTTATCTGTGCATAGTT

TTGTATATTACTTTGTATTTCTGGGCAGCATTTAACTACTGTAAGTGTAGGCCTTAAGTTTTTTGCACGACAAAGGTTCT
TGAGCACAGGATGACAGATACATTACGGTTTTTTGTGCTTGTAGTACAAATTGAATATCTACTGTATGTTCTGGTTAGGT
TACACAAGTTGCAGTGGCTCTCTAAATATAAAAACTTTGTGTAATACTGGTGTGCCACTGGTGACTGCTGCACGGCACGT
AATGGTGCACATAATTTTATCTCTGATGCTAATGCCCCAAGCTTTATTGTTGGCACCTAAATTCCTAACCTTTAGTGCAA
CTTCTAAGGACACTTGCATCATTTATGGGCCAAATTACTATGGATTTTTTTGGCAAATATCCTTCTAAATGTGGTGCAGA
TTTTGTCATTTCTAAAGTAGTGTAGGAAGAAGCAAAAAACTGGTATCATAACACCACAGGAAAGATGGAACTCAAATCAA
AGATGCTCCGCTAAACAGAAATAGGTGAGTAAAGTACAATTTAGGGGTTTATTTGGGGGAGGGGGTAACTGTACAGATGG
GGGGGTTAAAGTAAAAAAACTGCAAGGGGTGCATAGAGTGCAGCCCCAATATTATGCATTTTCAGGGCATTTGGCGGCCA
TGTACTTATGTGCAACCAGACATATGGGGTGTAATTTTATTCAGGGTAACTTGCTAATTGATATTGAGCAAGTTTTTTGA
TAGTTGCCATTGAGATTTTGGGGAGAAATCTAACTTTATATCTGTTTTTTCCTTGATTTCAGACTAAAGCGCAGACTTGT
ATARGGAACTGCAGCCACAATTTTSCMTGTAGAAARMGAAGARTGGTGTCTCTAAATAGCTGAAGTTGTSTATTTTTAAT
CAATRTATAGTTTATATGGGTTTTTTCACAGGTAAGGGGCAATTTTGTCTGAAAAACTTTGAGATGTGCACATAAATTGC
AGCCCCCAAAATTTCAACTGAAATTACCCTTATGCATTGCCCCTGTTTTGGGGCATTTGGTGGCCACGTCTTTATGTGCA
CCCATACATATGGGGTATCGTTTTATTCAGGGGAACTTGCAGATTGATGTTTAGTAAGTTTTTGGTAGTTGCCATGGAGA
TTTTGGGGAGAAATCTAGGTTTGTATCTGGTTTTTCCTTGATTTCTGACCAAAGCGCTGACTTCTGCAAGGGACTGCGGC
CACAATTTTCCATGTAGAAAGAGAAGAGTGGTGTCTCTGAATAGCTGAAGGTGTGCACTTTTCAGAAATATATAGTTTGT
GGGGGTTATTTCACAGGTAGGGGGGGTTAACACTGAAAAACTGCAGGAAGTGCACATAGAGCGCAGCCCCCAAATTTTCA
ACTGAAATTGCCCTTATGCATTGCCCCTGTTTTGGGGCATTTGGTGGCCACGTCTTTATGTGCACCCATACATATGGGGT
ATCRTTTTATTCAGGRGAAGTTTGTCTTTCAAATTTGCCTTGGTTAGAAAATTTTTATGAGATTTTTTTTTGTCAAATCC
ACATTTGATCATGCGTCCAAGTTTACGTTTTAGAAAAAAAAAAATGTCAAAAAAAGCTCCAAATTTCGCAATGCACTGACA
AAAGGTATTTGGCTTTTGAGTGAAAACTACATTGCACCTAGAAACCTGAAGGTCTGTAGTTTCTAAAGATACCAAACATG
AGGGGATATTTTAGATTTACATATAAGTTATGCTGCATCAACTGTTACAAGCGCTTTTCCGCTTTGTTCTGGTGTGATAT
TGTACAAAGTATTGCTTTAGTTTGGGGGTTACTTCTGGACAGGAACTGTGGGGTACCACCACATATTTGGTATCGTTGGA
CTTGGGAGTATCAGGGCTTTTACAAACAACAAAAAAAGTGTGTAAAATTAACTTTTCTATGGAAAAAAAACTCAAAATAT
ACAGAAATTTTTCATAATTTTTTTTTTTTTACATATTTCACCCAAAATACACATCATATCTCCAGAAAAGTTATAAAATT
TGGTATGTATGTCGAAGCCCAATTAGTGATGAAAAAAACGATATATAATTTCCCTAGTTTCATGGAGGTTTTCCTACCAA
AAAACATTGTTAAAGTGAATGAGTACAAAATGCTTAAAAAACGTCTGGCACTGGGGGGAACTGAAATGACGAATTCGGCT
GGCACTTAAAGGG**TTAA**

>uribo2 (TIR + 2 IIR)

TTAACCCTTTGCCTGCCAACACCGTACGGCGTACGTGTTGGCAAGGCAATGCACTGAGTGCCAGCGACGTACCTTGTACG
TCTTCCTGCTTGTGGCGATGTGGATACTGATCGGCGGCTTCTTCCTCTTTCAAAAGCCGCCGATCGCTGCAGGATCCCCT
AGGCAACGAGCGCTGGGGGCTGGAAGTGACGCTGTTGCGTCAGATGACCCCAAGAAGGACCCCAGAGCAGCGGCTACGTC
ATAGACATGTGCCTGCTGCTGGGGTATTTAAACTCAGCCCCCCAGCGCCGCCCTCTGCCTCTCTGCTGGTCCTGGATGTT
GCTGGAGCCCTCCTGCTGTGTGCTGCTGTTCTCCTGCGGTCCTTGGCTGTGAGATTTGCTGTGCCTGCCCTCGGGTAGTC
CCTTTGCTTTTACATTACACACTCACTACTCACTTACTACACACTTACTACACACTTACTACACATTTTTACATTTTAGT
ACTTTACATTACTTTTTTTTTTGCAGATTTACACTAACTACACTTACTAACTTACATTTGCACTCACTACACTTACACACA
TACACTTACACACATACACTTACACTTACACACATCTATAAAAAAAAGCTTTTTGTGTCTTTGTGTCATTGTTTTTGTCCA
AAAAAACCTGTTTATTTGGTGTTGTGCCCTTTTATTCACTGTTTGTTATTTCATTTATTTGTTCACTGTTATTTGTTCAT
TGCCTTTTATTGTGATTTTATTGATTATTCTTTTCTGCGTTGCTCTTTTGTTATATACTTTTGCTTGTTACCATCATTTC
TGATTATTAGTGATTTTATTGTTTGAGTTTTGCTTTGCTCTTATTACTTGTTTTTGGTTGTACTTTGCATTTTCAGTTTT
ATATTACTGTTCATTGCTGGTTGTTTTGATTTTCAGTTATTTTATTGAATCTTGAGGTTTGCATTGTTTTTTATAACTGG
TTCTAACTTGTTTTTGCTTGTGCTTTGAATTTATAGTTTTTTATTCCTAATCATTACTGGTTCTCTTGTTATTGATTTGT
GATATTTTGACCACTAGTTACTACTAATTGCATTTTATTTGATTGTAAGTTCTTTGGCTAGTCTGATAGGGTTTTTGTTC
CAGAAAGTTCCAGAGAGTTCAGGGAGTTGAAGAGTTCAGGGGGTTGAGCTTATAGCATTTTTTTTTTCTTGGTCTGATTT
GTTTATATTATCAGTTGGGTTTTTTTCTCCAGTCTCTGCCTTGCCTGTTGCAGTAGTTTTACAAGTATCCCATTTTATTG
CTTTATTGCTTTTATTKCTKTATTATATATTTTTGCTTGTTACCACCATCTCATTGTTGTTTGAGTTTTGCTTTGACTTC
TGACTAGTTTTACTTCTGACTAGTTCTTGCTTGTACTTCTAGTTTTATATTGCTTTTCATTACTAGTTCAATTGATTTTC
AGTGATTTTATTTTGAATTTTGAGTTTTGCTTGGTTCTAATTACTGTTCATTGCTTGATTTTGCTCTAATTACTAGTTGT
AACTTGTTCTTCTTTGTACTTTGGATTTTCAGTTTTATATTGCTGTTCATTGCTCTTGATTTTCAGTGATTTTATTTTAT
CTTTATTTTGCATTGTTGTTATAACTGCCTCTAACTTGACCTCTAGTTCTTTGGCTAGTCTGATAGGGTTTTTTGTTCCA
GAAAGTTCCAGGGAGTTGGGAGGGAGTTCAAGGGGGTTGAGCTTATTAGCAGTTTTTTTTTTCTTGGTCTGGTTTGTTTA
TATTATCAGTTGGGTTTTTTTTCTCCAGTCTCTGCCTTGCCTGTTGCAGTAGTTTTACAAGCAATCTTTTTTTCTGGCTT
GATTGTTAGTTCTTTGTCACTGAGTTTTAGATTAGCCTGTACATATTTTGTTTGCAACATGGCAAAAAGGTTTTATTCTG
CTGAAGAAGCTGCTGCACATTGCATGGCCTCTAGCTCAGAGGAGTTTAGTGGCTCCGATTCTGAGTATGTGCCCCCTGCA
TCAGAGAGTGACTCTTCTACTGAGGAGTCATGGTGTAGCAGCAGCACTGTTAGTGCTCTAGAGGAGCCTATGGAGGTAGA
TGAGGATGTGGATGATTTGGAAGATCAGGAAGCGGGCGATAGGGCTGATGCTGCAGCTGGTGGAGAGCCTGCGTGGGGGC
CTCCTTGTAATTTTCCCCCTGAGATCCCCCCATTTACTACAGTCCCTGGCGTCAAGGTGGACACCAGTAATTTTGAGCCC
ATAAATTTTTTCCAATTATTTATGACTGAGGCCATCCTACAGGATATGGTCCTCTATACCAATGTATACGCCGAGCAGTA
TCTTACACAAAATCCCCTACCTAGGTATGCTCGAGCTCATGCATGGCATCCCACAGATATTGCCGAAATGAAGAGGTTTG
TGGGACTTACTTTAGCTATGGGTCTCATAAAAGCCAATTCTTTGGAGTCCTACTGGGATACCACTACTGTGTTGTCTATC
CCTGTCTTTTCTGCCACTATGTCTAGAAACCGGTATCAGCTACTGCTCCGGTTTCTGCATTTCAATAATAATGCTACAGC
TGTACCCCCTGATCAGCCCGGCCATGACAGGCTGCATAAATTGAGGCCCCTAATAGACAGCCTGTCTGAGCGCTTCGCAG
CAGTATACACTCCCTGCCAAAATATTTGTATTGATGAGTCCCTTCTCCTCTTCAAAGGGCGCCTACAATTCCGTCAATAC

ATCCCAAGCAAGCGTGCCCGCTATGGGATTAAATTTTATAAACTTTGCGAAAGCAGCTCGGGGTACACTAGCTATTTCCT
GATTTATGAAGGCAAGGACTCCAAATTAGATCCCCCTGGTTGTCCCCCTGACTTAACTGTCAGTGGAAAGATTGTATGGG
AGCTCATTTCTCCACTTCTGGGCCAAGGTTTCCACTTATATGTGGATAACTTTTATTCTAGTATCCCCTTGTTCACTGCC
CTATATTGTTTAGATACCCCAGCCTGTGGCACCATAAATCGTAACCGAAAAGGATTGCCCAGGGCTCTGCTTGATAAAAA
ATTGAATCGTGGAGAGACTTATGCCCTTAGAAAAAATGAGCTCCTTGCCATCAAATTTTTTGACAAAAAAAATGTTTTCA
TGCTGACATCAATCCACGATGAGTCAGTAATTAGAGAACAGAGAGTTGGTAGGCCCCCCAAAAATAAGCCTCTCTGTAGC
AAGGAGTACAGTAAGTACATGGGTGGTGTTGATAGAACTGACCAACTCCAACATTATTACAATGCCACCCGGAAAACAAG
GGCCTGGTACAAGAAAGTTGGTATATATTTGATACAAATGGCCCTCCGAAATTCATATATTGTGTATAAGGCAGCAGTAC
CAGGTCCCAAATTGTCCTATTATAAATACCAGCTGCAGATACTCCCTGCCCTGTTGTTTGGTGGTGTAGAGGAACAGACA
GTGCCTGAGATGCCCCCTAGTGATAATGTGGCCCGATTGATTGGGAAGCATTTTATAGATACGCTTCCACCCACACCTGG
CAAACAGAGACCCCAGAAAGGCTGCAAAGTGTGCCGCAAAAGGGGTATAAGACGAGATACAAGGTACTATTGCCCAAAGT
GCCCTCGCAACCCTGGGTTATGTTTCAAACCATGTTTTGAAATATACCATACGCAGCTGCACTACTGAAAGGTTTATAAA
TTGTGTATTGATTTGGATTGGATTTATAGTAAGCATTGGTATCCCACTTTGTGCAAAATTTATATAGTATACTTTGATTT
ATGTATTCAGGGCAGTTGTGGTTATTTTGGAATTTTCAGTACCCTAATATATTGCTTATTATGGATGGCACTGGGGGTAT
AGCCAATTGAGGTGTAAGGTAACGCACAGGGCGCCTACACATTTGTGGTGTGTGATGATTATGATGATTCATCTTGCCTG
CTGATGTTTTCATCTCCCCTGCTGATGCCACAAAGTTTTTTTTCTACTGACACCTCATAAGCTGCCCTTGCTATCTGTTT
TTTTGGACAATATTTCATATTTTACTAGTTTATTGGGTCTCCCTGAAATCTTGTTTTTTGTTATGGGCTCCAGCTTTTTT
GTAAAGTGATTTTTGCATGTTTAGGGTTTTTGTTTTGCAGTTGTGAAAGTAGTTTTGCTGATGGTTTGCTGATAGATGGG
GACAGTTTGCATTTCAATACTAATCAACTATTATCTACTAATCAACTAATTCAATGCTAATTGTAGTCTTTCTCTTTATT
ATGGATWTTAGTACTGAGCTGAAGGCATAACAGTGAAAGGAACAGGGAACTTGATTGACAACTTACTGAGGATATCTTTA
ATTTTGGGAAGCATATATTTTCTTCTTAGTCACCTAGCACCACAATTTAGCAAAAAGCTAAATGCTTACATCAGGGGTCG
TCAAACTTTTTAAGTGCTGGGTCCTTCCGCTCTCCGCTGGGTCTTTCCGCTCTCCGCTGCGTCCATCTGCTCTCCGCCTC
GCATCATATTGGGGAACAGGTAAATTATCTTGGGGGGCCGTATCAAGCCCAAGGGCCGTAGTTTGAGGATGTCTGGCTTA
CATGAACCGCAAGCGTGATCAGACTTGACATCACCTCTTTAGCTAATGCAAGCTTTATTGTTAGCATCTCAAATGCTCAC
AATTCAGTGAAGTACACATGAAACGAAAGGGCCGAAAGATGCTCAACTAAATAGATAAACGTAAGTAGAATTTATTTGGG
GCTAAATTGGGCAGGGTGTTTTAGTCATTATTGCAATTTCAAGGAGAAATCTAACTTTGGATCTTTTTTTATCCTTTATT
TTAGGCTAAACTGCAAAGTTCATCGGAGAACTGTGTCCACAATTCTATATGTAGGGAAGCAAGGATGGTTTTCCTGAATA
GCTGCAGGTGAGCACTATCAAGGAACATATAGTTTGTGGGGGTATTTTCCAGGTAAGGGGGGTGTTAAGCAGTACCACCG
CTCCCTATACGCAGAGTACACACTCTGCGTAGGGCACCAACTCCCAGGGGGGCACCCAGACAGTACCGCCGCTCCCTACG
CAGAGTGCGCACTCTGCGCAACAACATACCAAGTCTTCGCCAGCGCCAGTCAGCGCCAGCGGCTCCTTCTCTCTTATGCT
GCGGCAACAGGCCCTTTTATAAGGTTGCGCCCGTGCGTATGACGTCACACATCAGCGACGAGGCGCAACCTTATAGAAGT
GCCTGTTGCCGCGGCATAAGAGAGAAGGAGGCGCCGACGATCACTGGTCTGTTGGGGGTACTTTCTGTGGGGGCACTGTG
CGGGCTGGCTAATGTCTATGGGGGGTACTGTGTATGAGGCAATTGGGGGTACTGTTTTTGGGGGCACTGTTTATGGGGGC
AATTGGGGGTACTTTGTATGGGGGCACTTTGTGTGGGGGCTACTGTCTATGGAGGGTACTGTTTATGGGGGCAATTGGGG
GTACTTTCTATGGGGGCACTGTGTGTGGGGTCTACTGTCTATGGGGGGTACTGTGTATGAGGCAATTGGGGGCACTTTGT
GTGGGTGCTACTGTCTATGGGGGGTACTGTTTATGGGGGCAATTGGTTCTATTTTGGCCCCTATAGTAGGTGTTTTTTTT
TTTTTTTTTTTTTCTCGGTAATATTTGGGGGAGGGGGCACCAAAGTAAATTTCACCCATTTGGCCAGCAGCGGCCCTGGT
GTTAAGACTCAAAATCTAGCAGATATGCATTTATAGCATAGCCACCAGATTTAGCTATTGCCCTAGTTTTAGGGTGTTTG
GTGGGTGTGTCTTTTGATACTCACATATGTGGGGTATCGTTTGATTCAGAAGAAGCTGAAGATTGATATTGAGAAGGTTT
TTTGTAGTTGTCACGGCAATTTTAGGGAGAATTTTGACTTTGGATCTTTTTTTTTCTTCATTTCAGGCCAAACCGCAAAT
TTCTACGAAGAACTGCGTCCACAGTTTTTCATGTAGGGGAACAAGGATGGTACCGCTGAATAGCTGCAGGTGTGCACTTT
CAAGAAATATATGGTTTGTGGGGGCTATTTTACAGGTAGGGGGTGTTTTGACTAAAAAACTGCAAGGAGTGCACTTAGAG
AGTAGCCCCAAACTTTCCAGCTGAAATTGCTCGTATGTATTGCCCCTGTTTTGGGGTGTTTGGTGGCCCCGTCTTTAGGT
GCACACTTGCATGTGGGGTATCGTTTTGTTCGGGAGAATTTGCAGATTGATATTGAGCAGGTTTTTTGTAGTTGTCATGG
CAATTTTGGGGAGAAATTTAACTTTGGATCTTTTTTTTTCTTCATTTCAGGCCAAACTGCAAACTTCTACGAAGAACTGC
GTCCACAGTTTTTCATGTAGGGGAACAAGCATGGCACCGTTGAATAGCTGCAGGTGTGCACTTTCAAGAAATATATGGTT
TGTGGGGGCTATTTCACAGGTAGGGGGTGTTTTGACTAAAAAACTGCAAGGAGTGCACTTAGAGAGTAGCCCCAAACTTT
CCAGCTGAAATTGCTCGTATGTATTGCCCCTGTTTTGGGGTGTTTGGTGGCCCCGTCTTTAGGTGCACACTTGCATGTGG
GGTATCGTTTTGTTCGGGAGAATTTGCAGATTGATATTGAGCAGGTTTTTTGTAGTTGTCATGGCAATTTTGGGGAGAAA
TTTAACTTTGGATCTTTTTTTTTCTTCATTTCAGGCCAAACTGCAAACTTCTACGAAGAACTGCGTCCACAATTTTTCAT
GTAGGGGAACAAGGATGGCACCGCTGAATAGCTGCAGGTGTGCACTTTCAAGAAATATATGGTTTGTGGGGGCTATTTTA
CAGGTAGGGGGTGTTTTGACTAAAAAACTGCAAGGAGTGCACTTAGAGAGTAGCCCCAAACTTTCCAGCTGAAATTGCTC
GTATGTATTGCCCCTGTTTTGGGGTGTTTGGTGGCCCCGTCTTTAGGTGCACACTTGCATGTGGGGTATCGTTTTGTTCG
GGAGAAGTTGCAGATTGATATTGAGCAGGTTTTTTGTAGTTGTCATGGCAATTTTGGGGAGAAATTTAACTTTGGATCTT
TTTTTTTCTTCATTTCAGGCCAAACTGCAAACTTCTACGAAGAACTGCGTCCACAATTTTCCATGTAGGGGAACAAGGAT
GGCACCGCTGAATAGCTGCAGGTGTGCACTTTCAAGAAATATATGGTTTGTGGGGGCTATTTTACAGGTAGGGGGTGTTT
TGACTAAAAAACTGCAAGGAGTGCACTTAGAGAGTAGCCCCAAACTTTCCAGCTGAAATTGCTCGTATGTATTGCCCCTG
TTTTGGGGTGTTTGGTGGCCCCGTCTTTAGGTGCACACTTGCATGTGGGGTATCGTTTTGTTCGGGAGAAGTTGCAGATT
GATATTGAGCAGGTTTTTTGTAGTTGTCATGGCAATTTTGGGGAGAAATTTAACTTTGGATCTTTTTTTTTCTTCATTTC
AGGCCAAACTGCAAACTTCTATGAAGAACTGCGTCCACAGTTTTCCATGTAGGGGAACAAGGATGGCACCGCTGAATAGC
TGCAGGTGTGCACTTTCAAGAAATATATGGTTTGTGGGGGCTATTTCACAGGTAGGGGGTGTTTTGACTAAAAAACTGCA
AGGAGTGCACTTAGAGAGTAGCCCCAAACTTTCCAGCTGAAATTGCTCGTATGTATTGCCCCTGTTTTGGGGTGTTTGGT
GGCCCCGTCTTTAGGTGCACACTTGCATGTGGGGTATCGTTTTACTTGTGAGAACTTGTTCTTTCATATTTTACTTCATT
TGAAAATTTTTATTGGATATTTTTTTCTCAAATTCACTTTTGTCTCTGTGACTTTGATGGCATTTCATAAAATAAAAAAA

ATCCCAAAAAGTATTGTTGAATTTCGTAGTGATCTGTATGACGCCAACTGCATTTAGAGCAAAAAACTACCCCAGACCCA
AAAACCTAGAGGTGTGTAGTTTCTAAAAATACCTAACTTATGGAGGTCTTTCACTTCTATCATTAGATATACCATGTTAA
CATAGAGATGCGCTTATCGGTTTTGTTTCAATGTAAAATTGTAGAAAATGTTGTTTTATTATTGGGGTGTCTTCTGGGCA
AGAAAAGTGGGATACCAATACATATTTGGTATTGGTGGATTCAGCAGAATCGGGGCTTTTATGAATAATAAAATTTTTGT
CAGTAAATGTAATTTTTCTTGAAAAAAAAACACAAAATAAAAACATATTTTTATTTATTTTTTTTTTACATATTT
CACTCAAAATTTTTGTTACATCTCCTGAAAAAGTACATTTTTTTTTTACAGTGTTCAAGTCCAATTCGCTCCGGAAAAAA
CGATATATAATTTGCCTTATTTCATGTAGGCTTTCTTGTCAAAAAACCTAAGCAAATGTAATGAGTGCAAAATGTCTCAA
AATTGCTTGGCAGTAGATGTTCGCTTTTAGGGCAAATTGGCTGGCAGTGAAAGGG**TTAA**

Clade PGBD7

>NviPBLE (TIR including STIR)

TTAACCCTTTCGTGCACATAAGAGTTACAGCTATGGTGCGGGGAAACTGTTTATATTTCATAGGGTCCCGAGCCGCGCTG
AATCCGAATCCGAGATCAGATTTTGAAAATTTAAGATGGCGGATCCAATATGGCGGACGGGAATGTCAAAAAATCGAGAT
TTTGAAAAACAAAAAATTTTTTCGTATTCTACGTGAAAATACAAATAGAAACTGTTCTTGGTGTTTTTTCGATATCTCTC
TTAGTTCTTGAGTAAAAAATTCATAAACTTTGAAAATAACGCAGACTTATTTACCGTTCGTACACATAAGCGTTGCAGCC
ACATGAGTAGGAAACTGTTTATATTTCATAGGTCCCGAGCCGCGCTGAATCCGAATCCGAGGTCAGGTTTTGAAAATTCA
AGATGGCGGGTCCAATATGGCGCATACGGTCTTAGTTGTAGCACATACCTTCGGAAATGCAAGAAAGATCTTCTACACAT
TCATAGTAGACCTTCCATGCACCTTTTCACAAATGCCCTGAGCATTTTTTAGATGTTCACTAATGAATGCCGTGCTGATT
CATTACTTTGAATAAATATTCATTTCAGTTTTTTGAAAAAATTCAATTCAGCGATTAGTAGGCGTCAAAAATTTGCAAGA
ACACTGATTTCATCTTATGATGTAATTTTCATGCTCAGAATAACTTTCTCCCTGTGAATTTATTCCGATCCTTCTAATTT
GATTTTGTTCATAGTATTGGATAATTACGCTTCTCGTCAGCTCGACTGAATAACTATAATACGCATTTTGCTCAGACTAT
TGTATAACATTTTTCTTCGTTTTCTCCAATTCGTCTTACGGTTATTGGATAACAAATTTACAAAGCTTGTTTAACAACGC
ATTTTCAATATTCAGTGTCAAGTGCAATCCTCGGGATTTTCATAACGTTTCAATATTTTTGTAACGAATGACGAGTGTCA
GTGTTGCAATTAAGATTATCAATTGATCAAATTGAGTTCTTATCGAAAAAAAAAACTTAGAATATATTTTCATAAACATG
GCCAAGAGAGCCCGTAGCAAAAAGACAATCTTCCAGTCAATCCCGAAGAGTCGGATTTTAGTTCAGATAAATCCACTGAC
GATGCAGATGAATGTGATTCATCAGAATTGGATGAAAATGATGACAGTGATTCTTTGAATTCAAAATGGACTCAAAATGT
AGGCAATTTTGGTGTAGATTGGGAATTTTTTGATTCACCTTCAACTATCAACATACTTTCAATACCGAAGAAAAATGTCG
AATCATTTTTATTTTCTCAAATTTTCTCGACTTCGCTTTTAGACAGTGTAGTAACTGAGACTAATAAATATGCATTTCAA
AAAAAGAGTAAGAATTGGCATGATACGAATAACGAAGAAGTTTCAGCTCTCTTAGGAATGTGGATTTTGATGGGTATTCA
CAAATTGCCAAGAGTAAGAAATTATTGGAGTTCAGATCCATTTCTCAAAGTTGATTGTGTTGCAAATATTATGCCTTATA
CTCGATTTTCTAAGCTGACGGAAATGATACATTTTAATGATAATGATTCTATGCTCCCAAGAACTGACCCTAACTACGAC
AAGCTCCATAAAATCAGACCGTTGATCGATCATATCAACCAAACAAGTAATAAGAGTTATAACACCTCAAAGACCGTTTC
TGTAGATGAGTCAATGATCCCCTTCTCGGGACGGTCCTCTTTTATTCAGTACATGCCAATGAATCCAATAAAACGAGGGT
TCAAAGTCTGGTGTTTGGCTGACTCATCTACAGGTTACGTGGTCAAATCTGAGGTCTATACTGGAAAAGCAAGAGATGGG
TCGAAAAATCCACTTGGAGAACGTGTAGTTCTGAACTTAGTATCGAATATTCTGGGGGTTGGATATTTGGTAGCTTTCGA
TAACTTTTTCACTTCCGTCAACTTGGTTGAAGAGCTTCTTGATAAAGGCATCTACTCAGTTGGAACAGTTCGAAACAACA
GGAAAGGGTTGCCAGAAATGATGAAAGCAAAGGAGAAAATGAAAAGAGGAGATTTCAATTATAGAATCAAAAAAGGTATT
TCTGCGATCAAATGGATGGACAGGAAACCAGTCACAGTGTTGACGTCAGCTTTCAATCCTACGGAAACTGTCGAGATCAC
TCGGAAACTGAAAGATGAATCTCGGCTTCCTATATTTTGTCCAATTACGATCCAGGCTTATAACAAAATAATGGGTGGTG
TAGATAGGTTTGATCAATTGAGGGAAGTATATATGATAGAAAGAAGATCTAAAAAATGGTGGCATTCAATAATGTATTAT
TATATAGATATGGCTATAGTCAATTGTTATATTCTATACAAATGTTATAAGGAAGGTGAGTGTGAAGATCAATTAGCATT
TAGAATACAGCTAGCGAAACAATTGATTGGTGATTACAACGGCAGAAAAAGATTCAGAGATTCGAAATAATATGAATATA
TCCAAGTTATTTTGAGTATACTCAAATTGACGTCAATCTCTGAATCTTTTTCTAACACTAATTAAAAAAACGAAGGGAAG
AAATCTAAAAATTATCAAAAAAATCAAATAAAATACACAAAAAAATATAAAAATTTCAAAAACTACAAAAAAAAAAGAAC
TCATCAAAGTCCGAACTCCAACCTTCACGTGGTGTTGAGTGTAAGTAAGTGAAATCATCACTGAAACTAATAGAGTTAAG
GCTATGCAGCTGACGATTCAATTGATCCTCAGCCTGACCAGTGATATGAGAATTCCTATCTCCTACCTCATATAATTTTC
GAAAAATCTCACATAGAAATGTATTCCTGTTTTCGTTTTTTCGACACAGAGTACAGTATAATTATTCGTTCCTCAAAATC
TCGATTTTTTGACTTTCCCGTTCGCCATATTGGATCCGCCATCTTGAATTTTCAAAACCTGACCTCGGATTCGGATTCAG
CGCGGCTCGGGACCCTATGAAATATAAACAGTTTCCTACTCATGTGGCTGCAACGCTTATGTGTACGAACGGTAAATAAG
TCTGCGTTATTTTCAAAGTTTATGAATTTTTTACTCAAGAACTAAGAGAGATATCGAAAAAACACCAAGAACAGTTTCTA
TTTGTGTTTTCACGTAGAATACGAAAAAATTTTTTGTTTTTCAAAATCTCGATTTTTGACATTCCCGTCCGCCATATTG
GATCCGCCATCTTAAATTTTCAAAATCTGACCTCGAATTCGGATTCAGCGCGGCTCGGGACCCTATGAAATATAAACAGT
TTCCCCGCACCATAGCTGTAACGCTTATGTGCACGAAAGGG**TTAA**

>SmiPBLE2 (TIR + STIR)

TTAACCCTTTCGTGAGCCGTGGGAAGATTGCTTCCCACCTACAAGATCCTACAAAATACACCTGTGGGAAGCGCAATTCC
CATTATAATATATTGATACTGCTATTAGAACGTTGGGAACGGAATTTCCCGCTTCAATTTTGTGCTGCCATCTAGTGTTC
AATCGACTGAATGGAATTGGCGGTAATTTTGGTTTCGAATGCCTGAAAGCTTTATGTGAGATTTGTTCTACTAATTCTGT
TTTGATGTGACATTTGTACTGCTCAACGATTCTGCTGCAAACGTGAGTAAAACACTGTTTTTATTGTAATAACATTTTCT
TATTCTGTATAATTCTGTACATTTTGTGCCAGTAGCTATTGTATAATAAATAATAGGTGAAATGTGAAAATGCTATTGTA
TCATATAAAAGACAAGTGCAATTTGTATACTATTGAAAAATTAGAATAGTTAGATTGTAAGTATTATGTTTAGATAATAA
ATAAAGAAAATATTTTTTTTTTTTAACTATATTTACATATTTGTTGCATTTCTGCACAATTTATGCCAGTAACTATCTAA
TTTGAATAATAAATACTAAATCAAAAGTGAAATAATTTTAACGATCGTAGTTCAAAAAAAAAATTATTCTCGGCTTCTTC
TCATAACTTTCAGATGAGTAAGAAAAAGTTTTTGTCCACATCAGCTGCTTTGGATATGCTGTTTAGCCTCCCTAGTGACC

CAGATGATTCTGAAGTAGAAGCTGATGATGATGCATTTTCAGAATTACAATCGCTAGAAAATTTAGATAAACCTAGCACA
AGTTCTGGAAGTAATTGTGTGACGACTGAGGGACTGTCCTCTGCTGATGAAGTATTGCAGATTGAACCTGATGTCAACAA
TTCTTCTGAAAGTGATGATGAAATGAATGAAGAATGGAGCAATGAAGTAGATTTCTTTGACAACATTGATTCTTCTTTCA
ATAATGAGCCAAATCTGAAAATATCATTCAATTCAACTGACAAGGAAAAGGATTATTTTGGAACAATTTTTACTGAAGAA
ATTTTACAGAATATTGTTGACGAAACAAATTCATATGCAAATCAACCTAAGAAAAATAAGAGATTGGGTAGTAATACAAA
AGACATTATTTTGAACTGGAAAGAGATCACATCAGAAGAATTGAAAGCATGGATAGGTATGCATATACTCATGGGAATAC
ATCAACTGCCTGAACTAAAAAACTACTGGAGCACAGACCCTATTCTAAATGTTCAATCAGTAGCTAGAGTAATGACTGCA
AAAAGGTTCAAAAAAATAACTGAGACTTTACATGTGAATGATAACAGCACAAATCCACCACGTGGACAACCTGGTCATGA
TAAACTTCATAAGGTCAGACCAATCCTCTCACAACTCAATGATTCCATATGTAAAGCCTATGAACCTTCATCGACTTTAT
CAATAGATGAATGTATGATAGCTTTCAAAGGGAGAACTTCATTGAAGCAGTATATGCCAATGAAGCCTATCAAAAGAGGG
TACAAGGTTTGGTGTTTAGCAGACTCTAAGACAGGTTATATTCAGAAATTTGAAATTTATTGTGGGAAGGAAACGTCTGA
TAATAAGGCTAAAAAATTAAATTTAGGGGAAACAGTAGTTACTAAACTAACAGAAGATGTGAAACACACAGGTCGCCTTA
TAGCCCATGATAACTATTTCACATCATTTAGGCTTGCAAAAAAACTACTTGCAAGTGGTCTATTTTCTGTGGGGACTGTG
AAACCGCAAAGGAAGGGTCTACCACAAATGCTGAAAACTAATGACAAACTAAAAAGGGGAGAATCAAAATTCCTAACTAA
GGAAGGTGTTGCTGCAATAAAATGGATGGACAATAAGGCTGTTACCTTGTTGACCACTGCACATAATCCAGCATTTATGA
CATCTGTACATCGAACAAAAAAAGATGGTACTAAAGCAGAAGTCCCTTGTCCAAAAGCTGTTGCGGTATATAACGATGTA
ATGGGAGGAGTTGATCGCTTCGATCAAAGAAAAGAAAGATACCAAATAGGAAGAAGATCTGTAAAATGGTGGCATCGTAT
CTTTTATTTTCTTCTTGACCTTGGCATCATAAATAGTTTCATACTGTGGCAAGTGAACAAAAGAAACAGAAATTTAGACC
AGCTAACATTCCGTATAGCATTAGCTCGTCAGCTAATAGACGGATATTCCTCAAGAAAAAGGAAAGGACGTCCTGCTAGC
TTCCAAGCAAAGAAGTGTGCAGTTCCGGATGATGTGCGCTTAGCCAGTGTGGGAAATCATATGCCAAAGATGGTTTCCAA
CTATAGACGATGTAGGAAATGTAGCACAAAAGGAAAAGAAAAGAGGACCCGCTGCATCTGTGCAGAATGTGATGTACCCT
TATGCATTGGAACATGCTTCTCATCTTTTCATGGCAAATAATCACCATTAAATTTAATTTTCTTTATGACTGTAAATGGT
TCCGTGGGAAGGATACTTCCCACCCCTGAAATTTAAATGTAAAAATAAAATTCGGCAAAATAAAATGCATTTCCTTAAAT
GTTATTGCTTAATTGTTAATAAAACATAAAAAACTGGGAGGAATTCAAAAGAAAAATAAATGGCCCACGAAAGGG**TTAA**

Clade PGBD8

>Api1-pble (TIR + IIR)

TTAACAGGTTGACTGCCACGACTATTTTTACATACTTGGTAAAATGTTCATGTGACACATATATGGGTCACACTTGATTT
GCAGTTTTTTTTCCGTGACCCTTATGTGTGTCACACAATACATTTCGTATTTTTATCGTGTCTCACCGGTGAGTCATGTT
TTCAGTATTTTTATTATCAATTTTTTATCAGCAAAATTTTATAAATATTTCACCACTGATAGCGCATGGAGATTATAATT
TTTTTCGACATCACAAAATTAATTTATATTGCTCACTTATAATCTCCTTATAAATAAATGTATTTAATATCACTGTTAAA
ATATGTCATATGATTATTTAACACTAAGATATAGATAAGAATGTGTTATTTAAGCGTGACCCGTATTAGTGTCTCACCAT
ATCATATAAATAAAATAAGTCAAAATGTTAAATAACAACGGTTATTATTATTAATTATTTTTATTTTGTTATACTCTTAT
TTATTTTTAATATTTTTATTATATATTAATATTCATAAATATGAATTCCAATGATATTTTTGATGCCCTCAATGAGAGTT
ATGACGAGTCATTATTTTCAATGGACGATACTGACTCTGATCCAAATTTTAATCCTGTAGGAATCAGTGATTCAGATATA
GATCCAGAAGGCATTTCAGAGGTATAATTTTATTTTTAAGAAGATGAATTATTTTGTTATCTATATTACTAAAATTTGTT
TTTAGGGTGACTTAAATGTTGATATACCAACCCAAACATCAACAAATATTCTACCAGATTTAAAAAATTTGACATGGGGC
CCAGCAGATAATGATTTACATGATTTTACATTTAATCCTGATGGATTAGATATTGGAATTAATGATGATATTTTATGTAC
ACTTAACAACGGAACACCTATTGATTTTTATTATTTATTTATTGATGATGAAGTTATATCATTAATTGTCACTGAAACAA
ACAAGTATGCAAATAAAAAGTTAGCTTCACCAAACATAAAAAGACATTCTAGATTACAAAAATGGATTGACACCAACGAA
TCAGAAATAAAAAATTTCCTTGGAATCATAATGTATATGGGTACAGTAAACATGCCAACTATTGCTAGTTATTGGACATC
GTCTCATTTGCTAACTTCTCACTTAAGCCATATTATGAGTCGTAACCGATTTGAACTACTATTATCTCACATACATTTTC
AAAATAATGATGTCGCTAACCTAAATAACAGACTTTATAAGATTCAAAATATTTTAGATTTGTTAAATCAAAAATTTAAA
CAATGGGTTATTCCATCACATGATATGTGTATTGATGAATCAATGATAGCTTTTCAAGGTCGCCTTGTTTTTAAACAATA
TATTCAAACAAAACGATATAGGTATGGAGTTAAAGTTTTTAAACTGTGTGTTAGTCCATGCTACACACTCAAGTTCAAGA
TTTATTCTGGTAAAGAATCTGTGGTTGACAAAAATGATAGTGTTTCTTCCAGAATTGTTATGAACCTTGTAGATGAGTAT
TTAGATCTTGGGCGCACACTTTATGTAGACAATTGGTATACTAGTGTTTCCTTAGCACATCAACTTCTTGAACGAAAAAC
ACATTTAGTTGGAACACTCCGTTCAAATAGAAAGTATAATCCTGACTATGTTATCAAAAAAAAATTAAAAAAGGGTCAAG
TAATAGCTCAAAAAAGTTCATCTAAAGTTATGGTATTAAAATTTAAAGATAAAAGAGATTTACACATGTTAAGCACAAAA
CATACAGATAATATGGTTATGGTTCCAAGAAGAGAAAATAAATCAAAGCCAGAAGTAGTATTAGATTATAACAGATGTAA
GTCTTTTATTGATCTTTCTGATCAAATGGCTTCATATGGCTCACCTCTTAGAAAATCATTAAAATGGTATCGTAAAGTAT
TTGAACTTCTTCTGAATACATCAGTTGTAAATGCTTTATGTTTGTTTCAAAAAACAACTGCAACACGCATTAAGATAACT
GATTTTAGATCAAATTTGATAAAATATCTGACATTCAAACCCAATATGAATCAAGAACTATCAAACAAACATGTACTCAC
AACAGCCAAAAGAAACAGATGTGTTCAATGTTACTTGAATATTGCTCAAGAAAAAGGAAGAAAATATTCTCAAAGTCATT
GCAAAAAAGTCACAACAAAATGTTTTATTTGCAACAAACAATTATGTGTTCCTTGCTTTTTTGTATTACACAAAACATAG
ACTGATTATTATTACCTACCTAAACTGTTAAATTATTATACTATTATTATGTTGTAATGTTTTTTTTTATGAAACAAGAC
TATTTACTAAACTAACAGTTAATATTATTTTATTTTATTTTTTATTGTTATTATTATGTTAATTTAAGTTGATTTTTTAT
GAAACAAGACTGTTTACAATAACAAAGAATTAATATTGTTTTATTTTAGTTTATTTTATACTGTTATTATAATGTTATGT
TGCTTTTATTATGAAATAAGACTTAACAAAGAGTTAATATTATTTGTTTAATTTTTAAATTGATTTTTATATTGTTATTA
CACTATGTTACTTATATTATGAAACAAGTTATCAATTACTATACAAATAAAGTGAATTTGTAAAATGTGTTTTTGTTTTA
TTTTATTATATTACCTACTTTTTTCATTGAGGGATCATTAAAGTCTTGCCAAAAAGTTTGTGTGACACACCCGTGGGTCA
CACCAAAAAATGGTTAAATTCTATATTAGAACACCATAAACTATATAATACCTATTTTACAAAAAATTATCAAAAAATTA
TTTGACAAGTTATGACAAAAACAGAATTAACAAAATCATTAATTATCTCAAAAACATTTTGACAAACATCATATGACCCA
CATAAGGGTCTCGTGGCAGTCAACCTG**TTAA**

>Api3-pble (TIR + IIR)

TTAACACGTTGACTGCCATGCGGCAGCCGGCGGTCGCATCGCTATCTTTTCTACAAGGCCACGTTCTATTTTCGATGTTA
TTCCATAAATGTACATTGGCATATTTATAGAACAAATAATCGGAAAAAAATAATTCCGATATTATTATCTTTGTTATGTC
TTTCTGCGTAAGAATATGATATTTAAAGTGTATGGTAGCCGGCGGTCGCATGGCGTTGTCATTATAATATAATTAGTCAG
TTGCCATATGACCGCCGGCAGTCGTACATAATAAAATTTCGTGTGATATTTTATTGTAGTTTTTTATTTTTAATTATGGC
GTCACGTCCAAAGAGACTTAAAGTAAATGACGATACGGTACGTCAATTATTAGATACTGATGAGAGTGGTAGTTCATTTA
GTGATTTCGATGACACAGATGAAGACGAAACTTACAACCCAATTATAACTTCTGATTCAGATGAACAAAGTGAAACCAGC
GATACCGAACATTCACAAGACACAATATTACCAGTTTCAAATCCTGGTGTTCAAATTTGGTCCAATGTAACAGGTAATAA
TCAAAAACAGCTAGTTTTTAATGGTTTTCCGGGATTAAAAATTAATATATCAGCTAATGCCACTGTTTGGGACTGTTTTA
ATTTATTTTTCACTGATGACATTATTGAACTTATTGTTATTGAAACGAACAGAAACGCGGAACAGTTTTTATCAAGTAAT
AGAATTTCCAAGTCTAGCAGATTTACAAAATGGGTTCCTACAGATTCAAAAGAGATTAAATTGTTTTCGGGTTTATTAAT
ATGGATGGGACTCGTTGAAATGCCAAATTTAGGTAGTTACTGGAGCACCAAACATAGATATAAAAACTATGTGGCACAAA
AAACAATGTCTAGAAATCGCTTCGAATTACTTCTAAGATTTTGGCACTTTTCAGACAATAAAAAAGCCCCTGAAGGTGAT
CGTATTTATAAAATACGTAATATAATTGATAAAATTGTAAACCGTTTTCAAAATGTAATGGAACCAGGTGAGGTATTGGC
CGTCGATGAGACAATGGTACCTTTTCGTGGCAGATTATTATTTAGACAATATATACCGGGAAAAGCCCATAAGTACGGTG
TAAAATTGTTTAAACTTTGCGGAACAAATGGTTATACTTATAACGTTCAAGTTTATGGAGGAAAGAGTCAAGTGGATGGA
AAGGGATTGGGCTGTAGAGTTGTTCTAGATCTAAGTAGAAGATATTTAAACATGGGACGTACAATGGTCACGGATAATTT
TTACACTTCAATTACACTTGCAAACGAACTTTTAAGTTATGATACACACCTAGTTGGAACTCTGCGATCAAATAGAGTAA
AATTACCCGAAATAACAAAAAAAAAGTTAAGTCCCGGAGAAATAATCGGAAAGGAAAATTCAGACGGTATTATTGTTGCG
AAATGGCGTGATAAAAGGGATGTAACTATGTTGTCAACAAAACATAATATTGATATGATTGACACAGGAAAAAAGAACAA
AAAAAAAGAATCGATAGTTAAGCCTAAGATAATTGTCGACTACAACTCCGGAAAAGCGGGAATTGATCTATCGGATCAAC
TGTCTAGCTATAGTACAGCTGTTCGTAAGTCTATACGTTGGTACCATAAGGTGGCGACAGAAATACTGTTTGGAACTGCT
ATGGTAAATGCTTTAATAGTTTACAATACAATTAATTCTGATAAAAAAATGAAGATAACACAATTTAGAGAAACATTAGT
TGATACAATACTCGGATTAGACAATCAAGGGCCACAAGAAGTTCAACAAGCAAGAGTTAACAGCAAACACATATTCCAAG
AAACAACAGAAAAATGTGGTAGAAACAGAAAAATCAGAAAGAGATGTTTCCACTGCTATCAAAGTAGAACAGTTATAGTT
GGATCGAAACAGGCATCGAAGGAAGTAAAAAAGATCACTACATTCTGCGCACCATGTAAATCTTATACATGTATAAGTTG
TTTTAATCTGCATCACACAAACTAAGGATTTTAATTATTTTTTACATATATTATTTATAAATACGTCTCGATAATTAGGT
TTTGTATTAAATTTATAATTTGTATTATGTTTATATTTAAGAATTGAAATTTGAAACATATTATTTGAATAAAAAAAGTG
ATTGAATTTTTTTTTATATGCTTATTGTAGTGATATTTTACGTATAGTAATAAATGTATATAAAATAATCGGTTTGCGGC
AAATAATATATCATTTTACGCGAAATTTTCTCCTCTACGCCATCATTTAAACATTATCTACATAACTTTGACGCCACAAA
AAATATTCCCCTTTACAAATAAGTGTGTTTATGGCAGCCCGCGGTCATATGGCGTATAAATAATACAAAGGTGCGACCGC
CGGCGTCCGTAAAAATACTGTTGTATGCTAGTAGTCTGCGGACGCCGGCGGCCACATGGCAGTCAACGTG**TTAA**

>Atr-pble1 (TIR + IIR)

TTAACACGTTCATCGCCATGGTGGGTCAAAATGACCGGCACTTGATCTAGCCATTTGGGCCATGGGGGATCAAAATGACC
GGCACGCCAGACCTACCTTTGCGAAGCCCGTTCGCGCCGCAGAATGCAAAATCTGTGTATTCAGCCATACAATGTGACCG
GCGCGGCCCCAATGAACACCTGTAACTACTTGGTTTTTTGATTTTTTGATGGCATACCTGTTTTTAGGACTTAGAGAAAA
AAGTTAATGAAAAAACATAGGATTTCGAATCATATATTTTTATTTCTGTTATAAGTTGCATTTATGCATAAAAATGAAAA
GACAACAGAATATTATTAATTTTAGGCATGTTTTTATCAAAATTTAAGAAAAAAAGAAAGACAATAGAACAAAAAATATG
TATTAGGAACGTTTTTTATCAAACCACAATTACAAATCGTCATACGAACAACAATCAAACTAAAGATTAATTCAACTCTA
AATGGATATTTAAACATTTTAGTGGCATTTGTCAAAACATTGACGACATAAATGGGGATTGCCGGGACAAATAGTGCAAA
ACCAAGAAACCTTCTTTGTTTTGTTCTTGGCGACATTTCTTCCATGCAACTCAACATTTGATTTGTAACAATCACGACAA
AATTTTCTTGTTGATACTGCTGCACCTTCTTTTTTCTTAAGCTCGTGTTTTTGTCTTCTGGGTCGTGGTGTGGTCAATCC
CGTATTTTCTTGACTACCCTCAATCAACTGCTTTGCCAACAGTCGCCTAAAATTGACGACAGATATCTTCTTCTTCGTTG
TACTTTCATATAGTACTCTTGCATTTACTACAGCAGTATTGAGTAATAATTCGATACCGAGCTTCTTATACCACTTTAAC
GTCTTTCGCAATGGACTCGAATAAGCACTCATTTGGTCAGACAAGTCAACGGCTGCTTTGGCTTTATTATAATCAATCAC
AATTTTCGGTTTAATTGTTTCAACCCCTCTTTTGTTCTTTTTTTTTTGTAAAACCAATTGAATGTTTGGTTGACAACAAAC
ACACGTCCCTCTTATCCTTCCACTTCATAACCGTGACTCCGTTGTTGTTTTCTTTTGCCATAAACTCACCCGGTTTCAGA
TTTCCAGTCATGACGTCCTTTGGTAAGTTCTTCCTGTTTTTTCGTAGGGTCCCTACTAAATGAGTATCGTGTTGAACTAA
CTTCTCAGCTAGATCTAAGCTTGTATACCAATTATCTACGAACATGGTATGACCTTTTCCAAAGTAGTTACGTGCTAAAT
TCAAGACAACGTTGGTTGGAGTATTATCCTGTGAATCTAATGTCTTTCCTCCATATACCTCCACCTTTGTCGTATAGCCA
GGAGTAGTACACAATTTAAATAGCTTGATGCCATATTTATGTCTTTTCGACTTATTATATTGGCGGAAAATGATACGCCC
ACGAAATGGTATCATACTTTCATCGACGCATAAACATTCATCGGGAGTATAATGAATTGAAAATGTTTGGTTCAATATTT
CAACCAATCTTCGTATTTTACCCAGCCGGTCATTTTTTGGAACTTCTTCGTTGTTGGAAAAGTGGAGCATTCTAAGCAAA
ACTTCGAAGCGGTTGCGACTCATAGTCTTTTTTGCAAAAGATTGACCCAACATTTCATCATTAGACCAATAGTCTACGAT
TTTTGGTAACTTCACTATACCCATCCAAATAATCACGCCTAAAAATCGTTTCATTTCTTCTTTGTCGGTTGGCGTCCAGA
GCCGTAGTCTCGACTTATCTTTAATGCCACCTTTATCGATACTCTGTTGGGCAAATATATTCGTCTGATGTGCTATATTT
TCAAATAACTGGTCAGTGACTAAGAGGGAATAGAAATCAATAGGTGTAGCATTGGCTAGTTGATTTCGAAGTGAGTAAGG
TTTCAGTCCCGGTATTTCGGTAAAAGGAATTAAACGAGGCTGCTTACCACGTGGCTGGTACCACTCACATTCAGGTGTAT
GGCCGGATAAAGCTGGAGACAAATCACGGTCATCTTCTTCACTGTCCGACAAAATAACTCTACGTTTTCTGGTACGTGGA
ATATTTATCACTTCATCACCCTCACATAAGACATGGTCAGTACTGCAATGTGACCTCTCTAAACTGTCAATTACATGTAA
TTGTTCTTGTGTAAACTCAAAGTCAAAAAAAGTGTCCATTTTAACAATAAAACTATTTTTTCTGCTAAACAAAATCTTAG
TGTGCATTTGATGACCAAAACAAGGTGCGAACTAACAAAAACAATCCGTGATTGGCTTGAGTCATAGGGAGAGGTGAAAG

GGGTGAAGGGAGGCGGGAACAACAACCGCACGCACGGTGTACTCCCGCGCTAGCATCCATTAGGCGCGCGGTACCCGTGC
GTAGTAGACGTGTCTATAACGTTTGTCTATGTCTCATCGATCGAAAATTACGCGTACATAAGTACGCTGACCGGCATTGT
GACCGGCACGGCCTGAATGTAATAATTATTCCGTTTAGGCCGCGCCGGGTCATAATGACCGGCACTTAAAAAAACCCTAA
AATAAAGCCAAAATAAACAAATATCACTGTAAAACTTTTATATATGAACAAATAAACCTATAAATAATCATAAAAACCAA
TGACATTTTTTTATGATTTTGTAAAATAATAGTCGGCGATGAACGTGTTAA

Clade PGBD9

>Pfl-pble (TIR + STIR or IIR; incomplete element in 3')

TTAACCCTTTGCAGTCCAGAGCGCCCTCAAACGCCCTATGGGGGTTGGCCAAAGCGCCCAAAAGCACCCGAACATTCCAC
ACTTCCGGGTTGATTTGAAAATGCGGGAAACGTAACGGTTTTCCGTGAGGTACGTATCATCAGTTTCTGGTGACGTACTA
AGCAAATGTGAGGCTATATTTTCCCGCTCATAAATATTCAGGTTGCCATTAAAAGATTTTGAATAAAATCGATCTGAGTT
TCCTGGGATGCGTCGCCAAGTTGAATTTTAAATTCTTGTTTACGTCTTTCCGTGGCGTTTACCGTCTATGAAAAATTATG
GATCCTGGTCGGGCAAATGCCGAAGTCATGAGAATTTTCATGGAAGACGACGATGAGGAGGAATTTTTAGGCTTCCCAGA
CATAGAAACAGCGAGTGACATCGACGTAGGAGAAATTGATGCGGCAATACGGCAAGTACTGGAGTTTGAGGACGGCGAGC
TACAACCGGAAGCGCCTGATAACTTGCGTGAGATTCCTTGGTCGGCCGATCTTGAAAATTTCGATCTGCCAAATTTTGTG
GTCGCTGAATCTGGTCCAAATCATCGTTTAGGTGTCGGAGCTCAGCCGTTAGATTATTTTTTTCTGTTAATTTCGCCTTC
TTTTCTTCAACTGGTTGTCGATGAGACCAACCGCTACGCAACTCAGAGTCAACGCATCAATGGTGCCGACCCGTCCTGGT
ATGAGACGAAGGTCGATGAAATGCGGGCTTTCATCGCCGTGAATATTTTGATGGGAATCCACACCCTCCCAGACGTAAAC
GACTACTGGTCAACTGATGACAGATTGCAAGTACCGGGGATCGCAAAAGTACTGCCGAAGAACCAGTACAAGAAGCTGTG
GCACTATTTTCACATCAATAACAACGAACTCGATCTGCCCGCCGGTGACCCCAACCGTGATCGCATGTTCAAGGTTCGAC
CTCTCCTTGACCTGGTTCATTTTTCTTTTGCTGCCAGCTATGTGCCATTTCGCGAGCTGTGGATCGATGAGGCTATGGTC
AAATTCAAAGGCCATAGCTACATGATCCAGTATATGCCAGCCAAGCCAGTCAAATGGGGTTTCAAGGTGTGGATGCTTGC
TTGTCCCCGTACCGGATACTGCCTGAGTCTCGACCCCTACACCGGAAAGAAAGGCGGACAGGTAGCACAGAAGGGCCTTG
GATACTCAGTCGTCTGGAATCTTTCAAGGATGTACCACAACATGTGGCATATTTTATTTTACGATAATTTTTTTAGTTCA
GTTTTAAGTCGCGCCCTAGATCTTCAGAAACAAGGTACCCTGTGCTGCAGTACTGTCCGCACAAACAGACTTGGGTTTCC
AGAAGACCGGAATTTGAAGTTGCGGGACAACGCAAGCGAAGCCCGGCAGTTTGGTAGACTGGTGGCAACATCATGGCAAG
ACAAGGGCACAGTCAACATCTTGTCAACATCATCGCAGCCAGAAATAACAACAATCAGACGGAGAGGTGGTGATGTATCA
TGTCCGAAGAGTGTGATAGACTACAACAAGTTCATGGGCGGAGTTGACCTTTCTGATCAACTCGCCACTTACTATCCACT
CGGAGGCAAGCATTACAAATACTGGAAATATTTGTTCTGGTTTATCATCAACAGAGCAATTTGTAATGCTTGGATCATGT
ACAAGATCGCCAACACACCACCTCCGAGGCATCGATATCGCCAAAAAGACTTCCGTCTTGATCTGTCAACTCAACTGGTG
GCTGGTTTCACTGCTAGGAAAAGACCAGGCCCTCCAGCAGCACCAGTCCTACTTCCATCTTCACTGGCTGCGGAAAACCT
GCCTTGCCATGAGTGTGTGAGAATCGACACTGCCAGAGGAAAAAAGACATGTGTCCTCTGTTCCAAACTTGGAAAGAAGA
CAGCCTCTGGCAGGGGTGTGCAAACATCATGGATGTGCTGTAAATGCAAGATAGCACTGTGTAAATTTGACTGTTTTGTC
AAATTTCATGCCATCCAAATGTGAGGTCAACATCTCTCACACATGCCAAACGGCCCTTTTCAGGGCCACTAAACCCTCAA
AAAAGAGAACTACCGTACACACATATATTTATTTTGAAATAAAGACATAGTCACACATTAATAAATCCTTTCATTTGTTA
TAAATATTGAAAATTTTGTTTTCAATATTACTTTTACTTTATTATAAATTCAAAAATTGTTCTTCAAAGTAACATCTTGT
AAAAAAGGAACACACAAACTAATATGAAT

>Cgi-pble (TIR + IIR)

TTAACTCATTCAGCCCCACTGACGAGTATTCTCGTCATATTTATGATGCGTTATTTAGCCGCTGACGAGTATAGTCGTCT
GGGTTTTCCCCGGCAATATGGCCGATGACGACTATACTCGTCGTTTTACCTGTATGTTAAACAAGCAGTGATTGGTCGAA
ACTAGAGGATTTCATGTACACACGGTTATACCAACCAATGAAATACATGTTTAAATGTCTTTGTGTGAAGTGTTCTTTAT
TTAGATGCGATATAAATAGCTGTAAACAAAGGAATTCGGCACATGTGTTAAATTTATCGATCGAAAATGGCGGACGTATT
GCGGGAAATATTACAAGACGATAGTGGCGAAGAATTTGAGGGTTTTGAATTGGAAGATGTAGAAAAAGCAGATAATACTT
ATGTCTATCATCTGCTTGCGTTGGATGATGCCGTNGAACAAGGATCAGACCTCGAAGTGTCCGATATTGAATCGGAATCA
GACCGCGAGTCAGAAACTGAAAGTAGTGACGATGACGATATCGTGCTGGCAGATTACAATCGTTGGACAGAGAATTTACG
GGACATTAATGTTCCTGCATTCAATGGACCACAGCCTGGACCTACACAAACACTCGAAGCGGATAAAAATGAAATAGATT
TCTTAAATTTATTGTTTCCTGAGGAGTTATACTCGAAGATATCCGAGGAAACAAACGCTTATGCTCAACTCTGCATCAGG
GCTAAAGCAAATCCTGGATGGTACGACACAACCCCCGAGGAAATACGTGCTTTTCTTGGATGTTTAGTCGTAATGGGGAT
TATTAGTGCACCTGCACAAGATCTCTATTGGTCAAAGGACAAACTGTTTCATTTGTCATGTATTGAAGAAAGGTTCACAA
GAACAAGATTCGAAAATATTCAGAGGTATTTTCACGTGGCCGACACCACACAAAATCCACCAAGGAACCAGCCTGGGCAT
GACAAACTTGTTCATGTACGAACAATAATGGAAACAATCAGAGAAAATTTCAAACGACAATACAATGCGCACAAAGAGGT
TTCCATTGACGAGGCCATGATTGGATTCTCAGGCCGGCTTGGTTTCAAGCAGTATGTTCCACTGAAACCTACCAAGCGTG
GAATAAAAGTGTGGGTTCGTGCTGATCCACACAATGGATTTATGAACGACTTTCAAGTCTACACTGGAAAGGAAGGAAAT
GTTCCAGAGGCTAACTNAGGTGGCAGAGTTGTCCTGGACCTGATGGACTCTGTTTTGAATCTCGGACATCATGTGTATTG
TGACAGTTTTTTCTCGTCCCCCGATCTGTTTTTGAAGCTTTGGAATGAAGGAACCGTTGCGTGTGGNACAGTAAAACCAA
ACAGGAAAGGANTGCCAAAGGACTTGGACAAGATTAAATTNAAAAACCAAGGNGACCACATCATCAAACAGAAGGCAAAC
CTTGTGGTGAGTGTATGGAGGGACAAGAGGAAACNGACAATCCTGTCAACTAACACCAATCAGAGTGACCAAGAGGTACA
GAGAAAACAGAAAGATGGGACCNTGAAGAATGTAACGTGTCCAATGTCAGTCAAGCTTTACAATGCTTACATGAATGGTG
TAGACCATGCTGATCAGCTTAGGTCCACATACAACATCGCAAGGAAAAGCCTAAAGTGGTGGAAGTACCTGTTCCTGTAC
GATGTGTGTATTGTAAATGCGTTCATCTTGATGCGCGAGTCCCCAAATCATGGCCCGAGGACCAAAAACAACAGGGTGAG
GACCAGAACTCAACTGGAGTTCAGAANGAAGGTAGCCCATCAGCTGATTNGACTTTTTTCTATGAAAAGGAAAAGNAAAT
CCGAGTTCAATGTTGAGCATGTGATGNAACATCACTGGCCNAGCACCATGCCGAAGAAGAGGAGGTGTAAACAATGCACA
AAACCTNGTGTCAGACGAGAACCATTGAGTGGATGTTCCCAGTGTGATGTCAACCTATGCAGCGACTGCTTCAAACCATT

CCNTTTGGAAAACTTTCCAGACATTTATGGACAATAAACANAGAAATATCATTGTTTATTATGATTTTGAGAAGTGTTTC
CTTTTGTTTTTCTTTATGGAAATAAACAATTTTTAAGGTATGGTAATAAAAGTTATATCTTTAGAATCTGCATTTATTCT
TCTTTGTTCCCTGAAAATATGAAAAACAGATGTGAATAATTAATGAAGCAGTGGCTGTTTATCTANTGTCTGCTGNCCGC
TCATCAGCAGCAGGGATTCTGCTGGGGCTTATCTCAGATAAGGCGGCTGGGGCTGAATGAGTTAA

>Bgl-pble (TIR + IIR)
TTAACTCATTAGCTACTAGGCCCTTACCCCAGCTTCATTGCCTGAAATACTAAGCAAAGGGAGGTAACTCTAATTTTCAC
ATAGAGTCATCTTACAAAAAAATTTATTAAAAATACTAAATATCATCTAATTACTTTTAAAAAATCTTTAAAAAAAAGAG
AAATGAATGACCTTTCCAAAAATATATATTTTCAATTATTTTTATAAACAAATAAATGATAAAAATCAACATAAACACAA
AAGTTTTTGCTTTTCTACTTTACTGCTGCCTGAGAAACTCATGAAATTTGGAATAACAGAATAACAGAGTCCTCTGTGAA
GATGAACTTGACACATATGACACCCAGTTCTGGTCTCACTGGTTCTGCCACTTGCTGTTTTGTTCCCATCTTTAGAGCAC
TGAAAGCAGCGGCGTTTTCGGCCAGGCATAGGCTGCAACATGTGATGGCTGTTGTATGGCTTTGATCTTCCTGCAACACT
TGGTGTGTCTAGTAGTGTAACTCCTGTTGCTGCTGTCAGAAGCTGATCAGCTACTTTCTTGTGAAACAATAAGTGATTTT
GAGATAAAGAATCTGTAGAATCATTGGGTCTAGATTTTTTCATCAATAAAAAGCTGTTGATCATAGAGATGTCTATCAGA
TACCATAGCATGTATCTCCACCACTTGTTGCTTACACGTCCAACTGGGTAGTATGAACGATGCTGGTCTGCCAGGTCTAC
TCCAAACATGCCACTATTATAATTGACAACTGGTAGAGGTATTTCCACATCTAGCATGCCTGTTTTAGACCTACGCTCAA
CTGTGGTCATGACAGGGTCCGAGTTTGTGGACAGGATATTGACTTGCCGCTTATCGGTCCACTGTGTAGCAACAATCAGC
CCTTTCTGTCTCATCTTCAAATCTCCTTTCTTTTGTTTTGATTTAGGCAGAGGCCAACCCTTTCTACGAGCTCTGATGGT
AGCACAGCTATAGGTCTTTCTTTCCAACAAGTCCTCAGCTAACTTCAAAGAGGAAAAAAAATTGTCAAAATACACTTCAT
GGTATTTGTCCAAGTATGGCTCAGTCATCGACATGACCACATTGTATCCCAGGCCATTGGGAGATGAATAATTTGATTTG
CCTGTGTACACCCTAAAATTTAGTACATAGCCTGATGAGGCTTCTGCACAACACCAAATCTTGATGCCCCAAGGAGTAGG
CTTGTTCTTCAAATACTGCTTCATAGCGAGGCGACCACGAAAGCCTACCATTGCCTCATCAACACTCAGTTTGCGCTGAG
GTTTGTATACTGCTTGACATGTACTCCTTACATGATCCAGTAGAGGCCTAATTTTGTACAATGGGTCATAGTTGTCACTA
TTTCTTGGTGGAAAGTGTGCACTGTCATTGAGGTGAAAATATTGCCTGATCTTTTCAAACCGAGTGCGGCCCATTACATC
AGCTATTGCTGGCACCCTCCAGCTTTTGTCCGAGGACCAGTAGAGCTTAGTACTTGGAATCCAATGAATACCAAACATAA
TATTGATGTACAGGAAGGTGTAAATATCTTGAGCAGTCACTGGTTTCCACTGGGAATCCCTCTGGCCTTTTGCTACCTGC
TGTTGCTCTGCATATCTATTGGTTTCTGTAACCATCAACCCTATCATTTCATCCGTAATAAATAATTTGAAATAATCTAA
AGGACTGCAGTCTTCTTCTAGATAGTGAACACCCCCAACTATTTCATCAAATGGTAGCATTTGATCTGAAACACTGGTCA
AAGTATTGGTCCAAGATAAAGATTCAGTACTAGATCTAGGATCAGCGACATTATCATTAGGACTAGCTTCAGTTACAGTA
CTAACACTAACAGAATTGTTTACATCTAGATCTAGATCTATATCACCGTCAGTATCGCTATCAGATTCTACTTCTAGATC
TAGGTCTAAATTATCTAAATCTAAATCAGATTCTGCCTCATTATCTGAATCTATAGACCCTCTAGTATCAGCAATAAGCG
AGTCAAGAAACTGAAACGCCATTGTAGCAGAAGTCAAAATTGAGAGATCAAAGCGACTGAAATTTCACTTACTTTTGAAG
TCGCATAGAAAACGAAAGAGAAATCAGAAAAACATCCGTTTTTTTTTGCTGAAACGATTTACATTCCTACTTTTTAAAAA
ACCCAGCCAGTGACAAAAAACCTCTCGCTTACTGTGAGAAAAAGAGAGCTGAACACGTCACGTTGTGGCAGGAGGCGTGG
CGGGCACATTTGGACGCATTAGTAGTGAGGGGGGGATGGGCGCATTTGGTCGCATTAGTAGCTAATGAGTTAA

pble outgroup
>IFP2 (TIR + STIR)
TTAACCCTAGAAAGATAGTCTGCGTAAAATTGACGCATGCATTCTTGAAATATTGCTCTCTCTTTCTAAATAGCGCGAAT
CCGTCGCTGTGCATTTAGGACATCTCAGTCGCCGCTTGGAGCTCCCGTGAGGCGTGCTTGTCAATGCGGTAAGTGTCACT
GATTTTGAACTATAACGACCGCGTGAGTCAAAATGACGCATGATTATCTTTTACGTGACTTTTAAGATTTAACTCATACG
ATAATTATATTGTTATTTCATGTTCTACTTACGTGATAACTTATTATATATATATTTTCTTGTTATAGATATCAAGCTTA
GATCTGCCAATTGCTCGAGCCAgctagcTGGGGATCCAGACATGATAAGATACATTGATGAGTTTGGACAAACCACAACT
AGAATGCAGTGAAAAAAATGCTTTATTTGTGAAATTTGTGATGCTATTGCTTTATTTGTAACCATTATAAGCTGCAATAA
ACAAGTTGGGGTGGGCGAAGAACTCCAGCATGAGATCCCCGCGCTGGAGGATCATCCAGCCGGCGTCCCGGAAAACGATT
CCGAAGCCCAACCTTTCATAGAAGGCGGCGGTGGAATCGAAATCTCGTGATGGCAGGTTGGGCGTCGCTTGGTCGGTCAT
TTCGAACCCCAGAGTCCCGCTCAGAAGAACTCGTCAAGAAGGCGATAGAAGGCGATGCGCTGCGAATCGGGAGCGGCGAT
ACCGTAAAGCACGAGGAAGCGGTCAGCCCATTCGCCGCCAAGCTCTTCAGCAATATCACGGGTAGCCAACGCTATGTCCT
GATAGCGGTCCGCCACACCCAGCCGGCCACAGTCGATGAATCCAGAAAAGCGGCCATTTTCCACCATGATATTCGGCAAG
CAGGCATCGCCATGGGTCACGACGAGATCCTCGCCGTCGGGCATGCGCGCCTTGAGCCTGGCGAACAGTTCGGCTGGCGC
GAGCCCCTGATGCTCTTCGTCCAGATCATCCTGATCGACAAGACCGGCTTCCATCCGAGTACGTGCTCGCTCGATGCGAT
GTTTCGCTTGGTGGTCGAATGGGCAGGTAGCCGGATCAAGCGTATGCAGCCGCCGCATTGCATCAGCCATGATGGATACT
TTCTCGGCAGGAGCAAGGTGAGATGACAGGAGATCCTGCCCCGGCACTTCGCCCAATAGCAGCCAGTCCCTTCCCGCTTC
AGTGACAACGTCGAGCACAGCTGCGCAAGGAACGCCCGTCGTGGCCAGCCACGATAGCCGCGCTGCCTCGTCCTGCAGTT
CATTCAGGGCACCGGACAGGTCGGTCTTGACAAAAAGAACCGGGCGCCCCTGCGCTGACAGCCGGAACACGGCGGCATCA
GAGCAGCCGATTGTCTGTTGTGCCCAGTCATAGCCGAATAGCCTCTCCACCCAAGCGGCCGGAGAACCTGCGTGCAATCC
ATCTTGTTCAATCATGCGAAACGATCCTCATCCTGTCTCTTGATCAGATCCGAAAATGGATATCCAAGCTCCCGGGAGCT
TTTTGCAAAAGCCTAGGCCTCCAAAAAAGCCTCCTCACTACTTCTGGAATAGCTCAGAGGCCGAGGCGGCCTCGGCCTCT
GCATAAATAAAAAAAATTAGTCAGCCATGGGGCGGAGAATGGGCGGAACTGGGCGGAGTTAGGGGCGGGATGGGCGGAGT
TAGGGGCGGGACTATGGTTGCTGACTAATTGAGATGCATGCTTTGCATACTTCTGCCTGCTGGGGAGCCTGGGGACTTTC
CACACCTGGTTGCTGACTAATTGAGATGCATGCTTTGCATACTTCTGCCTGCCTGGGGAGCTGGGGACTTTCCACACCCT
AACTGACACACATTCCACAGAATTCGTCGACATCGATACCAAAAGTTTTGTTACTTTATAGAAGAAATTTTGAGTTTTTG
TTTTTTTTTAATAAATAAATAAACATAAATAAATTGTTTGTTGAATTTATTATTAGTATGTAAGTGTAAATATAATAAAA

CTTAATATCTATTCAAATTAATAAATAAACCTCGATATACAGACCGATAAAACACATGCGTCAATTTTACGCATGATTAT
CTTTAACGTACGTCACAATATGATTATCTTTCTAGGGTTAA

Supplementary data 6.

>GFPT01020379.1

ttaacccttgaagcacacattaaggatttaatttatttttaatttttacaatgaaaacttatttct
aaatacatttaaaatttaatttgagcaaaaataa--ttttt-
cattaatattgagtgttgctaaaaaatggtggtcattcagtgaaaccactatgcaactaagcttcact
atatgtttacatggtgtggtatgcttcaaagcaatttttagcttcattgagacacaatgccacttcgc
atgtcttacaagaccatttacttcggtgtggctgtaacttc-
gtgctgcagtgggcacaacgtctgcttgaactcc----
ttacaggtaaatgagccaccttgtcaaatctttttcaagaggtacttgaggcttgaaatagctcact
ggctttggaattggttttcttcctctaggtgaagattgtgttgcaccaatcaaaccatctactacaga
aatacgaaatttctttagtgataaatgggtgccctcactttggagtgcataaatgacaaagccatttg
tcagtgcaacatccacaaagtgccaaaatatacggtgccaccacttgcggc--atttc----
cggcttatttcgtatagggattt-
caacatatcagctttatccacatagcccatgtgctgattgtaatctttcactagttgtggacaggtta
ctatttgtattgttccatctttgtttttctcttggcttcagaaataagcgaaggattgtggaaattg
gacaaaaatgtaatacatttgttgtctttccactttagccaagcaactcctgtaaatgaggtgcgaaa
ctcactatcacctttcttcatgttcttttcta---
tcttttgttcttttggcaagccagttctcacattccttactgttccgcatgcatatatcgcttggttc
tttagtgaaatcattagatcaacacttgtaaaataattgtcgaaataaacatgatgacccaacccaac
aagttcaagagtaaggtctttgacaactcgggcacctaatagcttctcaatattgccttcaaccttac
ctgtgtatatttgaaactgacagacaaaccccgactcatccgcacgcacccaaaccttgtagcctcgt
tttattggcttgttgggcataaattgtctgaaagaactccttcctttaaatttgatcatactttcatc
gatggactgatgattccctggtttgaagcaggttttgaaagtctctgaaagtgtatcaagcattggtc
ttagtttgtagagcttatcaaaagcattgtcgtctcttttggctgcaatacattgttgtttacatgt
aaatgcgtcaacaaccaaccaaaacgattcactggcattaggggagagataaatgaatccctaagctg
agggtaagatgaccaatggtccctgtaggagggtaaacgttttatccccatcataatgtttattgcta
tga---
aaatcttcatttcttcttcagttacagggacgaatggtttccccaactgtgtagcataaaggttagtt
tgaaaaactaaatgttcaataaaagaatctggaaataaacacaaaaaagtatcgacaggagattctgt
acttgttggtacattagggccaaagtcatttgtaaaattttgaatgttccaccgccttattttgtttt
cccacacaccatcta-atttttaaatcgctcactaaggggttcatcatcatcatcatcatc------
tgtattacaatcttcctgatttatatcagaataatttaggacattattgttattattgaaaccgttat
cttcctgtataaaaacaattgatctgtcctcactttcctgtaagcaatcttctaaatccgaatcacat
tc----
tccatcagaatcacactgtccatcaatagagtcagtatcactaggcacttctattagcaaacgttcaa
catcttcctctttgagtgatacgtttagcaataattttcttttat----------
ttatagacatgacgacactgaaatgagtaaaaatatacacatgaccttacttgcacagtggtagcaat
ttgaaaccacctagtaattag---g--
ttattaaaattaatatatttagataaatttaattatgaatgtaaataattataaagcaaaaaataata
gtattaaataatgatttgaatact-------
tacatagttatatatttattccttgaaaaatataacaaaaccaattagaataaacgtgtcacttcaat
aacaaaaacaaaatacatactatattggaggcatgaataccaacaaaaattactaataataaattttt
acaagaataaaactagcaaataaattcagatagatgattcataagtgctgcca-
acagttaaatgtgtctcaaataattaaataaaaggtggttaccaagtgaaaccactatgcttcaaagg
gttaa

>MSSY03005913.1b

ttaacccttgaagcacacattaaggatttaatttatttttaatttttacaatgaaaacttatttct
aaatacatttaaaatttaatttgagcaaaaataa--ttttt-
cattaatattgagtgttgctaaaaaatggtggtcattc----------


```
------------------------------------------------------------------
------------------------------------------------------------------
------------------------------------------------------------------
------------------------------------------------------------------
------------------------------------------------------------------
------------------------------------------------------------------
-----------------------------------------------------
nggtagcaatttgaaaccacctagtaattag---g--
ttattaaaattaatatatttagataaatttaattatgaatgtaaataattataaagcaaaaaacaata
gtattaaataatgatttgaatacttagggcctacatagttatatatttattcattgaaaa--------
----
caattagaagaaacgtgtcacttcaataacaaaaacaaaagacatactatattggaggcatgaatacc
tacaaaaattactaataataaatttttacaagaataaaactagcaaataaattcagctagatgattca
taagtgctgccataatgttaaatgtttctcaaataattaaataaaaggtggttagcaagtgaaaccac
tatgcttcaaagggttaa
>PGFK01000003.1
----ccctttgacgcacacattaagga--
taatttatttttatctttttacaattaaaacttatttctaaatatatttaaaatttcattttagcaaa
taata--tgttt-cattaagatcgagtgttgctaaaaaatggtg----------
aaaccactatgcaactaagctccactatatttttacatggtgtggtatgctacaaagcaatttttaga
ctcgttgagaaacaatgcaacttcgcatttcttacaagaccatttacttcggtgtggctgtaacttcc
gtgttgcattgggcacaacgtctgcttgaactcc----
ttacaggtaaatgagctacctttttcaaatctttttttcaagaggtactcgagactgaaaatagctcact
ggctttggaagtggttttcttcctctaggtgaaggttttgttggactaatcaattcatctactacaga
aatacgaaattcctttagggataaatgggagccctcactttagagtgcataaatgacaaaaccatttg
tcagtgcaacatccacaaagtgccaaaatatacggtgccaccatttg--------------
cggcttatttcgtaaagggattt-
caacatatcagctttatccacatagcccatgtgctggttataatctttcattagttatggacagacta
ccatctgtattattccatctttgtttttttctcttggcttcataaataagcgaaggattgcggaaattg
gccaaaaatgtaatacatttgttatctttccactttagccaagtaactcctgtaaatgaggtgtgaca
ctcactatcacctttcttcatgttctttttctacctttttttttttctttttggcaagccagttctcacat
cccttattgtgccgcatgcatatatcgcttggttctttagtgaaatcattagctcaacacttg-----
------------------------------------------------------------------
------------------------------------------------------------------
------------------------------------------------------------------
----------------------------------------ct-
atgatttttt------------------------------------------------------
------------------------------------------------------------------
------------------------------------------------------------------
------------------------------------------------------------------
ttttcatttcttcttcaattacagggacgaatggtttccacaactgtgtagcgtaaaggttagtttga
aaaactaaatgttcaattaaagaatctggaaataaacacaaaaaattatcaacaggagattatttact
tgttggtacaatagggccaaggtaattagtaaaatgttgaatgatccaccaccttatttgttttccc
acacactatctatttttttttaatcgttcactaagtggttcatcatcattatcatcatc------
tttaatacaatcttcctgatttatatcagaataatttaggacattattgttattattgaaaccgttat
cttcctgtataaaaacaattgatctgtcctcactttcctgcaagcaatcttctaaatccgaatcacgt
tc----
tccgtcagaatcacactgtccatcaatagagtcagtatcactaggcacttctattagcaaacgttcaa
catcttcctctttgagtaatacgtttagcaattatttttatttttatttta-----
tttatagacatgacgacactgaaatgagtaaatatatacacatgacctatcttgcacagtggtagcaa
tttgaaaccacctagtgattagttagatttatttaaatatatgtattttaatg-----
acctattaatgtaaataattataaagcaataaacaatagtattaaaagatgatttgaatatt------
-
taaatagttctatatttattccttgaaaa-------------------------------------
```

------
aacaaactacatactatattataggcatgaataccaacaaaaattactaataataaatttttacttga
ataaaactagcaaataagttatgatagatgattaataagtgctgcca-
acagttaaatgtatttcaaataattaaatgaaaggtggttactaggggaaaccactatgcttcaa-
ggtttaa
>PGTA01000295.1
TTAACCCTTTGAAGCACACATTAAGGATTTAATTAATTTTTAATTTTTTACAATGAAAACTTATTTCT
AAATACATCTAAAATTTAATTTGAGCAAAAATAA--TGTTT-
CATTAATATTGAGTGTTGCTAAAAAATGGT---
CATTCAGTGAAAGCACTATGCAACTAAGCTTCACTATATGTTTACATGGTGTGGTATGCTACAAATCA
ATTTTTAGATTCATTGAGACACAATGCCACTTCGCATGTCTTACAAGACCATTTGCTTCGGTGTGGCT
GTAACTTCT-TGCTGCAGTGGGCACAACGTCTGCTTGAACTCC----
TTACAGGTAAATGAGCCACCTTGTCAAATCTTTTTTCAATAGGTACTTGAGGCTTGAAATAGCTCACT
GGCTTTGGAATTGGTTTTCTTCCTCTAGGTGAAGATTGTGTTGCACCAATCAAACCATCTACTACAGA
AATACGAAATTTCTTTAGTGATAAATGGGTGCCCTCACTTTGGAGTGCATAAATGACAAAGCCATTTG
TCAGTGCAACATCCACAAAGTGCCAAAATATACGGTGCCACCACTTGCGGC--
ATTTCTTTCCGGCTTATTTCGTATAGGGATTTTCAACATATCAGCTTTATCCACATAGCCCATGTGCT
GGTTGTAATCTTTCACTAGTTGTGGACAGGTTACTATTTGTATTGTTCCATCTTTGTTTTTTCTCTTG
GCTTCAGAAATAAGCGAAGGATTGTGGAAACTGGAGAAAAATGTAATACATTTGTTGTCTTTCCACTT
TAGCCAAGCAACTCCTGTAAATGATGTGCGAAACTCACTATCACCTTTCTTCATGTTCTTTTCTACC-
--
TTTTGTTCTTTTGGCAAGCCAGTTATCACATTCCTTACTGTGCCGCATGCATATATCGCTTGGTTCTT
TAGTGAAATCATTAGATCAACACTTGTAAAATAATTGTCGAAATAAACATGATGACCCAACCCAAAAA
GTCCAAAAGTAAAGTCTTTGACAACTTGAGCACCTAATAGCTTCTCAATATTGCCTTCAACCTTACCT
GTGTATATTTGAAACTGACAGATAAACCCCGACTCATCCGCACGCACCCACACCTTGTAGCCTCGTTT
TATTGGCTTGTTGGGCATAAATTGTCTGAAAGAACTCCTTCCTTTAAATTTGATCATACTTTCATCAA
TGGACTGATGTTTTCCTGGTTTGAAGCAGGTTTTGAAAGTCTCTGAAAGTGTATCAAGCATTGGTCTT
AGTTTGTAGAGCTTATCAAAAGCATTGTCGTCTCTTTTTGGCTGCAATACATTGTTGTTTACATGTAA
ATGCGTCAACAACCAACCAAAACGATTCACTGCCATTAGGGGAGAGATAAATGAATCCCTAAGCTGAG
GGTAAGATGACCAATGGTACCTGTAGGAGGGTAAACGTTTTATTCCCATCATTATGTTTATTGCTATG
A---
AAATTTTCATTTCTTCTTCAGTTACA----------
>PGTA01003554
TTAACCCTTTGAAGCACACATTAAGGATTTAATTTATTTTTATTTTTTTGCTATGAAAACTTATTTCT
AAATACATTTCAAATTTAATTTTAGCAAAAAAATAAAGTTTTCATTCAGATTGAGTGTTGCTAAAAAA
TGGTGGTCTTTCAGTAAAACCACTATTCAAATAAATTCCAGTATATTTTTACATGGTGTGGTATGCTA
CAAAGCAATTTTTATATTCATTGAGACACAATGCCACTTCGCATGTCTTACAAGACCATTTACTTCGG
TGTGGCTGTAACTTCT-TGCTGCAGTGGGCACAACGTCTGCTTGAACTCC----
TTACAGGTAAATGAGCCACCTTGTCCAATCTTTTTTCAAGAGGTACTTGAGGCTTGAAATAGCTCACT
GGCTTTGGAAGTGGTTTTCTTACTCCAGGTGAAGATTGTGTTGCACCAATCAATCCATCTACTACAGA

```
AAT--GA----------GA--
AATGGGAGCCCTCACTTTAGAGTGCATAAATTACAAGGCCATTTGACAGTGCAACACCCACAAAGTGT
CAGAATATACGATGCGG-CATTTCCGGCTTATTTC--------------------------
AACATATCAGCTTTATCCACATAGCCCATGTACTGGTTGTAATCTTTCACTAGTTGTGGA-------
CC---
TGTATTGTTCCATCTTCGTTTTTCCTCTTGACTTCAGAAATAAGCGAGGGATTGTGGGAATT------
---------
ACATTTGTTGTCTTTCCACTTTAGCCAAGCAACTCCCGTAAATTAAGTGTGAAACTCACTATCACCTT
TCTTCGTGTTCTTTTCTACC---
TTTTGTTCTTTTGGCAAGCTGGTTCTCACATTCATCACTGTGCCGCATGCATGTATCGATTGGTTCTT
TAGTGAAATAATTGAATCAACACTTATAAAATAGTTGTC---
ATAAACATCATGACCCAAACCAAAAAGTTCAAAAGTTAGACCTTTGGCAACTCGAGCACCTAATAGTT
TCTCAATATTGCCCTCAACCTTCCCTATGTATATCTGAAACTGAC--------------
TTACCCGCACGCACCCACACCTTGTAGCCTCGTTTTATTGGATTGTTGGGCATAGATTGTCTGAAAGA
ACTCCTTCCTTTAAATATGATCATACTTCCATCAATGGACTGATGTTTTCCTGGTTTGAAGCAGGTTA
TGAAAGTCACTGAAAGTGTATCAAGAATTGGTCGTAGTTTGTAGAGTTTGTCAAATGCATTGTCTCCT
CTTTTTAGCT-
TAATACTTTGTTGTTTACATGTAAATGTGTCAACAACCAACCAAAACGATTCACTGGCATCAGGGGAG
AGATAAATGTATCCCTAAGCTGAGGTTGATATGACCAATGGTCCCTTTAGGAGGGTAAACGTTTTATC
CCCATCATAATGTTTATTATTATGA---
AAATTTTCATTTATTCTGCAGTTACAAGGATGAATAGTTTCCCCATCTGTGTAGAGTAAAGGTAAGTT
TGAAAAACTAAAGGTTCACTGAAAGAATCTGGATATAAACACAAAAAAGTATTGACAGGAGATTCTGT
ACTTGTTGGTACATTAGGGCCAAAGTCATTTGTCAAATTTTGAATGTTCCACCACATTATTTTGTTTT
CCCACAAACCATATA-TTTTTTTAAATCCCTCACTAAGTGGTTCATCATCATCATCATCAT------
ATTTATTACTATCTTCCTGACTTACATCAGAATAATTTAGGACATTATTGTCATTATTGAACCCGTCA
TCTTCCTGTATAAAAACACTTGATCTGTCCTCACTTTCCTGTAAGCAATCTTCTCCTTA-
GAATCACACTCACCTTCC----------------------
ATAGAGTCAGTATCCCTAGGCACTTCTACTAGCAAACGTTCAACATCTTCCTCTTTGAGTGATAGACT
AAGTTTAGCAATCATTTTATTTTA-----TTTATAGACATGAC---
ACTGAAATGATTACAAATATGCACACAGCTTAACTCGCACAATGGTAGCAATTTGAAGCCACCTAGTA
ATTAG-----TTTATTTAAATATAT-TATATAT-
TATATATATATTT------------------------------------------------------
------------------------------------------------------------------
------------------------------------------------------------------
------------------------------------------------------------------
------------------------------------------------------------------
>PGFK01000005.1
TTAACCTTTTGAAGCACACATTAAGGATTTAATTTATTTTTATTTTTTTGCTATGAAAACTTATTTCT
AAATACATTTCAAATTTAATTTTAGCAAAAAAATAAAGTTTTCATTCAGATTGAGTGTTGCTAAAAAA
TTGTGGTCTTTCAGTAAAACCACTATTCAAATAAATTCCAATATATTTTTACATGGTGTGGTATGCTA
CAAAGCAATTTTTATATTCATTGAGACACAATGCCACTTCGCATGTCTTACAAGACCATTTACTTCGG
TGTGGCTGTAACTTCT-
TGCTGCAGTGGGCACAACGTCTGCTTGAACTCCCTCCTTACAGGTAAATGAGCTACCTTGTCCAATCT
TTTTTCAAGAGGTACTTGAGGCTTGAAATAGCTCACTGGCTTTGGAAGTGGTTTTCTTACTCCAGGTG
AAGATTGTGTTGCACCAATCAATCCATCTACTACAGAAATGAGAAAT----------------
GGGAGCCCTCACTTTAGAGTGCATAAATTACAAGGCCATTTGACAGTGCAACACCCACAAAGTGTCAG
AATATACGATGCGG-CATTTCCGGCTTATTTC--------------------------
AACATATCAGCTTTATCCACATAGCCCATGTACTGGTTGTAATCTTTCACTAGTTGTGGA-------
C--
CTGTATTGTTCCATCTTCGTTTTTCCTCTTGACTTCAGAAATAAGCGAGGGATTGTGGGAATT-----
----------
ACATTTGTTGTCTTTCCACTTTAGCCAAGCAACTCCCGTAAATTAAGTGTGAAACTCACTATCACCTT
TCTTCGTGTTCTTTTCTACC---
TTTTGTTCTTTTGGCAAGCTGGTTCTCACATTCCTCACTGTGCCGCATGCATGTATCGATTGGTTCTT
```

TAGTGAAATAATTGAATCAACACTTATAAAATAGTTGTC---
ATAAACATCATGACCCAAACCAAAAAGTTCAAAAGTTAGACCTTTGGCAACTCGAGCACCTAATAGTT
TCTCAATATTGTCCTCAACCTTCCCTATGTATATCTGAAACTGAC--------------
TTACCCGCACGCACCCACACCTTGTAGCCTCGTTTTATTGGATTGTTGGGCATAGATTGTCTGAAAGA
ACTCCTTCCTTTAAATATGATCATACTTCCATCAATGGACTGATGTTTTCCTGGTTTGAAGCAGGTTA
TGAAAGTCACTGAAAGTGTATCAAGAATTGGTCGTAGTTTGTAGAGTTTGTCAAATGCATTGTCTCCT
CTTTTTAGCT-
TAATACTTTGTTGTTTACATGTAAATGTGTCAACAACCAACCAAAACGATTCACTGGCATCAGGGGAG
AGATAAATGTATCCCTAAGCTGAGGTTGATATGACCAATGGCCCCTTTAGGAGGGTAAACGTTTTATC
CCCATCATAATGTTTATTATTATGA---
AAATTTTCATTTATTCTGCAGTTACAAGGATGAATAGTTTCCCCATCTGTGTAGAGTAAAGGTAAGTT
TGAAAAACTAAATGTTCACTGAAAGAATCTGGATATAAACACAAAAAAGTATTGACAGGAGATTCTGT
ACTTGTTGGTACATTAGGGCCAAAGTCATTTGTCAAATTTTGAATGTTCCACCACATTATTTTGTTTT
CCCAAAAACCATATATATTTTTTAAATCCCTCACTAAGTGGTTCATAATCATCATAATCATCATCATA
TTTATTACTATCTTCCTGACTTACATCAGAATAATTTAGGACATTATTGTCATTATTGAACCCGTCAT
CTTCCTGTATAAAAACACTTGATCTGTCCTCACTTTCCTGTAAGCAATCTTCTCCTTA-
GAATCACACTCACCTTCC--------------------
ATAGAGTCAGTATCCCTAGGCACTTCAACTAGCAAACGTTCAACATCTTCCTCTTTGAGTGATAAGTT
TAG-----CAATCATTTTATTTTA-----TTTATAGACATGAC---
ACTGAAATGATTACAAATATGCACACAGCTTAACTCGCACAATGGTAGCAATTTGAAGCCACCTAGTA
ATTAG-----TTTATTTAAATATAT-TATATAT-
TATATA----------

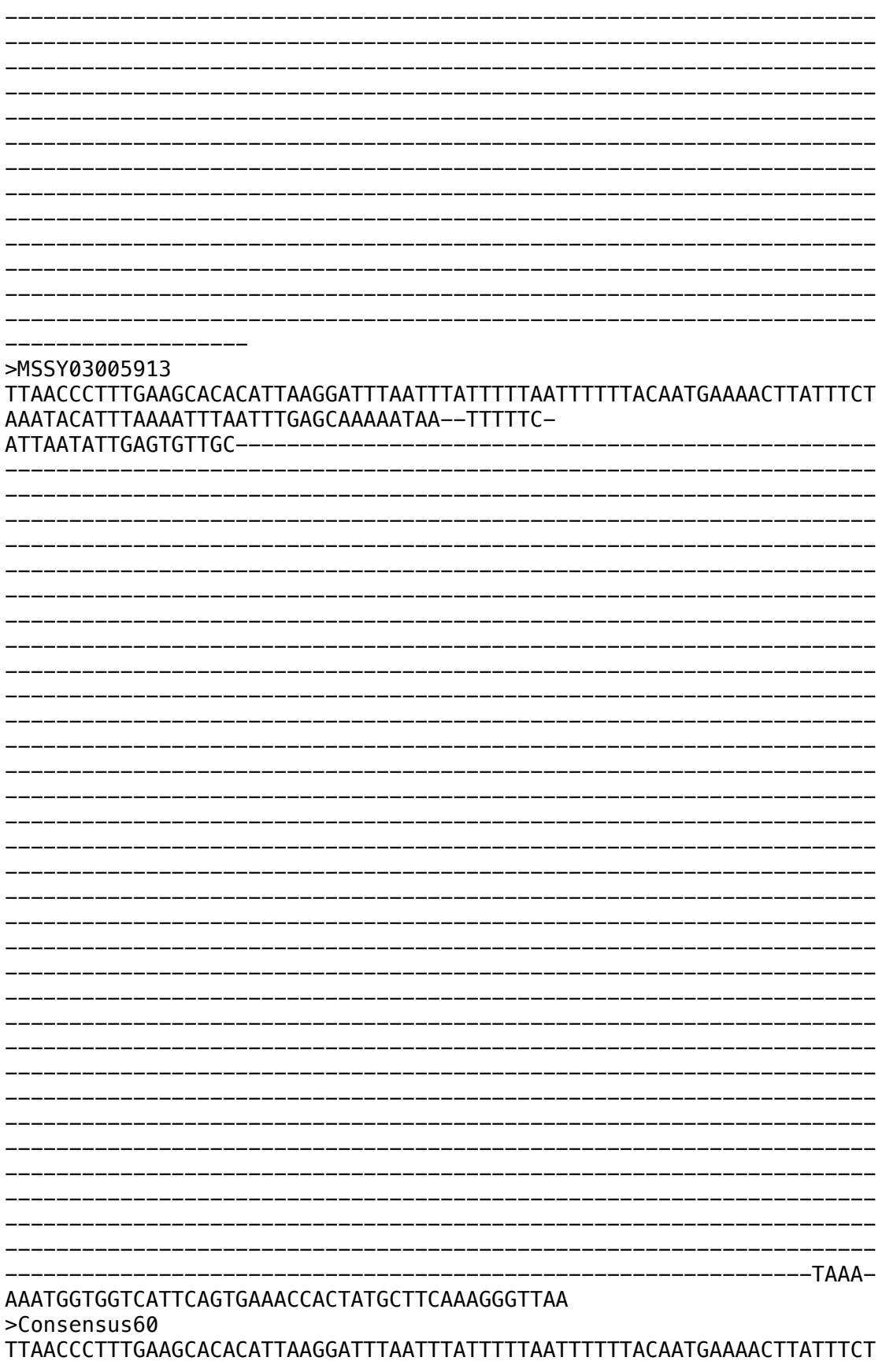
>MSSY03008721
TCAGCCCTTTGACGCACACATTAAGGA--
TAATTTATTTTTATCTTTTTACAATTAAAACTTATTTCTAAATACATTTAAAATTTCATTTTAGCAAA
--TAATATGTTT-CATTAAGATCGAGTGTTGCTAAAAAATGGTG----------
AAACCACTATGCAACTAAGCTCCACTATATTTTTACATGGTGTGGTATGCTACAAAGCAATTTTTAGA
NTCGTTGAGAAACAATGCAACTTCGCATTTCTTACAAGACCATTTACTTCGGTGTGGCTGTAACTTC-
GTGCTGCAGTGGGCACAACGTCTGCTTGAACTCC----
TTACAGGTAAATGAGCTACCTTTTCAAATCTTTTTTCAAGAGGTACTCGAGACTGAAAATAGCTCACT
GGCTTTGGAATTGGTTTTCTTCCTCTAGGTGAAGGTTTTGTTGGACTAATCAATTCATCTACTACAGA
AATACGAAATTCCTTTAGTGATAAATGGGAGCCCTCACTTTAGAGTGCATAAATGACAAAACCATTTG
TCAGTGCAACATCCACAAAGTGCCAAAATATACGGTGCCATCATTTGCGGCTTATTTC----------
-----GTAAAGGGATTTT-
AACATATCAGCTTTATCCACATAGCCCATGTGCTGGTTATAATCTTTCATTAGTTATGGACAGACTAC
CATCTGTATTATTCCATCTTTGTTTTTTCTCTTGGCTTCATAAATAAGCGAAGGATTGCGGAAATTGG
CCAAAAATGTAATACATTTGTTATCTTTCCACTTTAGCCAAGCAACTCCTGTAAATGAGGTGCGAAAC
TCACTATCACCTTTCTTCATGTTCTTTTCTACCTTTTTTTTTTCTTTTGGCAAGCCAGTTCTCACATT
CCTTACTGTTCCGCATGCATATATCGCTTGGTTCTTTAGTGAAATCATTAGCTCAACACTTG------
---

CTATGATTTTTTTTTTCATTTCTTCTTCAATTACAGGGACGAATGGTTTCCACAACTGTGTAGCATAA
AGGTTAGTTTGAAAAACTAAATGTTCAATTAAAGAATCTGGAAATAAACACAAAAAATTATCAACAGG
AGATTATTTACTTGTTGGTACAATAGGGCCAAAGTAATTAGAAAAATGTTGAATGATCCACCACCTTA

TTTTGTTTTCCCACACACTATCTA-
TTTTTTTTAATCGTTCACTAATTGGTTCATCATCATTATCATCAT------
ATTTATTACCATCTTCCTGATTTATATCAGAATAATTTAGGACATTATTGTTATTATTGAAACCGTTA
TCTTCTTGTATAAAA-
CATTTGATCTGTCCTCACTTTCCTGTAAGCAATCTTCTAAATCCGAATCACGTTC----
TCCATCAGAATCACACTGTCCATCAATAGAGTCAGTATCACTAGGCACTTCTATTAGCAAACGTTCAA
CATCTTCCTCTTTGAGTAATACGTTTAGCAATTATTTTATTTTATTTTATTTTATTTATAGACATGAC
GACACTGAAATGAGTAAATATATACACATGACCTATCTTGCACAGTGGTAGCAATTTGAAACCACCTA
GTGATTAGTTAGATTTATTTAAATATATGTATTTTAATA-----
ACCTATTAATGTAAATAATTATAAAGCAATAAACAATAGTATTAAAAGATGATTTGAATATT------
-
TACATAGTTCTATATTTATTCCTTGAAAAA----------

ACAAACTACATACTATATTATAGGCATGAATACCAACAAAAATTACTAATAATAAATTTTTACTTGAA
TAAAACTAGCAAATAAGTTATGATAGATGATTAATAAGTGCTGCCA-
ACAGTTAAATGTATTTCAAATAATTAAATGAAAGGTGGTTACTAGGGGAAACCACTATGCTTCAA-
GGTTTAA
>MSSY03005238
TTAACCCTTTGAAGCACACATTAAGGATTTAATTTATTTTTAATTTTTTACAATGAAAACTTATTTCT
AAATACATTTAAAATTTAATTTGAGCAAAAATAA--TGTTT-
CATTAATATTGAGTGTTGCTAAAAAATGGTGGTCATTCAGTGAAACCAC----------

TACAAAAATTACTAATAATAAATTTTTACAAGAATAAAACTAGCAAATAAATTCAGCTAGATGATTCA
TAAGTGCTGCCATAATGTTAAATGTTTCTCAAATAATTAAATAAAAGGTGGTTAGCAAGTGAAACCAC
TATGCTTCAAAGGGTTAA
>MSSY03000031



AAATACATTTAAAATTTAATTTGAGCAAAAATAAAATGTTTTCATTAATATTGAGTGTTGCTAAAAAA
TGGTGGTCATTCAGTGAAACCACTATGCAACTAAGCTNCACTATATTTTTACATGGTGTGGTATGCTA
CAAAGCAATTTTTAGATTCATTGAGACACAATGCCACTTCGCATGTCTTACAAGACCATTTACTTCGG
TGTGGCTGTAACTTCTGTGCTGCAGTGGGCACAACGTCTGCTTGAACTCCCTCCTTACAGGTAAATGA
GCNACCTTGTCAAATCTTTTTTCAAGAGGTACTTGAGGCTTGAAATAGCTCACTGGCTTTGGAANTGG
TTTTCTTCCTCTAGGTGAAGATTGTGTTGCACCAATCAATCCATCTACTACAGAAATACGAAATTNCT
TTAGTGATAAATGGGAGCCCTCACTTTAGAGTGCATAAATGACAAAGCCATTTGTCAGTGCAACATCC
ACAAAGTGCCAAAATATACGGTGCCACCATTTGCGGCTTATTTCTTTCCGGCTTATTTCGTANAGGGA
TTTTCAACATATCAGCTTTATCCACATAGCCCATGTGCTGGTTGTAATCTTTCACTAGTTGTGGACAG
NNTACCATCTGTATTGTTCCATCTTTGTTTTTTCTCTTGGCTTCAGAAATAAGCGAAGGATTGTGGAA
ATTGGNCAAAAATGTAATACATTTGTTGTCTTTCCACTTTAGCCAAGCAACTCCTGTAAATGANGTGN
GAAACTCACTATCACCTTTCTTCATGTTCTTTTCTACCTTTTTTTGTTCTTTTGGCAAGCCAGTTCTC
ACATTCCTTACTGTGCCGCATGCATATATCGCTTGGTTCTTTAGTGAAATCATTAGATCAACACTTGT
AAAATANTTGTCGAAATAAACATNATGACCCAANCCAAAAAGTTCAAAAGTNAGNNCTTTGNCAACTC
GAGCACCTAATAGNTTCTCAATATTGCCNTCAACCTTNCCTNTGTATATNTGAAACTGACAGANAAAC
CCCGACTNANCCGCACGCACCCACACCTTGTAGCCTCGTTTTATTGGNTTGTTGGGCATANATTGTCT
GAAAGAACTCCTTCCTTTAAATNTGATCATACTTNCATCAATGGACTGATGTTTTCCTGGTTTGAAGC
AGGTTNTGAAAGTCNCTGAAAGTGTATCAAGNATTGGTCNTAGTTTGTAGAGNTTNTCAAANGCATTG
TCNNCTCTTTTTNGCTGNAATACNTTGTTGTTTACATGTAAATGNGTCAACAACCAACCAAAACGATT
CACTGGCATNAGGGGAGAGATAAATGNATCCCTAAGCTGAGGNTNANATGACCAATGGTCCCTNTAGG
AGGGTAAACGTTTTATCCCCATCATAATGTTTATTNCTATGATTTAAATTTTCATTTCTTCTTCAGTT
ACAGGGACGAATGGTTTCCCCAACTGTGTAGCGTAAAGGTTAGTTTGAAAAACTAAATGTTCAATNAA
AGAATCTGGAAATAAACACAAAAAAGTATCGACAGGAGATTCTGTACTTGTTGGTACATTAGGGCCAA
AGTCATTTGTAAAATTTTGAATGTTCCACCACCTTATTTTGTTTTCCCACACACCATCTATTTTTTTT
AAATCGCTCACTAAGTGGTTCATCATCATCATCATCATCATCATATTTATTACNATCTTCCTGATTTA
TATCAGAATAATTTAGGACATTATTGTTATTATTGAAACCGTTATCTTCCTGTATAAAAACANTTGAT
CTGTCCTCACTTTCCTGTAAGCAATCTTCTAAATCCGAATCACNTTCACCTTCCNTCAGAATCACACT
GTCCATCAATAGAGTCAGTATCACTAGGCACTTCTANTAGCAAACGTTCAACATCTTCCTCTTTGAGT
NATACGTTTAGCAATTATTTTATTTTATTTTATTTTATTTTATAGACATGACGACACTGAAATGAGTAA
AAATATACACATGACCTTACTTGCACAGTGGTAGCAATTTGAAACCACCTAGTAATTAGTTAGATTTA
TTTAAATATATATATTTAGATAAATTTAATTATGAATGTAAATAATTATAAAGCAAAAAACAATAGTA
TTAAATAATGATTTGAATACTTAGGGCCTACATAGTTATATATTTATTCCTTGAAAAATATAACAAAA
CCAATTAGAAGAAACGTGTCACTTCAATAACAAAAACAAAATACATACTATATTGGAGGCATGAATAC
CAACAAAAATTACTAATAATAAATTTTTACAAGAATAAAACTAGCAAATAAATTCAGATAGATGATTC
ATAAGTGCTGCCATACAGTTAAATGTATCTCAAATAATTAAATAAAAGGTGGTTACCAAGTGAAACCA
CTATGCTTCAAAGGGTTAA

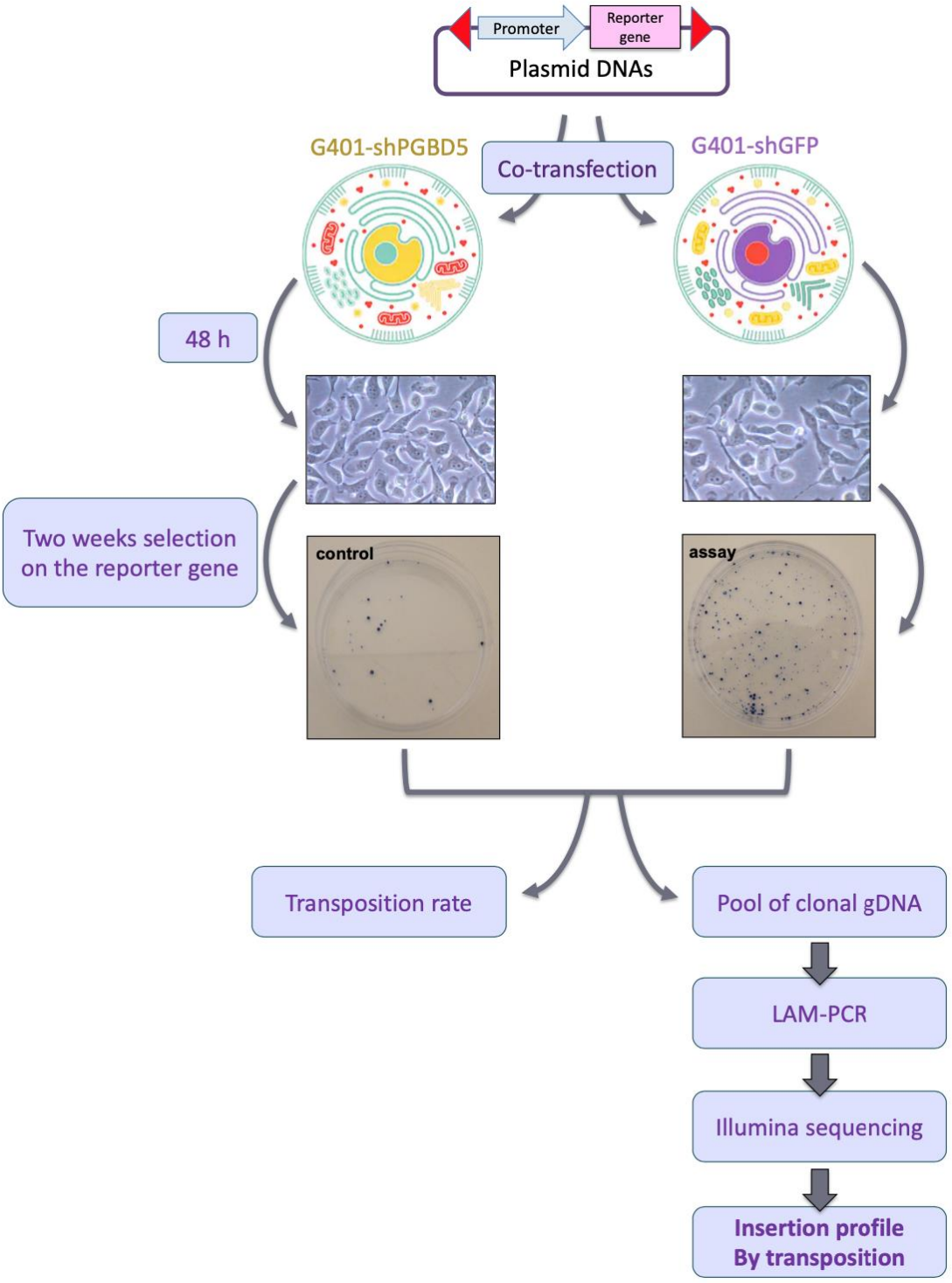


Figure 11 - Schéma récapitulatif des différentes étapes des tests de transposition dans les cellules G401.

Chapitre III. Pble humains et pgbd5

1. Transposition des *pble* humains par PGBD5

Le troisième volet des travaux dans lesquels je me suis impliquée avait pour objectif l'évaluation des capacités de PGBD5 à transposer des *pble* humains. D'après les données Repbase, il existe 5 *pble* humains : *MER75*, *MER75A*, *MER75B*, *MER85* et *Looper*. *MER75A* et *MER75B* sont des séquences consensus internes de MER75, ce sont des variants. *MER75* et *MER85* sont décrits comme des éléments non autonomes de respectivement 514 pb et 139 pb. *Looper* est décrit comme un élément autonome de 1556 pb (**Tableau 1**).

Nous avons testé les capacités de PGBD5 à transposer les *pble* humains dans deux lignées de cellules G401 transmises à notre équipe par le Dr Alex Kentsis. La fabrication de ces deux lignées a été décrite dans (Henssen *et al.*, 2017). Brièvement, il a été montré que les cellules G401 étaient capables d'exprimer l'isoforme Hs524 a un niveau suffisamment faible pour ne pas altérer leur viabilité, mais suffisant pour médier la transposition d'*Ifp2* (Henssen *et al.*, 2017). Deux clones de cellules G401 ont été produits par modification génétique pour exprimer respectivement un substrat ARN, un ARN « small hairpin » (shRNA), qui permet d'inhiber l'expression de PGBD5 (shPGBD5) ou de la GFP (shGFP). Pour le test d'intégration, les cellules G401-shPGBD5 (PGBD5-) et G401-shGFP (PGBD5+) sont transfectées avec un plasmide source de transposon. Le plasmide porteur du gène de résistance à la néomycine est flanqué en 5' et en 3' par les extrémités des transposons qui contiennent les TIR (en rouge **Figure 11**). Après incubation à l'étuve (37 °C, 5 % CO_2) pendant 48h, les cellules sont transférées en boite culture de 10 cm de diamètre pour être sélectionnées pendant deux à trois semaines avec 2 mg/mL de G418. Les clones sont ensuite colorés pour être comptés, ou récupérés pour des analyses moléculaires en population faites par LAM-PCR.

Tableau 1 - Distribution des *pble* humains dans hg38.

Pble	**Taille (pb)**	**0 TIR**	**1 TIR**	**Élément complet**	**Total**
MER75	514	354	476	182	1012
MER85	139	136	355	576	1067
Looper	1556	0	119	1243	1362

Ces données proviennent d'une réannotation effectuée avec l'outil MiteAnnot développé au laboratoire.

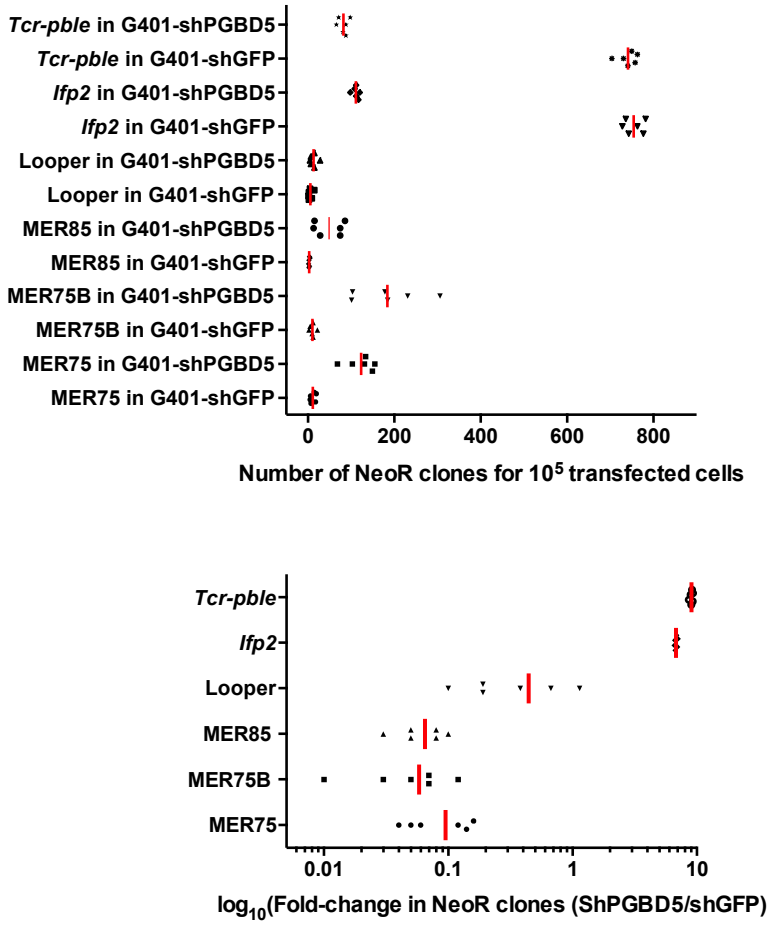


Figure 12 - Résultats de transpositions dans les cellules G401 shGFP et shPGBD5

(a) Nombre de clones NeoR obtenus après intégration chromosomique dans les cellules HeLa GFP (shPGBD5) et Mm523 (shGFP). (b) Taux d'intégration chromosomique entre les deux lignées G401 (shGFP/shPGBD5) obtenus pour une cassette NeoR avec les différents *pble*. Les lignes rouges représentent la médiane et chaque point représente la valeur de chaque réplique.

Les tests de transposition (**Figure 12**) ont été réalisés avec les deux clones de cellules G401 pour tester les deux *pble* d'invertébrés (*ifp2* et *Tcr-pble*) et quatre *pble* humains (*MER75*, *MER75B*, *MER85* et *Looper*). Le nombre de clones NeoR obtenus avec ces deux lignées cellulaires nous indique la capacité de PGBD5 à transposer les deux *pble* d'invertébrés mais à priori pas ceux de vertébrés. Cependant, nous avons observé que la quantité de clones obtenus avec les *pble* humains *MER75*, *MER75B* et *MER85* est environ 10 fois plus faible dans les cellules G401 shGFP que dans les cellules G401 shPGBD5.

Ce résultat est inattendu, car nous nous attendions à observer des nombres de clones NeoR similaires dans les deux types de cellules. Deux hypothèses permettent d'expliquer ce résultat : 1) le résultat est lié à une différence de capacité à intégrer au hasard le transposon donneur de transposon dans les deux clones cellulaires ; 2) La quantité d'ADN de plasmide donneur de transposon (500 ng d'ADN plasmidique) transfecté dans les cellules G401-shGFP est trop élevée et en présence de PGBD5 conduit à un taux de transposition trop élevé qui en conséquence est cytotoxique. Pour trancher entre ces deux hypothèses, nous avons testé la possibilité d'obtenir plus de clones NeoR dans les cellules G401-shGFP que dans les cellules G401-shPGBD5 en transfectant des quantités moindres d'ADN de plasmide donneur de transposon. Les résultats obtenus montrent que la quantité de clones diminue proportionnellement à la quantité d'ADN transfecté. C'est donc l'hypothèse 1 qui a été retenue pour expliquer nos résultats. Cette hypothèse est aussi soutenue par le fait que dans des tests d'intégration aléatoire réalisés avec un plasmide pBS-NeoR (vide de toute séquence de *pble*) une différence en nombre de clones d'un facteur 1,25 à 2 est observée entre cellules G401-shPGBD5 et G401-shGFP. Des répliques expérimentales complémentaires devront être réalisées avant de pouvoir publier ce point.

La capacité de PGBD5 à mobiliser certains *pble* soulève plusieurs questions et notamment la cible de reconnaissance de la protéine. Quel est le point commun entre ces éléments ? Nous avons donc cherché les caractéristiques communes aux *pble* pouvant servir de cibles de liaison à PGBD5. Cette partie est développée davantage dans la discussion de mon premier article. Nous avons pu mettre en évidence comme caractéristique commune la présence d'IR aux extrémités de chaque *pble*. Ces IR ont la capacité, selon les contraintes physiques subies par l'ADN (superhélicité négative) de former de structures secondaires (SII) et tertiaires (SIII) de types épingle à cheveux ou cruciformes. Les structures cruciformes peuvent être reconnues/liées par des protéines comme les protéines « High-mobility groupe » (HMG) qui sont impliquées dans les processus de régulation ADN-dépendant (transcription,

Tableau 2 - Conservation des séquences TIR (Terminal Inverted Repeat) et TSD (Target Site Duplication) des *pble* après événement de transposition par Mm523 dans les cellules HeLa.

Transposase	Tpson (NeoR)	Nombre de Jonctions	TIR & TSD conservés	TSD conservés (TIR inc.)	TIR conservés (TSD inc.)	TSD & TIR incomplets
Mm523	Ifp2	1479	9.9 % (147)	5.4 % (80)	4.3 % (64)	80.4 % (1188)
Mm523	MER75	389	11 % (43)	1.5 % (6)	3.9 % (15)	83.6 % (325)
Mm523	MER75B	307	2.3 % (7)	3.9 % (12)	1.9 % (6)	91.9 % (282)
Mm523	MER85	1313	2.6 % (34)	2.7 % (36)	0.3 % (4)	94.4 % (1239)
Mm523	Looper	193	3.6% (7)	3.1% (6)	1.6% (3)	91.7% (177)

Inc. pour incomplet.

recombinaison, etc.). Notre hypothèse est que PGBD5 serait capable de reconnaître un TSD ainsi qu'une structure cruciforme.

2. Sites d'insertion de PGBD5

Les résultats concernant les sites d'insertion de PGBD5 lors de la mobilisation de *ifp2* sont détaillés dans l'**article 1**. Pour rappel, les tests d'intégration ont été réalisés sur les cellules HeLa puis sur les cellules G401 de tumeur rhabdoïde humaine pour lesquelles l'expression endogène de PGBD5 s'est avérée avoir peu d'impact sur la viabilité des cellules (Henssen *et al.*, 2017). Lors de cette étude, nous avons observé que PGBD5 avait plus souvent tendance à déclencher une mauvaise intégration dans les cellules G401, ce qui est conforme avec l'idée selon laquelle PGBD5 présente des activités nucléasiques facilitant la recombinaison autrement que par transposition dans les tumeurs rhabdoïdes humaines (Henssen *et al.*, 2017). Des analyses de LAM-PCR ont été réalisées à partir de clones résultant de tests d'intégration réalisés avec les *pble* humains MER75, MER75B, MER85 et Looper, une source de PGBD5 Mm523 et des cellules HeLa. Les résultats des LAM-PCR (**Tableau 2, ligne 2**) révèlent que des évènements de transposition propres, ce qui indique que la transposase PGBD5 Mm523 est capable d'insérer proprement le transposon *Ifp2* dans 9,9 % des cas, avec des TIR et TSD intacts. Ces insertions propres pourraient permettre aux éléments insérés d'être mobilisés de nouveau par PGBD5. Pour le *pble* MER75, on obtient des résultats similaires à ceux obtenus avec *Ifp2*. Dans les autres cas de *pble* (MER75B, MER85 et *Looper* ; **Tableau 2, ligne 3 à 4**), seuls 2 à 3 % des éléments ont été mobilisés proprement. La grande majorité des éléments ont été insérés avec des TSD et TIR incomplets ce qui empêchera une nouvelle mobilisation par la transposase.

3. Sites de liaison de PGBD5

Les résultats précédents suggèrent que PGBD5 interagit physiquement avec les *pble* pour les transposer ou les recombiner. Des analyses ChIP-seq ont été conduites afin de vérifier si PGBD5 est capable de se lier ou non à l'ADN des *pble* présents dans les génomes de vertébrés et s'il existe des sites alternatifs de liaison. Les aspects techniques du traitement et de l'analyse des données de ChIP-seq sont développés dans la section IV.4. Ces analyses permettent d'identifier les sites de liaison de la protéine et de les situer par rapport au contexte génomique (présence dans les gènes, les promoteurs, les séquences non codantes, les *pble*). Elles permettent également d'identifier des motifs de liaison propres à PGBD5 avec des outils d'analyse de motifs tels que RSAT peak-motifs, GLAM2 et MEME-ChIP.

Tableau 3 - Résultats du ChIP-séquençage à partir de différentes lignées cellulaires.

Espèce	Tissu/Type cellulaire	Pics détectés	Co-localisant avec IR	Test de permutation (q-value)
Souris	Hippocampe Dorsal (HD)	851	99.29% (845)	3.29303e-8
	Hippocampe Ventral (HV)	873	98.97% (864)	9.544853e-7
	Cellules de Leydig	432	97.22% (420)	0.032688
Humain	G401	618	96.12% (594)	5.008677.e-7
	T98G	283	84.45% (239)	0.468
	HCT116	1306	97.93% (1279)	<1.e-100
	H4	0	-	-
	8MG	0	-	-

Nos travaux se sont portés sur la recherche de sites de liaison de PGBD5 dans deux organismes : la souris & l'homme. Chez la souris, nous avons travaillé dans les cellules de Leydig, dans l'hippocampe dorsal (HD) et le ventral (HV), des tissus dans lesquels nous avons vérifié par ailleurs que le gène *pgbd5* est transcrit. Chez l'homme, les travaux ont été réalisés sur 6 lignées cancéreuses : des cellules HeLa et HEL293 dans lesquelles le gène *pgbd5* n'est pas transcrit comme témoins négatifs, et des cellules T98G et 8MG (glioblastomes), H4 (neuroblastome) et HCT116 (cancer colorectal) dans lesquelles le gène *pgbd5* est transcrit à des niveaux variables dans les banques de données de RNA-seq disponibles. Les données précédemment obtenues en cellules G401 [reins ; (Henssen *et al.*, 2017)] ont été réanalysées.

3.1. *Souris*

Le nombre de sites de liaisons obtenus dans les trois tissus chez la souris est détaillé dans le **Tableau 3**. Après obtention des pics, aucune co-localisation de ces derniers avec *Looper*, le seul *pble* identifié dans le génome de la souris jusqu'à présent, n'a pu être mise en évidence. Des recherches de motifs conservés ont été réalisées avec RSAT peak-motifs, GLAM2 et MEME-ChIP sur les trois échantillons. Elles n'ont pas permis de mettre en évidence un (ou plusieurs) motif nucléotidique conservé de liaison.

Nous avons ensuite étudié la co-localisation des pics avec la présence de motifs structuraux particuliers, des petits palindromes (mettre ici taille des IR et des spacers) pouvant former des structures cruciformes. Les caractéristiques organisationnelles des répétitions inversées (IR) retenues pour nos analyses sont celles des IIR ou des STIR trouvés dans les régions terminales des *pble*. Nous observons qu'un grand nombre de pics (84 à 99 %) co-localisent avec de telles IR (**Tableau 3**). Des tests de permutation ont été effectués pour déterminer si l'enrichissement était dû au hasard. Les q-value des tests pour les tissus HD, HV et les cellules de Leydig (respectivement 3.29^{e}-8, 9.544^{e}-7 et 0.032) montrent que les co-localisations avec les IR ne sont pas dues au hasard ($p<0.05$). On retrouve donc un enrichissement en IR présentant des caractéristiques de STIR ou d'IIR au niveau des sites de liaison de PGBD5 dans la chromatine des tissus murins.

3.2. *Homme*

Les données de ChIP-seq obtenues avec la lignée de cellules humaines G401 sont disponibles (Henssen *et al.*, 2017). La lignée G401 est une lignée cellulaire dérivée d'une tumeur de Wilms localisée dans les cellules embryonnaires du rein. L'analyse de ces données nous a permis d'identifier 618 pics avec notre approche (**Tableau 3**). Ces 618 pics co-localisent

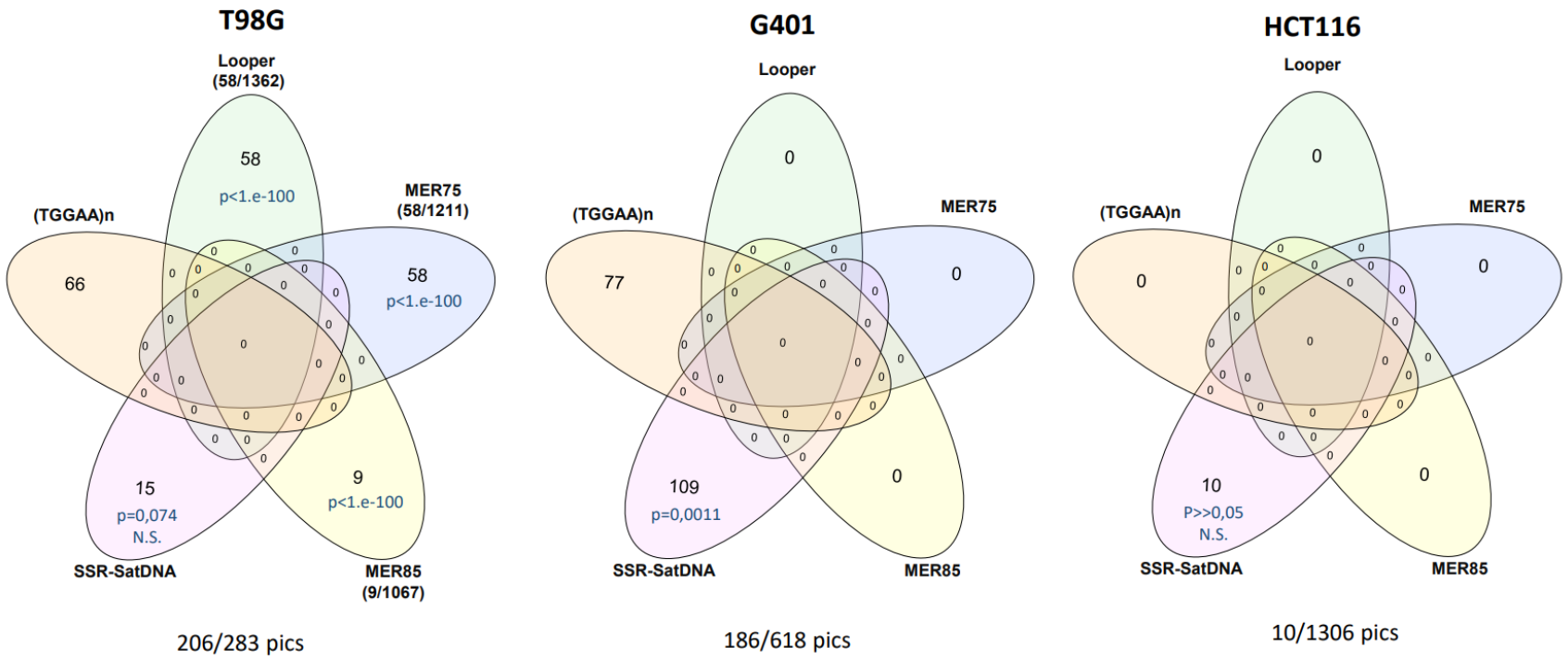


Figure 13 - Diagramme de Venn des différentes co-localisations des pics obtenus pour trois lignées cellulaires cancéreuses.

avec 100 % des 614 pics décrits dans (Henssen *et al.*, 2017), validant ainsi notre approche analytique. Une analyse des motifs révèle que 109 pics couvrent des séquences de microsatellites et 77 pics couvrent des motifs répétés en tandem (TGGAA)n (**Figure 13**). Ces motifs correspondent à de l'ADN satellite 3 humain qui est principalement localisé dans les centromères (Talbert and Henikoff, 2020). En fonction de la configuration locale de l'ADN chromosomique la séquence de ce satellite a été trouvée être capable de former des structures en épingle à cheveux sur leurs brins riches en G, ce qui suggère la possibilité que cette structure secondaire soit impliquée dans la fonction génomique (Brázda *et al.*, 2011; Kaushik *et al.*, 2016; Guiblet *et al.*, 2018). Il faut noter que les IR présentent dans les STIR et les IIR de *pble* auraient la même aptitude à former des structures en épingle à cheveux.

Aucune colocalisation des pics avec les *pble* humains (*MER75*, *MER85* et *Looper*) n'a été mise en évidence. Enfin, nous observons que plus de 96 % des pics colocalisent avec des IR. Les résultats des tests de permutation (**Tableau 3**) indiquent que la localisation des pics des cellules G401 avec les IR n'est pas due au hasard.

Nous avons procédé à la même analyse sur des données produites par l'équipe sur des cellules HeLa, HEK293, T98G, H4, HCT116 et 8MG.

Avec les deux lignées obtenues utilisées comme contrôle négatif, HeLa et HEK, car *pgbd5* qui n'est pas transcrit, aucun pic n'a été obtenu. Des résultats identiques ont été obtenus avec les lignées H4 et 8MG alors que le gène *pgbd5* y est transcrit. Deux hypothèses permettent d'expliquer ces résultats négatifs : 1) il existe une régulation traductionnelle des ARNm *pgbd5* qui prévient l'expression de PGBD5 dans ces cellules ; 2) il existe une régulation de la localisation cellulaire de PGBD5 qui prévient son transport actif dans le noyau. Des résultats préliminaires obtenus par les autres membres de l'équipe indiquent que ce serait le premier mécanisme qui serait impliqué dans les cellules H4 et 8MG.

Des pics ont par contre été obtenus avec les deux lignées cellulaires qui expriment la protéine PGBD5. Pour la lignée cancéreuse T98G, l'analyse des données de ChIP-seq a permis d'identifier 283 pics de liaison. Parmi ces derniers, 66 pics couvrent à des motifs répétés en tandem de type (TGGAA)n et 15 couvrent des séquences microsatellites. Les tests de permutation nous montrent que la co-localisation des séquences microsatellites avec des pics de liaison de la protéine PGBD5 peut être considérée comme étant due au hasard ($p=0.074$). Les pics co-localisant avec des motifs répétés en tandem de type (TGGAA)n sont de l'ADN satellite 3 qui ont aussi été trouvés colocaliser avec des pics PGBD5 dans les cellules G401. Nous avons également étudié la co-localisation des pics avec les 3 *pble* humains (MER75, MER85 & Looper). Nous avons pu identifier 58 pics colocalisant avec *MER75*, 9 pics

colocalisant avec *MER85* et 58 pics colocalisant avec *Looper*. Malgré le faible nombre de *pble* colocalisant avec les pics, il existe plus de 1000 annotations pour chacun des 3 types de *pble* au sein du génome humain (**cf**), les tests de permutation (p-value < 1.e-100) montrent que ces colocalisations ne sont pas dues au hasard. Nous avons également recherché la présence d'IR au niveau des sites de PGBD5. Mais malgré le fait que plus de 84 % des sites co-localisent avec des IR, la co-localisation peut être considérée comme étant due au hasard (q-value = 0.468). Ce résultat statistique peut être dû au faible nombre de pics obtenus, mais aussi à l'importance des pics co-localisant avec l'ADN satellite 3 qui sont capables de former des structures en épingle à cheveux à partir de motifs autres que des IR.

Des analyses similaires ont été réalisées à partir des données obtenues avec la lignée HCT116. Nous avons identifié 1306 pics. Parmi ces derniers, 10 séquences correspondent à des séquences d'ADN satellite ou microsatellites (mais pas d'ADN satellite 3). Les autres séquences ne nous permettent pas d'identifier un motif particulier et ne co-localisent pas avec les *pble* humains. Cependant, plus de 97 % des sites co-localisent avec des IR et les tests de permutation montrent que cette co-localisation n'est pas due au hasard (q-value < 1.e-100).
Le point commun principal de notre analyse ChIP-seq est que les sites cibles de PGBD5 co-localisent en très grande majorité (>95 %) avec des IR. Ces structures sont capables de former des structures tertiaires cruciformes. Cependant, le rôle de ces structures lors de la transposition et/ou de recombinaisons dépendantes de PGBD5 n'est pas encore élucidé.

Chapitre IV. Outils informatiques mis en place

Les analyses conduites durant ma thèse ont nécessité l'élaboration de scripts informatiques pour le traitement et l'analyse de données. Avant le développement d'un outil, une veille bibliographique a été conduite pour évaluer le choix d'un outil ainsi que la nécessité de développer un moyen propre à notre question.

1. Recherche de *pble*

1.1. *Démarche et sélection de candidats*

Afin de connaitre l'origine des *pgbd*, nous avons entrepris la démarche de rechercher les *pble* proches de l'élément ancêtre à l'origine de ces domestications (voir axe II. Origine de *pgbd5*). J'ai établi cette recherche sur la base d'un modèle caractéristique d'un *pble*. À savoir une homologie avec la transposase domestiquée d'intérêt et la présence de TIR à chaque extrémité de l'élément.

La recherche de *pble* proches de l'élément « ancestral » à l'origine de chaque *pgbd* a été faite en deux étapes : (1) Recherche de séquences de *pble* homologues aux *pgbd* dans les différentes espèces hôtes à l'extérieur du clade d'intérêt. Si un *pgbd* est présent seulement chez les primates, alors la recherche de l'élément ancestral s'est faite dans toutes les espèces à l'exception des primates. (2) Recherche pour chaque *pble* candidat de la présence de TIR (signature forte des éléments de la famille *piggyBac*) aux extrémités de l'élément.

Les recherches d'homologie entre la séquence protéique de chaque PGBD et les séquences nucléotidiques des *pble* candidats à identifier ou des transposases qu'ils codent, ont été réalisées en utilisant l'outil BLAST (en particulier blastp et tblastn). Chaque recherche a été lancée à distance sur le site du NCBI. Pour chaque *pgbd*, une recherche dans les bases de données protéiques et nucléotidiques a été conduite en utilisant les banques de séquences des espèces animales présentes en dehors du clade spécifique du *pgbd* utilisé comme séquence requête. La complexité de cette analyse fut notamment dans le lancement des analyses BLAST sur les invertébrés qui regroupent un très grand nombre d'espèces dans les banques de données. Il fut alors nécessaire de découper les requêtes en fonction de chaque identifiant taxonomique (Taxonomy ID) qui est disponible dans le site Taxonomy Browser (https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi) du NCBI. Les résultats ont été filtrés pour ne retenir que les homologies de séquences [*HSP* (High-scoring Segment Pair)] ayant un score de « positives » supérieur à 50 % et non pas un score d'identité supérieur à 50 %.

Le score d'identité est pour les résidus qui sont identiques entre la requête (« query ») et le sujet (« subject »), quand les deux sont parfaitement alignés. Le score de « positives » fait référence aux résidus qui sont similaires entre eux.

Chaque séquence de transposase présentant une similitude élevée avec au moins une PGBD, a été ensuite validée pour la présence d'un domaine « transposase » en utilisant l'outil Prosite du site d'Expasy (https://prosite.expasy.org/).

Afin d'identifier de potentiels *pble*, la dernière étape a été de rechercher la présence de TIR de part et d'autre de l'ORF codant la transposase. Une veille bibliographique nous a montré que malgré la richesse en nombre d'outils pour la détection de transposon à TIR, aucun ne pouvait être utilisé pour répondre à notre question. En effet, nous n'avions aucun élément de référence assez proche de nos éléments à rechercher aux vues de leurs diversités pour faire une recherche basée sur l'homologie de séquence ni aucun modèle propre à nos TIR. Il a donc fallu faire une recherche sans apriori de séquences IR. Pour répondre à cette question, j'ai créé un script permettant, à partir de séquence génomique, d'identifier des *pble* flanqués par des TIR. G2tir (genome2tir) est disponible sur github (https://github.com/Leelouh/g2tir). J'ai choisi de coder ce script en Perl, car c'est un langage qui permet de traiter facilement de l'information de type textuel.

1.2. *G2tir*

Cet outil permet d'identifier des séquences flanquées par des TIR à partir de fichiers de séquences sous format fasta. Il est nécessaire de lui donner un motif sous la forme « IUPAC ». Le script est découpé en plusieurs fonctions. Si l'on cherche un élément flanqué par des TIR, alors, une fonction va permettre de récupérer le format IUPAC et de le décliner en plusieurs motifs. Une fonction va permettre de générer une séquence inverse complémentaire qui va être cherchée à l'autre extrémité. On peut également donner au script un score, la distance de Damerau-Levenshtein qui reflète la distance entre deux chaînes de caractères. Elle représente le nombre minimum d'opérations nécessaires pour passer de la chaîne A à la chaîne B. Une opération est définie par une insertion, une suppression ou une substitution. Le module CPAN choisi pour calculer cette distance se nomme « Text:: Levenshtein : : XS ». La fonction utilisant ce module est importante, car elle permet d'utiliser le format IUPAC tout en contrôlant les substitutions. En effet, à partir du motif nucléotidique IUPAC « ARCTS » on peut décliner 5 séquences : A**A**CT**C**, A**A**CT**G**, A**G**CT**C**, A**G**CT**G**, A**A**CT**G**. Si l'on cherche à extraire de notre génome des séquences ayant des TIR identiques, alors, le score à mettre en option sera de 0.

On obtiendra alors des séquences avec les TIR déclinés dans notre fichier de résultats, mais seules les séquences aux TIR exactes entre le TIR1 et le TIR2 seront retenues. Si l'on a un motif précis, mais que l'on souhaite intégrer des substitutions au sein de ce motif, alors une fonction a été ajoutée « motifDegenere ». Une seule base à la fois sera dégénérée au sein du motif, il faudra utiliser l'option « --score » pour savoir si l'on souhaite une seule substitution entre nos deux extrémités ou deux, voire plus. Le script commence par chercher le TIR1, ensuite, par défaut, à partir de la position 43 après le TIR1 trouvé et jusqu'à 6000 bases après le TIR1, il recherche le TIR2. Pour changer ces distances, il faut utiliser les options « --minL [int] » et « --maxL [int] ». Les TIR2 candidats sont ensuite comparés et le couple est retenu ou non en fonction des options données au lancement du script. Le script génère par défaut un fichier tabulé contenant les positions des TIR1 et 2 et leurs séquences. Il est composé de plusieurs options. L'option « --fasta » permet de générer un fichier au format fasta contenant ces séquences. L'entête de chaque séquence extraite reprend le nom de la séquence d'origine, la position de début et celle de fin. L'option « --TSD [int] » permet d'afficher dans le fichier tabulé le nombre de nucléotides souhaités en amont du TIR1 et en aval du TIR2. En effet, si la séquence extraite est un candidat alors, les nucléotides en amont et en aval de cette séquence ne devraient pas être identiques. L'option « --printScore » imprime pour chaque séquence la distance de Levenshtein associée à chaque couple de TIR.

1.3. *Estimation de la contrainte sélective sur les exons d'un gène*

Un des éléments factuels qui permet de révéler la domestication d'un gène issu de transposons est la valeur de son ratio dN/dS (=ka/ks). Les substitutions nucléotidiques dans les gènes codant les protéines peuvent être synonymes ou non synonymes. Le ratio dN (nombre de substitutions non synonymes) sur dS (nombre de substitutions synonymes) permet d'estimer la « balance » entre les mutations neutres, la sélection purifiante (= sélection négative/naturelle) et les mutations bénéfiques. Un ratio supérieur à 1 révèle une sélection positive (plus de substitutions non synonymes qu'attendu par chance, la fixation par dérive génétique de mutations non synonymes devient alors faiblement délétère ou avantageuse, cela [« entraîne un changement » de la séquence]. La sélection sera purifiante si le ratio dN/dS est inférieur à 1 [« sélection contre le changement »]. Si le ratio est proche de 1, on considèrera qu'il y a une absence de pression de sélection [« pas de sélection, évolution neutre de la séquence »] (Hurst, 2002). Afin de pouvoir calculer le dN/dS des transposases domestiquées, il est nécessaire de

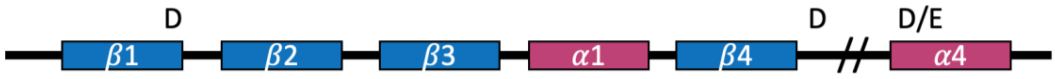


Figure 14 - Schéma du positionnement de la triade catalytique des transposases DDE en fonction des éléments de structure secondaire.

En bleu, feuillets bêta ; en rouge, hélices alpha (adapté de Yuan & Wessler, 2011).

fournir en entrée un fichier d'alignement des séquences codantes basé sur l'alignement des codons. La première étape fut d'aligner les séquences protéiques avec l'outil MAFFT. À partir de cet alignement, à l'aide de l'outil pal2nal.pl, nous avons procédé à un alignement des séquences nucléotidiques. Nous avons également construit un arbre à partir de l'alignement protéique à l'aide de RaxML (raxmlHPC-PTHREADS-SSE3, matrice PROTGAMMABLOSUM62). Cet arbre est nécessaire pour le calcul du dN/dS. Une fois les fichiers créés, il a fallu construire un fichier de configuration qui est donné en entrée à codeml, l'outil de calcul du dN/dS. Ce fichier de config nécessite donc un fichier d'alignement des séquences nucléotidiques à partir de l'alignement des protéines (seqfile), un arbre sans bootstrap dont les séquences possèdent le même identifiant que dans le fichier seqfile (treefile). Dans ce fichier, nous avons défini l'option seqtype à 1. Ainsi codeml effectue l'analyse phylogénétique avec la méthode du maximum de vraisemblance (ou Maximum Likelihood « ML ») des séquences d'ADN codant pour des protéines en utilisant des modèles de substitution de codons.

1.4. *Phylogénie des transposases de pble et des pgbd*

Dans le cas de la famille piggyBac, les séquences de transposases domestiquées et sauvages présentent des taux d'identité et de similitudes qui peuvent être faibles (~20 %). Afin de pouvoir obtenir un alignement de séquences permettant de calculer les liens de parenté, nous avons fait le choix d'aligner exclusivement le domaine issu de la transposase piggyBac (en excluant les domaines propres à chaque transposase) et de procéder à un alignement multiple prenant en compte la structure secondaire et le type de résidu. L'importance de la structure secondaire réside dans le fait que les résidus de la triade catalytique des transposases DDE sont distribués comme décrit dans le schéma de la **Figure 14**. Pour aligner les séquences, nous avons utilisé l'outil web d'alignement de séquences multiples Praline (https://www.ibi.vu.nl/programs/pralinewww/). Les jonctions entre les blocs conservés dans l'alignement ont été vérifiées et corrigées dans certains cas à la main. L'arbre phylogénétique est basé sur un alignement couvrant environ 400 résidus. L'arbre a été généré par une version Pthreads appelée raxmlHPC-PTHREADS-SSE3 version 8.2.12 de RaXML (Stamatakis, 2014). 1000 bootstraps ont été utilisés pour déduire le support des branches et nous avons utilisé le modèle de substitution d'acide aminé PROTGAMMABLOSUM62.

2. Pipeline d'analyse des sites d'insertions obtenus par LAM-PCR

Les lectures obtenues après séquençage Illumina des produits de LAM-PCR sont pairées et font une taille de 250bp. L'information pairée n'a pas été utilisée dans le processus d'analyse d'écrit ci-dessous. Après vérification de leur qualité (outil FastQC) et nettoyage (Outil TrimmomaticSE), plusieurs étapes de pré-analyses ont été faites avant la détection de variants.

- Construction de la banque BWA du plasmide sans le transposon. L'utilité de cette banque est de permettre de retirer les lectures qui s'alignent sur le plasmide et de ne conserver que celles qui s'alignent sur l'extrémité du transposon et l'ADN chromosomique qui le juxtapose.
- Alignement avec bwa mem des lectures sur la banque du plasmide
- Retrait des lectures s'alignant sur le plasmide avec samtools view [option -b (bam) -f 4]
- À partir du fichier bam obtenu, séparation des lectures non alignées avec bamToFastq en deux fichiers fasta (« reverse » et « forward »)
- Construction de la banque BWA avec bwa index contenant le génome hg38 et le transposon d'intérêt
- Alignement des lectures non alignées sur le plasmide sur cette seconde banque
- Extraction des lectures illumina s'alignant à la fois sur le transposon et l'ADN chromosomique (*split reads*) avec samtools view ainsi qu'un script de l'outil Lumpy extractSplitReads_BwaMem puis tri des fichiers bam.
- Lancement de Lumpy sur les lectures fragmentées qui génère un fichier de résultats au format vcf

Après nettoyage et analyse des données par l'outil Lumpy, nous avons extrait les informations qui nous intéressaient.

De façon générale, Lumpy est un outil de détection de variants structuraux à partir d'un alignement au format BAM produit avec l'outil d'alignement BWA. Pour la détection de ces variants, cet outil utilise deux informations principales dans l'alignement : (1) les lectures illumina fragmentées (*split reads*) au sein d'une même lecture, les parties 5' et 3' s'alignent sur des régions non contigües ; et (2) les lectures pairées discordantes qui au sein d'une même paire, vont avoir la lecture 1 alignée sur un chromosome et la lecture 2 alignée sur l'élément d'intérêt. Dans les résultats de Lumpy, tous les variants seront confondus, qu'ils soient exclusivement propres au génome (c'est à dire, des lectures discordantes mappant uniquement sur le génome) ou spécifiques d'une insertion du plasmide dans le génome. J'ai écrit un script permettant

l'extraction de ces événements d'insertion (https://github.com/Leelouh/lumpy2site/blob/master/extract_bnd_with_evid.py). Le script permet d'extraire et de récupérer toutes les preuves (lectures) propres à chaque événement d'insertion. Il prend en entrée le fichier VCF de sortie de Lumpy ainsi que le nom de la séquence de plasmide (de *pble*) utilisée pour l'analyse et présent dans le VCF. Il génère alors un fichier VCF du même format que celui de la sortie de Lumpy.

Une fois tous les événements d'intérêt récupérés avec les preuves associées, il était nécessaire d'identifier la position du site d'insertion sur le génome ainsi que la position sur l'élément inséré (insertion propre au niveau du TSD ou tronquée). J'ai écrit un script lumpy2gff.py (https://github.com/Leelouh/lumpy2site/blob/master/lumpy2gff.py). Ce script prend en entrée un fichier vcf (sortie de Lumpy) et la taille du transposon.

Pour chaque événement, il va analyser (« *parser* ») toutes les preuves associées. Il va regarder dans toutes les lectures preuves si les positions au niveau du transposon sont plus proches du début ou de la fin de l'élément (grâce à la taille donnée en entrée). Une fois cette information sauvegardée, il va chercher la preuve la plus représentative de l'élément et regarder en cas de discordance au niveau des positions, quelle ligne commence à la position 1 par exemple. Une fois la meilleure ligne sélectionnée, elle est segmentée et combinée avec la ligne de l'événement pour générer une seule ligne au format GFF qui nous donne la position de début et de fin du read aligné sur le chromosome, le sens du variant, le site (« breakpoint ») de l'insertion, l'identifiant de la lecture, la position de début et de fin d'alignement de la lecture sur le transposon, le site d'insertion au niveau du transposon (en cas d'insertion propre par exemple, le site sera en position 1 ou la taille maximale du transposon).

3. Annotation des motifs IR dans un génome

Afin de vérifier si les séquences nucléotidiques couvertes par les pics de ChIP-Seq contiennent des motifs palindromiques (ou inverted repeats – IR –) pouvant éventuellement former des structures tertiaires intra brin de type tige-boucle (« hairpin »). Deux annotations des IR sont disponibles dans les bases de données, celles sur hg19 et mm9 (Cer *et al.*, 2013). Cependant, ces annotations ne concernent que les IR n'ayant pas de mésappariements (« mismatch/MM »). Comme il est tout à fait possible d'avoir l'apparition d'une structure tige-boucle en présence d'un mésappariement, nous avons décidé de compléter l'annotation des IR dans les génomes de l'homme, de la souris et de réaliser celle du génome du poisson zèbre. D'après la littérature, il existe deux outils vraiment importants pour cette tâche : IRF et

detectIR. Le souci d'IRF est que les options disponibles ne permettent pas de répondre à notre problème, à savoir annoter les IR ayant 0 et 1 mésappariement et pouvant présenter un espacement entre les deux répétitions inversées d'une taille variable. En effet, IRF se base sur une matrice de score (un peu comme BLAST) avec une pénalité pour l'ouverture d'un gap et pour la présence d'un MM. Le calcul du score ne peut pas se faire aussi « précisément » que si on chercher réellement 0 ou 1 mésappariement. DetectIR quant à lui est un package Matlab dont je ne connais pas le langage.

La solution finalement retenue est un des deux outils EMBOSS : Einverted ou Palindrome, permettant de produire ce type d'annotation. Einverted qui est aussi basé sur une matrice de score et qui présente les mêmes défauts que IRF. Nous avons donc opté pour Palindrome qui peut répondre au cahier des charges de notre question. Cependant, certains paramètres sont à prendre en compte. Pour la formation d'une structure tige-boucle, différents cas sont possibles. Si l'on cherche une IR de taille 6 et que l'on autorise un MM de 1 alors on se retrouve avec un IR apparié sur 5 nt. Si cette IR ne possède pas d'espacement, il faudra retirer à la taille totale, 4 nucléotides afin de permettre la formation d'une boucle. Ce qui nous donnera au final, un IR de taille 3 (si le MM est au centre de l'IR, au plus proche de l'espacement). Avec un IR de taille 3, il n'est pas possible de voir apparaître une structure tige-boucle. Nous avons alors décidé de lancer Palindrome pour un motif de taille entre 5 à 15 nucléotides, avec un gap de 6 nucléotides et un nombre de MM allant de 0 à 1. Afin de pouvoir respecter les conditions précédentes, c'est-à-dire ne pas retenir les résultats suivants :

- IR de 6bp avec 1 MM
- IR de 6bp avec 0 MM mais un espacement strictement inférieur à 4
- IR de 7bp avec 0 MM mais un espacement strictement inférieur à 4
- IR de 7bp avec 1 MM mais un espacement strictement inférieur à 4

La sortie de Palindrome n'est pas standard, c'est un fichier texte avec pour chaque IR une sortie de la forme :

```
10481     agccgg      10486
          |  ||||
10501     tgggcc      10496
```

J'ai donc écrit un script en python permettant à partir de la sortie de Palindrome de vérifier le nombre de MM et de calculer la taille de l'espacement (https://github.com/Leelouh/pal2gff/blob/main/pal2gff.py). Ce script filtre et génère le résultat sous format GFF qui est un format standard d'annotation. L'analyse d'annotation d'IR suivit de

pal2gff.py a été lancé sur les génomes complets de l'homme (hg38), de la souris (mm10) et du zebrafish (dr11).

4. Définition informatique des profils de liaison à l'ADN ou à la chromatine d'une protéine par analyse de données ChIPseq

La technique ChIP-seq permet l'analyse des interactions entre une protéine et l'ADN chromosomique ou la chromatine. Les régions de la chromatine fixée qui sont spécifiquement liées par la protéine étudiée sont sélectionnées à l'aide d'un anticorps. Après une déprotéinisation puis une mise en banque Illumina, l'échantillon est soumis à un séquençage à haut débit dans l'objectif de déterminer les régions du génome qui sont le plus souvent liées par la protéine contre laquelle l'anticorps est dirigé. Après une analyse de la qualité des lectures Illumina de ChIP-Seq avec FastQ, les lectures sont nettoyées avec l'outil TrimmomaticSE. Les lectures nettoyées sont ensuite alignées contre la séquence du génome étudié avec l'outil Bowtie2. Une fois l'alignement effectué, la définition des pics de ChIP-seq a été faite avec la suite d'outils pepr. Les régions présentant un enrichissement en lectures alignées sur le génome de référence sont identifiées. Un pic est défini comme une région où le signal est enrichi par rapport au bruit de fond. PePr est un pipeline de détection de pics (« peak-calling »). Cette suite se compose de trois étapes :

(1) Pepr-preprocess : Cette fonction de pré traitement exécute uniquement l'étape d'estimation des paramètres et produit un fichier de paramètres à l'usage de l'utilisateur. Les utilisateurs peuvent utiliser/modifier le fichier de paramètres et ensuite exécuter la fonction principale de PePr avec celui-ci.

(2) PePr : Fonction principale de PePr qui permet la détection des pics.

(3) PePr-postprocess : Fonction de post-traitement de PePr qui à partir des pics PePr supprime les faux positifs suspects en utilisant les formes des pics.

Une fois les résultats PePr obtenus, seuls les pics ayant une q-value < 5.10^{-2} ont été retenus. La p-value est une mesure de la signification statistique de détection du pic. La q-value de chaque pic détecté est calculée en ajustant la p-value avec une correction de Bonferroni pour contrôler le taux de faux positifs de l'ensemble des pics détectés. Plus la q-value et la p-value sont faibles, plus il est probable que le pic détecté reflète la présence d'un site réel de liaison entre l'ADN ou la chromatine, et la protéine étudiée. Il a été ensuite possible de comparer les positions de ces pics avec différentes annotations disponibles comme celles des LAD (Lamina-

associated domains), certaines régions géniques (promoteur, UTR3, gène, intergénique), les positions des *pble*, et la présence d'IR. Ces comparaisons sont faites avec l'outil intersect de la suite bedtools. Afin de vérifier que ces sites détectés n'étaient pas dus au hasard, des tests de permutation peuvent être effectués en utilisant la suite bedtools.

Conclusion

Les ET sont des acteurs majeurs de la dynamique des génomes. Ils peuvent être impliqués dans la régulation de l'activité des gènes, mais peuvent aussi être la source de nouveaux gènes apportant de nouvelles fonctions. Les gènes des éléments transposables codant pour une transposase, une intégrase, une transcriptase inverse ainsi que les protéines de structure et d'enveloppe ont été recrutées à plusieurs reprises par leur hôte au cours de l'évolution. Mon travail de thèse a pour objet l'activité de la transposase domestiquée PGBD5 dans le génome des vertébrés ainsi que son histoire évolutive.

Le premier chapitre traite de la capacité de PGBD5 à transposer des éléments. Les séquences amino-acides des transposases de *pble* peuvent être découpé en trois régions : un domaine N-terminal (~100aa), un domaine central (~400aa) contenant une triade catalytique DD[D/E] et un domaine riche en cystéine (CRD ; ~90aa). D'après la littérature, il a été montré avec le *pble Ifp2* que le CRD de la transposase PB (IFP2) se liait au motif 5'-TGCGT-3'/3'-ACGCA-5' présent dans la STIR de l'élément *Ifp2* entre les positions 178 et 199 (Morellet et al, 2018). Il a aussi été montré que le CRD était nécessaire pour la dimérisation de PB (Sharma et al, 2018), le retrait des sept derniers résidus étant suffisant pour produire une protéine monomèrique qui se lie toujours aux extrémités *Ifp2* in vitro. Enfin, ce CRD contient un signal de localisation nucléaire (NLS) qui est nécessaire pour la médiation de la localisation nucléaire de PB (Keith et al, 2008).

Le rôle du CRD a été identifié comme essentiel pour la transposition d'*Ifp2*, car l'utilisation d'une PB tronquée de son CRD dans un essai d'intégration effectué sur des cellules de mammifères n'avait pas entraîné d'augmentation de l'activité d'intégration par rapport aux contrôles faits sans transposase (Morellet et al, 2018). Une autre étude propose que la dimérisation de PB servirait à empêcher une transposition excessive d'*Ifp2*, car la suppression du CRD conduit le PB monomère à être plus actif dans l'excision de transposition (Sharma et al, 2018). Mais il est surtout possible que l'absence du CRD provoque une rétention dans le cytoplasme car il contient le motif de localisation nucléaire (NLS) de cette protéine Keith et al, 2008). PGBD5 ne possède pas de séquence CRD à son extrémité C-terminale. Nous avons étudié l'activité de transposition de deux transposases déficientes en CRD (PB.NLS-1-558 et PGBD5). Nous avons cherché à déterminer si le PB.1-558 fusionné à un virus simien 40 (SV40) NLS était capable d'assurer la transposition de l'*Ifp2* dans les chromosomes des cellules humaines. Tout d'abord, nous avons montré que les deux PB.1-558 et PB.NLS-1-558 avaient un effet négatif sur l'obtention de clones dans les essais d'intégration. Nous avons aussi découvert que PB.NLS-1-558 était capable de réaliser la transposition des éléments *Ifp2*.

Ensuite, nous avons comparé les propriétés de PB.NLS-1-558 avec celles d'une autre protéine *piggyBac* déficiente en CRD, l'isoforme Hs524/Mm523 de PGBD5 humaine et murine. Nous avons aussi évalué la qualité des extrémités *Ifp2* néo-intégrées dans les chromosomes par PB et les deux protéines déficientes en CRD, PB.NLS-1-558 et PGBD5. Enfin, nous avons déterminé si les trois protéines utilisaient des pools de sites d'insertion chromosomiques similaires ou différents, quel que soit l'état des extrémités de transposons Ifp2. J'ai montré que PB sans son CRD et PGBD5 étaient capables de transposer l'élément *Ifp2* de manière précise/exacte et imprécise. Ces événements « imprécis » peuvent résulter d'événements de transposition incorrecte, d'intégration aléatoire ou d'événements de recombinaison dépendants de l'activité nucléasique de PB.NLS-1-558 et PGBD5. Cependant, nous avons été confrontés à des problèmes de cytotoxicité nécessitant le réglage entre la balance du niveau de protéine exprimé et des systèmes de réparation de la cellule hôte pour maintenir l'intégrité du génome. Cette balance est propre à chaque type cellulaire aux vues des variations d'activité rencontrées au sein des différentes lignées. Cette forte cytotoxicité sur les cellules utilisées confirme qu'elles possèdent bien une activité nucléasique comme cela est décrit dans les travaux de Henssen et al, 2017 (Nat Genet). Malgré cette difficulté j'ai constaté que les transposases déficientes en CRD fonctionnaient à un taux d'intégration significatif, mais inférieur à celui de la transposase « sauvage » *piggyBac*.

Lors de cette étude nous avons également pu observer que les capacités des transposases à mobiliser les *pble* produisaient des événements d'intégration qui n'étaient pas dus au hasard. En effet, bien qu'il soit difficile de montrer une spécificité stricte de sites d'insertions, nous avons démontré que ces transposases possédaient des préférences pour insérer les *pble* proche des sites d'initiation de la transcription (TSS pour *transcription start site*) et plus particulièrement proche des gènes ayant des fonctions inhérentes au système nerveux central. Des expériences supplémentaires seront nécessaires pour confirmer ce point. En effet, les gènes impliqués dans le fonctionnement du système nerveux sont, en moyenne, plus longs que les autres gènes (Zylka, Simon and Philpot, 2015). On pourrait supposer que les sites d'insertions trouvés étaient préférentiellement situés dans les gènes exprimés par les neurones à cause de leurs tailles. Cependant, notre étude tient compte de la couverture du génome ainsi que de la densité en TSD (motif TTAA) montrant que nos résultats ne sont pas liés à ces facteurs.

La fonction de l'insertion des *pble* proches des gènes ayant des fonctions inhérentes au système nerveux central reste encore mystérieuse et nécessitera une investigation. Cela pourrait résulter de l'accessibilité des gènes exprimés par les neurones ou de leur association

avec les facteurs de réparation de l'ADN et de remodelage de la chromatine qui soutiennent la transposition de l'ADN.

En conclusion, le résultat à retenir de cette étude est que le domaine CRD n'est pas nécessaire pour la transposition bien qu'il demeure nécessaire pour induire des événements propres/exacts notamment avec des insertions précises aux extrémités des *pble*.

Le **second chapitre** est découpé en deux axes : (1) retracer l'histoire évolutive des PGBD et notamment PGBD5 et (2) étudier la capacité de PGBD5 à mobiliser l'élément se trouvant à son origine.

Pour le premier axe, nous avons montré qu'il y a eu 8 à 9 événements de domestication d'éléments de la famille *piggyBac* sur une période couvrant plus de 600 millions d'années chez les vertébrés. Tous les gènes que nous avons identifiés font l'objet d'une sélection purifiante, indiquant que l'apparition de mutations est contre-sélectionnée afin de favoriser la conservation de la protéine codée par ces gènes dans la population d'hôtes. Ce type de sélection souligne aussi l'importance fonctionnelle de ces gènes pour l'hôte.

À partir des données disponibles dans les banques de données, nous avons pu mieux caractériser l'origine et l'histoire de certains gènes *pgbd*. *Pgbd3* et *pgbd4* étaient jusqu'alors considérés comme étant spécifiques à la lignée primate. Nos analyses révèlent que ces deux gènes ne sont en fait présents que dans le sous-clade des Haplorihini et sont absents chez les Strepsirhini. Cependant, nous avons identifié une domestication spécifique des Strepsirhini, celle de *pgbd8*.

Nous avons identifié aussi une domestication, *pgdb7* qui est propre aux afrothériens. Comme d'autres *pgbd* déjà caractérisés, ce gène code pour une protéine qui a la particularité d'être composé d'un domaine *piggyBac* et un autre domaine, SCAN. D'autres PGBD comme PGBD1 et PGBD3 sont composés d'autres domaines qui sont respectivement les domaines Zn-SCAN et CSB. La domestication a donc conduit à plusieurs reprises à l'acquisition de domaines fonctionnels.

Chez PGBD3 il n'y a pas de conservation stricte du domaine catalytique DDD, mais Gray et al, 2012 ont montré que les protéines CSB-PGBD3 et PGBD3 (épissage alternatif) pouvaient se lier aux éléments MER85 *in vitro*.

La conservation (ou la « cooptation ») de ces domaines dans des fusions avec des transposases *piggyBac* permet à ces nouvelles protéines d'acquérir de nouvelles capacités.

La caractérisation approfondie des *pgbd* nous a permis d'identifier que *pgbd3* et *pgbd4* étaient respectivement flanqués en 5' et en 3'par les extrémités des *pble* non autonomes MER85

et MER75. Cela suggère deux choses : (1) l'absence de TIR à lui seul n'est pas un critère pour caractériser un phénomène de domestication. (2) *pgbd3* et *pgbd4* partagent une histoire respectivement avec les *pble MER85* et *MER75*.

L'ensemble de notre analyse nous a conduits à mettre en évidence la présence d'un à plusieurs *pble* d'origine pour chaque *pgbd* domestiqué. Ces *pble* proviennent a priori essentiellement de génomes d'insectes ou même de certains tuniciers. Cela peut être expliqué par les transferts horizontaux d'éléments transposables entre certains vertébrés et leurs parasites. Par exemple, il a été montré (Gilbert, Schaack and Feschotte, 2010) qu'une punaise hématophage (Rhodnius prolixus) vectrice du trypanosome responsable de la maladie de Chagas (*Trypanosoma cruzi*) partageait une famille d'éléments transposables (*OCl*) avec le singe écureuil *(Saimiriboliviensis)* et trois *(OC1, hAT1* et *SPIN)* avec l'opossum (un marsupial) *(Monodelphis domestica)*. Ces deux espèces de mammifères étant des hôtes de cette punaise. Bien que 900 millions d'années séparent les insectes des mammifères, il a été montré que les séquences consensus des ET de la punaise, du singe et du marsupial étaient très similaires. La forte conservation de ces ET n'étant pas le résultat d'une évolution verticale sous une sélection purifiante, ils ont très probablement été transférés horizontalement entre le parasite et l'hôte durant des repas sanguins de la punaise, soit de manière directe ou par l'intermédiaire de virus transmis par la punaise (Gilbert, Schaack and Feschotte, 2010).

D'autres cas de domestication de *pgbd* devraient être mis en évidence à mesure que les technologies de séquençage et d'assemblage de génome s'améliorent. En effet, les données de certaines espèces comme certains arthropodes, tuniciers ou même céphalocordés restent peu étudiées ou du moins, peu disponibles dans les bases de données publiques.

Dans le second axe, nous avons identifié le *pble* ancêtre de *pgbd5*. Ce *pble, Tcr-pble,* provient du phasme *Timema cristinae* et possède un gène de transposase flanqué par deux TIR (ses capacités à se mobiliser lui-même n'ont pas été testées en condition *in vitro* dans le cadre de mon travail).

Étant donné que la transposase de PGBD5 semble plus proche de Tcr-pble que de PB (IFP2) (cf distance phylogénétique des transposase article 2 figure 3), nous nous attendions à ce que PGBD5 transpose plus efficacement *Tcr-pble* qu'*Ifp2*. De façon donc inattendue, j'ai montré que PGBD5 transposait avec la même efficacité *Tcr-pble* et *Ifp2*. Les isoformes Mm523 et Hs524 des PGBD5 murins et humains sont capables de stimuler l'intégration de *Ifp2* et de *Tcr-pble* dans les cellules HeLa et G401 sans différences apparentes, du moins dans nos conditions expérimentales. Ils sont capables de faire des transpositions précises, mais dans la

plupart des cas on retrouve des transpositions imprécises et des événements d'intégration par recombinaison probablement dépendante de l'activité nucléasique de PGBD5.

L'identité de leur séquence primaire est très faible, indiquant que la reconnaissance spécifique ne peut pas reposer uniquement sur la séquence primaire. Mon analyse de l'organisation des extrémités des deux *pble* m'a permis d'identifier la présence de motifs capable de former des structures secondaires non-B, à savoir des répétitions inversées dans les motifs terminaux des deux *pble*. Des études antérieures ont démontré que les régions allant de l'extrémité extérieure aux STIR aux deux extrémités d'I*fp2* étaient nécessaires pour la transposition (Solodushko, Bitko and Fouty, 2014; Troyanovsky *et al.*, 2016). Ces STIR (sub-terminales inversions répétées) et IIR (internal inversions répétées) pourraient donc être un élément important pour la liaison des transposases à leurs cibles ADN. Nous avons pu remarquer que pour les 24 *pble* étudiés, on retrouvait la même organisation aux deux extrémités. Lorsqu'une torsion de l'ADN telle qu'une super-hélicité négative est appliquée à l'ADN, une propriété intéressante des IR contenus dans les STIR et IIR est qu'ils peuvent s'extruder sous forme d'une structure tige-boucle (structure en épingle à cheveux) sur un des deux brins d'ADN, ou une structure cruciforme si les 2 brins sont impliqués d'ADN non-B l. Notre hypothèse est que PGBD5 serait capable de reconnaître ces structures et tout ou partie du reste de l'ITR. Une analyse similaire faite en LAM-PCR avec CSB-PGBD3 et MER85 montre qu'un palindrome interne imparfait de 16 pb est essentiel pour la liaison de la protéine dans les 59 pb à l'extrémité des éléments MER85 (Gray et al, 2012). Cette étude révèle des similarités entre la protéine de fusion CSB-PGBD3 et PGBD5. C'est pour l'heure, les seules transposases piggyBac de vertébrés décrites pour posséder cette caractéristique commune de liaison.

Le **troisième chapitre** concerne les capacités de PGBD5 à mobiliser les *pble* humains. À partir des tests d'intégration, nous avons observé que bien qu'il y a 10 fois plus de clones dans les cellules G401 shPGBD5 que dans les cellules G401 shGFP pour les différents *pble* humains ; MER75, MER75B, MER85 & Looper. Cela suggère que le rôle de PGBD5 n'est pas de transposer ces *pble* humains.

Outre la transposition des *pble*, nous avons déterminé les sites de liaison de la protéine PGBD5 par analyse ChIP-Seq dans plusieurs types de génomes cellulaires. Nous avons observé une faible colocalisation entre les pics et la présence d'un *pble*. Cependant, nous avons pu mettre en évidence une forte colocalisation entre les pics et la présence d'IR ayant les caractéristiques des IR présentes dans les STIR et les IIR.

Le **quatrième chapitre** concerne les outils créés ou utilisés pour effectuer mes analyses.

Les outils ainsi que les pipelines développés étaient nécessaires pour répondre aux objectifs de ma thèse. Afin de rechercher des *pble* des recherches sur les bases de données publiques ont été faites suivies de la création d'un programme pour caractériser des pble possédant certaines caractéristiques en fonction des paramètres utilisés (motif des TIR, taille du pble, conservation des TIR, etc.). Des pipelines ont été mis en place pour estimer la contrainte sélective sur les exons d'un gène ainsi que pour faire les analyses phylogénétiques. Pour les analyses de LAM-PCR j'ai pu mettre en place un pipeline ainsi que la création de deux scripts pour extraire les événements d'intérêt et les mettre en format d'annotation standard, le format gff qui est très utilisé dans l'analyse de données. J'ai ré-annoté les IR dans 3 génomes (hg38, mm10, dr11) à l'aide de l'outil Palindrome et j'ai développé un script permettant de transformer le fichier de résultat en format gff. Les outils créés ont tous été déposés sur la plateforme de partage github (https://github.com/Leelouh).

Discussion et Perspectives

Les *pble* sont-ils les vraies cibles physiologiques de PGBD5 chez les vertébrés ?

Les résultats obtenus durant ma thèse ont soulevé d'autres questions. Malgré le nombre de *pble* humains présents dans le génome, peu semblent reconnus par PGBD5. De plus, PGBD5 n'est pas capable de les mobiliser pour la transposition et la recombinaison avec la même efficacité que les *pble* d'insectes testés. La raison peut être que ce ne sont pas les cibles premières de PGBD5. En effet, on retrouve plus de 4000 *pble* dans le génome de l'homme. Si PGBD5 était capable de tous les mobiliser sans contrainte, cela présenterait très probablement un risque pour la cellule hôte.

Le cas de RAG1/RAG2 indique que lors de la domestication d'une transposase, ses cibles le furent également. Pour PGBD5, il est actuellement difficile de vérifier une telle hypothèse. Si PGBD5 est très conservé au sein des vertébrés, il n'en est pas de même pour les *pble* dont les « espèces » varient d'une espèce à l'autre. Chez l'humain, on retrouve trois « espèces » de *pble* (*MER75*, *MER85* et *Looper*). Chez la souris, une seule « espèce » de *pble*, *Looper*, semblable à celui de l'homme est annotée. Chez le poisson-zèbre, il y a sept « espèces » *pble* qui ont une origine différente aux *pble* humains/murins. Chez le poulet, aucun *pble* n'a été répertorié jusqu'à présent. L'absence de *pble* dans certains génomes (malgré la présence de PGBD5) et le manque d'efficacité de PGBD5 à mobiliser les *pble* humains suggèrent que la protéine se lie à d'autres cibles de liaison ADN.

Cependant, les propriétés des cibles ADN de liaison pour PGBD5 doivent se rapprocher en séquence ou en structure de certains éléments contenus dans la séquence des extrémités des *pble*. En effet, nous avons pu mettre en évidence que parmi les 24 *pble* étudiés dans ma thèse, tous comportent des IR à proximité des extrémités. Notre hypothèse est que PGBD5 est capable de reconnaître ces motifs en tant que cible de liaison.

Jusqu'où un mécanisme de recombinaison programmée impliquant PGBD5 pourrait-il être domestiqué à partir d'éléments dérivés d'ET ?

Dans la continuité de mes travaux, il serait donc intéressant de mener d'autres analyses afin de vérifier les capacités des IR à passer d'une configuration d'ADN B à une configuration d'ADN non-B. Ceci permettrait de voir si c'est bien la formation d'une structure cruciforme qui est reconnue par la transposase PGBD5 au moment d'initier la transposition ou la recombinaison.

Le ciblage des sites de liaison de PGBD5 ne dépend pas seulement de la présence de ses cibles, mais aussi de l'accessibilité de celle-ci dans la chromatine. En effet, avec une transposase domestiquée dans un génome, il peut être nécessaire que son expression soit contrôlée ou que

ses cibles ne soient pas disponibles tout le temps afin que les événements de recombinaison qu'elle catalyse soient contrôlés par l'hôte. C'est ce qu'on retrouve notamment avec les transposases domestiquées RAG1 et RAG2. En condition physiologique, ces dernières ne sont exprimées qu'au cours de la lymphogénèse ; ce qui leur permet de procéder aux recombinaisons des domaines V et D, puis D et J des gènes codant les IgH et les IgL. Ces deux recombinaisons successives sont contrôlées par les réarrangements chromatiniens qui ouvrent et ferment la chromatine pour diriger l'accès à l'ADN (Fugmann, 2010; Huang *et al.*, 2016).

Il pourrait également être intéressant de tester la transposition des autres *pble* identifiés dans notre étude avec les autres transposases PGBD présent chez l'homme, en présence ou non de PGBD5. En effet, PGBD1 à 5 sont présentes chez l'homme et forme ainsi un système complexe de transposases *piggyBac* qui pourraient interagir les unes avec les autres. Il pourrait donc être intéressant de voir si elles jouent un rôle seul ou en combinaison avec d'autres protéines dans la mobilisation des *pble* présents dans un génome. Cela permettrait de vérifier si tout ou partie des PGBD s'associent entre elles ou avec d'autres protéines pour former des complexes comme ceux retrouvés chez les protozoaires ciliés lors de la différenciation du micronucleus en macronucleus. Par exemple, on sait que PGBD3 et PGBD4 sont flanqués par les extrémités respectives de MER85 et MER75. Il serait intéressant de voir si ces 2 protéines sont capables de mobiliser ces éléments non autonomes et si elles sont toujours capables de se mobiliser les gènes qui les codent.

Comment et avec quels types de moyens pourrait-on approfondir l'analyse des sites de liaison de PGBD5 dans la chromatine ?

Les données de ChIP-Seq ont été traitées en utilisant des outils spécifiques pour la recherche de motifs d'intérêt. Les analyses restent compliquées à cause des biais et des artéfacts locaux pouvant être introduits lors du séquençage, de l'alignement, la structure de la chromatine ou encore par les variations du nombre de copies d'un gène. Dans notre cas, nous avons utilisé des logiciels de détection des régions enrichies. Les logiciels font donc appel à des algorithmes qui vont identifier les motifs les plus présents au sein de la liste des séquences (MEME). Mais il ne s'agit pas de détecter un motif précis, mais plutôt un pattern moyen. D'autres outils permettent de comparer les séquences vis-à-vis de base de données de sites de liaison de facteur de transcription déjà existants (ex. : JASPAR). Les limites principales de ces logiciels sont qu'ils ont été faits pour identifier le facteur de transcription principal du jeu de données. Cependant, certains facteurs de transcription possèdent de nombreux cofacteurs et les sites de liaison fonctionnels ont tendance à former des groupes chez les eucaryotes, des régulations en

cis. Il existe d'autres méthodes/pipelines sous Bioconductor qui cherchent à limiter les biais rencontrés. Aux vues du nombre de pics obtenus, nos signaux sont peut-être fondus dans la masse.

Afin de procéder à une analyse plus poussée sur les extrémités des *pble* il serait intéressant de mettre au point un outil utilisant des méthodes de *machine learning* pour générer un modèle représentatif de nos données. Ce modèle aurait pu être étendu à nos pics de données ChIP-seq. En effet, plusieurs études récentes proposent de nouveaux outils comme par exemple l'outil CNN-peaks (Oh *et al.*, 2020) qui est un pipeline de détection des pics ChIP-Seq utilisant des réseaux neuronaux convolutifs qui imitent l'inspection visuelle humaine. Utiliser de tels outils permettrait d'apporter une autre approche pour l'identification de motifs de liaison de PGBD5.

PGBD5 est une recombinase domestiquée probablement pour faire des réarrangements génomiques programmés et spécifiques dans certaines cellules saines. Est-ce que les cellules cancéreuses sont véritablement un bon modèle d'étude ?

Le rôle de PGBD5 n'est pas tout à fait élucidé. Bien que PGBD5 est fortement exprimé dans le système nerveux des mammifères (Pavelitz *et al.*, 2013) et que les bases de données actuelles soutiennent cette conclusion (https://www.proteinatlas.org/search/PGBD5), il manque cependant des données concernant la traduction des ARNm *pgbd5* en protéine. Cependant, si l'acquisition au cours de l'évolution d'un mécanisme permettant de déclencher des réarrangements programmés et irréversibles de l'ADN de certaines cellules du système nerveux central peut être considérée comme un avantage pour les espèces qui le possèdent, alors PGBD5 semble être le candidat idéal pour jouer un tel rôle. Le mécanisme de PGBD5 est probablement régulé au cours du temps et son fonctionnement se fait sans doute en interagissant avec d'autres protéines de l'hôte.

Une des limites de nos études concerne les cellules utilisées. Les analyses de transposition et de ChIP-Seq ont été réalisées dans deux lignées de cellules dérivées de cancer, HeLa et G401. Les cellules cancéreuses sont des cellules ayant subi des lésions de l'ADN avec des échecs de réparation, c'est-à-dire des mutations constitutionnelles (héréditaires) dans les gènes de réparation de l'ADN, les gènes contrôlant la prolifération cellulaire et l'apoptose. Une cellule cancéreuse possède alors des mécanismes corrompus de développement, de maintien, de prolifération, de signalisation et de dialogue avec le système immunitaire (Cieślik and Chinnaiyan, 2018; Chatsirisupachai *et al.*, 2019). À ces caractéristiques s'ajoute l'impact de la

culture in vitro qui ajoute davantage de dérives dans les mécanismes de fonctionnement ainsi que dans le phénotype des cellules par rapport à leur tumeur d'origine.

Il est donc légitime de se demander si ce que l'on observe en termes de transposition, de recombinaison et de reconnaissance de l'ADN dans les cellules cancéreuses reflète exactement ce qu'il se passe dans les cellules saines. J'ai évoqué plus tôt l'intérêt d'étudier les interactions entre PGBD5 et d'autres protéines comme ce qui a été montré pour la protéine de fusion CSB-PGBD3. Les co-facteurs protéiques nécessaires à une action enzymatique correcte de PGBD5 peuvent être absents ou présents à des concentrations très différentes modifiant alors totalement les relations allostériques de PGBD5 avec des co-facteurs potentiels et donc leurs spécificités d'action.

Lors de nos études, nous avons relevé que la transposition propre est le type d'événement le moins fréquent (**tableau 2**). Ce résultat peut être dû au fonctionnement spécifique au sein des cellules cancéreuses. Nous n'avons pas identifié de motif précis lors de nos analyses ChIP-Seq, cela peut également être dû aux lignées de cellules. En effet, selon les lignées de cellules et les tissus, nous avons observé de grandes différences. Pour rappel, nous n'avions trouvé aucun pic pour les lignées de glioblastomes H4 et 8MG, 238 pics pour T98G et plus de mille pics pour HCT116 alors que dans les tissus sains murins, nous avons détecté 400 à 800 pics selon les tissus étudiés.

Afin de pallier à cette diversité de résultats, il serait intéressant de faire des tests de transposition *in vivo* dans les tissus étudiés comme l'hippocampe et les cellules de Leydig avec comme source de transposon *Ifp2* ou d'autres *pble* puis d'analyser les événements avec les procédures de LAM-PCR ainsi que les analyses *in silico* mises au point durant ma thèse.

Nous connaissons les variations en termes de *pble* au sein des différentes espèces. Il serait intéressant de développer du ChIP-seq dans plusieurs tissus et espèces animales pour étudier les sites de liaison à l'ADN et identifier les points communs au sein des espèces.

Le futur du projet autour de PGBD serait donc d'étudier les interactions de PGBD5 directement dans les cellules qui sont son siège d'action et non plus dans les cellules cancéreuses qui ont néanmoins été un moyen expérimental pratique et nécessaire pour « débroussailler » le projet.

Références Bibliographiques

A Smit, A. F., Riggs, A. D. and Olson, M. V (1996) 'Tiggers and other DNA transposon fossils in the human genome (interspersed repeats/pogo/mariner/Tcl/centromere protein CENP-B)', 93(February), pp. 1443–1448. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC39958/pdf/pnas01508-0090.pdf%0Ahttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC39958/.

Alzohairy, A. M. *et al.* (2013) 'Transposable elements domesticated and neofunctionalized by eukaryotic genomes', *Plasmid*, 69(1), pp. 1–15. doi: 10.1016/j.plasmid.2012.08.001.

Anxolabéhère, D., Kidwell, M. G. and Periquet, G. (1988) 'Molecular characteristics of diverse populations are consistent with the hypothesis of a recent invasion of Drosophila melanogaster by mobile P elements.', *Molecular Biology and Evolution*. doi: 10.1093/oxfordjournals.molbev.a040491.

Arensburger, P., Piégu, B. and Bigot, Y. (2016) 'The future of transposable element annotation and their classification in the light of functional genomics - what we can learn from the fables of Jean de la Fontaine?', *Mobile Genetic Elements*, 6(6), p. e1256852. doi: 10.1080/2159256X.2016.1256852.

Arkhipova, I. R. (2017) 'Using bioinformatic and phylogenetic approaches to classify transposable elements and understand their complex evolutionary histories', pp. 1–14. doi: 10.1186/s13100-017-0103-2.

Aronson, A. *et al.* (1960) 'Biophysics', *Carnegie Institution of Washington*, 59, pp. 229–279.

Bailey, A. D. *et al.* (2012) 'The conserved Cockayne syndrome B-piggyBac fusion protein (CSB-PGBD3) affects DNA repair and induces both interferon-like and innate antiviral responses in CSB-null cells', *DNA Repair*, 11(5), pp. 488–501. doi: 10.1016/j.dnarep.2012.02.004.

Barau, J. *et al.* (2016) 'The DNA methyltransferase DNMT3C protects male germ cells from transposon activity', *Science*, 354(6314), pp. 909–912. doi: 10.1126/science.aah5143.

Baudry, C. *et al.* (2009) 'PiggyMac, a domesticated piggyBac transposase involved in programmed genome rearrangements in the ciliate Paramecium tetraurelia', *Genes and Development*, 23(21), pp. 2478–2483. doi: 10.1101/gad.547309.

Belancio, V. P., Hedges, D. J. and Deininger, P. (2008) 'Mammalian non-LTR retrotransposons: for better or worse, in sickness and in health.', *Genome research*, 18(3), pp. 343–58. doi: 10.1101/gr.5558208.

Biémont, C. *et al.* (1994) 'Population dynamics of the copia, mdg1, mdg3, gypsy, and P transposable elements in a natural population of Drosophila melanogaster', *Genetical Research*, 63(3), pp. 197–212. doi: 10.1017/S0016672300032353.

Biémont, C. *et al.* (1997) 'Transposable element distribution in Drosophila.', *Genetics*, 147(4), pp. 1997–9. doi: 10.1016/0091-7435(75)90040-7.

Biémont, C. and Vieira, C. (2006) 'Junk DNA as an evolutionary force', *Nature*, 443(7111), pp. 521–524. doi: 10.1038/443521a.

Bird, A. (2002) 'DNA methylation patterns and epigenetic memory', *Genes & Development*, 16(1), pp. 6–21. doi: 10.1101/gad.947102.

Bire, S. *et al.* (2016) 'Mariner Transposons Contain a Silencer: Possible Role of the Polycomb Repressive Complex 2', *PLoS Genetics*, 12(3), pp. 1–38. doi: 10.1371/journal.pgen.1005902.

Blumenstiel, J. P. (2019) 'Birth, School, Work, Death, and Resurrection: The Life Stages and

Dynamics of Transposable Element Proliferation', *Genes*, 10(5), p. 336. doi: 10.3390/genes10050336.

Bohne, A. *et al.* (2012) 'Zisupton--A Novel Superfamily of DNA Transposable Elements Recently Active in Fish', *Molecular Biology and Evolution*, 29(2), pp. 631–645. doi: 10.1093/molbev/msr208.

Bouallègue, M. *et al.* (2017) 'Molecular Evolution of piggyBac Superfamily: From Selfishness to Domestication', *Genome Biology and Evolution*, 9(2), pp. 323–339. doi: 10.1093/gbe/evw292.

Brandt, J. *et al.* (2005) 'Transposable elements as a source of genetic innovation: expression and evolution of a family of retrotransposon-derived neogenes in mammals', *Gene*, 345(1), pp. 101–111. doi: 10.1016/j.gene.2004.11.022.

Brázda, V. *et al.* (2011) 'Cruciform structures are a common DNA feature important for regulating biological processes', *BMC Molecular Biology*, 12(1), p. 33. doi: 10.1186/1471-2199-12-33.

Britten, R. J. (1996) 'DNA sequence insertion and evolutionary variation in gene regulation.', *Proceedings of the National Academy of Sciences*, 93(18), pp. 9374–9377. doi: 10.1073/pnas.93.18.9374.

Brosius, J. (1991) 'Retroposons--seeds of evolution', *Science*, 251(4995), pp. 753–753. doi: 10.1126/science.1990437.

Brosius, J. (1999) 'Genomes were forged by massive bombardments with retroelements and retrosequences.', *Genetica*, 107(1–3), pp. 209–38. Available at: http://www.ncbi.nlm.nih.gov/pubmed/10952214.

Bureau, T. E. and Wessler, S. R. (1994) 'Stowaway: a new family of inverted repeat elements associated with the genes of both monocotyledonous and dicotyledonous plants.', *The Plant Cell*, 6(6), pp. 907–916. doi: 10.1105/tpc.6.6.907.

Campillos, M. *et al.* (2006) 'Computational characterization of multiple Gag-like human proteins', *Trends in Genetics*, 22(11), pp. 585–589. doi: 10.1016/j.tig.2006.09.006.

Capy, P. *et al.* (1998) *Dynamics and Evolution of Transposable Elements*.

Capy, P., David, J. R. and Hartl, D. L. (1992) 'Evolution of the transposable element mariner in the Drosophila melanogaster species group', *Genetica*, 86(1–3), pp. 37–46. doi: 10.1007/BF00133709.

Carpes, M. P. *et al.* (2009) 'Molecular analysis of a mutant Anticarsia gemmatalis multiple nucleopolyhedrovirus (AgMNPV) shows an interruption of an inhibitor of apoptosis gene (iap-3) by a new class-II piggyBac -related insect transposon', *Insect Molecular Biology*, 18(6), pp. 747–757. doi: 10.1111/j.1365-2583.2009.00917.x.

Cary, L. C. *et al.* (1989) 'Transposon mutagenesis of baculoviruses: Analysis of Trichoplusia ni transposon IFP2 insertions within the FP-locus of nuclear polyhedrosis viruses', *Virology*, 172(1), pp. 156–169. doi: 10.1016/0042-6822(89)90117-7.

Cer, R. Z. *et al.* (2013) 'Non-B DB v2 . 0 : a database of predicted non-B DNA-forming motifs and its associated tools', 41(November 2012), pp. 94–100. doi: 10.1093/nar/gks955.

Chandler, M. *et al.* (2013) 'Breaking and joining single-stranded DNA: the HUH endonuclease superfamily', *Nature Reviews Microbiology*, 11(8), pp. 525–538. doi: 10.1038/nrmicro3067.

Chandler, V. L. and Walbot, V. (1986) ‘DNA modification of a maize transposable element correlates with loss of activity.’, *Proceedings of the National Academy of Sciences*, 83(6), pp. 1767–1771. doi: 10.1073/pnas.83.6.1767.

Charlier, D., Piette, J. and Glansdorff, N. (1982) ‘IS3 can function as a mobile promoter in E. coli.’, *Nucleic acids research*, 10(19), pp. 5935–48. doi: 10.1093/nar/10.19.5935.

Chatsirisupachai, K. *et al.* (2019) ‘A human tissue-specific transcriptomic analysis reveals a complex relationship between aging, cancer, and cellular senescence’, *Aging Cell*, 18(6), pp. 1–5. doi: 10.1111/acel.13041.

Chen, Q. *et al.* (2020) ‘Structural basis of seamless excision and specific targeting by piggyBac transposase’, *Nature Communications*, 11(1), pp. 1–14. doi: 10.1038/s41467-020-17128-1.

Chénais, B. *et al.* (2012) ‘The impact of transposable elements on eukaryotic genomes: From genome size increase to genetic adaptation to stressful environments’, *Gene*, 509(1), pp. 7–15. doi: 10.1016/j.gene.2012.07.042.

Cheng, C.-Y. *et al.* (2010) ‘A Domesticated piggyBac Transposase Plays Key Roles in Heterochromatin Dynamics and DNA Cleavage during Programmed DNA Deletion in Tetrahymena thermophila’, *Molecular Biology of the Cell*. Edited by K. S. Bloom, 21(10), pp. 1753–1762. doi: 10.1091/mbc.e09-12-1079.

Cheng, C. Y. *et al.* (2016) ‘The piggyBac transposon-derived genes TPB1 and TPB6 mediate essential transposon-like excision during the developmental rearrangement of key genes in Tetrahymena thermophila’, *Genes and Development*, 30(24), pp. 2724–2736. doi: 10.1101/gad.290460.116.

Chomet, P. S., Wessler, S. and Dellaporta, S. L. (1987) ‘Inactivation of the maize transposable element Activator (Ac) is associated with its DNA modification.’, *The EMBO journal*, 6(2), pp. 295–302. Available at: http://www.ncbi.nlm.nih.gov/pubmed/3034583.

Cieślik, M. and Chinnaiyan, A. M. (2018) ‘Cancer transcriptome profiling at the juncture of clinical translation’, *Nature Reviews Genetics*, 19(2), pp. 93–109. doi: 10.1038/nrg.2017.96.

Cordaux, R. *et al.* (2006) ‘Birth of a chimeric primate gene by capture of the transposase gene from a mobile element’, *Proceedings of the National Academy of Sciences*, 103(21), pp. 8101–8106. doi: 10.1073/pnas.0601161103.

Cordaux, R. and Batzer, M. A. (2009) ‘The impact of retrotransposons on human genome evolution’, *Nature Reviews Genetics*, 10(10), pp. 691–703. doi: 10.1038/nrg2640.

Cosby, R. L., Chang, N. C. and Feschotte, C. (2019) ‘Host–transposon interactions: Conflict, cooperation, and cooption’, *Genes and Development*, 33(17–18), pp. 1098–1116. doi: 10.1101/gad.327312.119.

Daimon, T. *et al.* (2010) ‘Recent transposition of yabusame , a novel piggyBac -like transposable element in the genome of the silkworm, Bombyx mori’, *Genome*. Edited by T. Bureau, 53(8), pp. 585–593. doi: 10.1139/G10-035.

Davies, D. R. *et al.* (2000) ‘Three-Dimensional Structure of the Tn5 Synaptic Complex Transposition Intermediate’, *Science*, 289(5476), pp. 77–85. doi: 10.1126/science.289.5476.77.

Delsuc, F. *et al.* (2006) ‘Tunicates and not cephalochordates are the closest living relatives of vertebrates’, *Nature*, 439(7079), pp. 965–968. doi: 10.1016/j.tet.2010.01.110.

Dewannieux, M., Esnault, C. and Heidmann, T. (2003) 'LINE-mediated retrotransposition of marked Alu sequences', *Nature Genetics*, 35(1), pp. 41–48. doi: 10.1038/ng1223.

Ehret, C. F. and De Haller, G. (1963) 'Origin, development, and maturation of organelles and organelle systems of the cell surface in Paramecium', *Journal of Ultrastructure Research*, 9, pp. 1–42. doi: 10.1016/S0022-5320(63)80088-X.

Eickbush, T. H. and Malik, H. S. (2002) 'Origins and Evolution of Retrotransposons', in *Mobile DNA II*. American Society of Microbiology, pp. 1111–1144. doi: 10.1128/9781555817954.ch49.

Elick, T. A. (1996) 'Excision of the piggy Bac transposable element in vitro is a precise event that is enhanced by the expression of its encoded transposase', *Genetica*, 98(1), pp. 33–41. doi: 10.1007/BF00120216.

Elick, T. A., Lobo, N. and Fraser, M. J. (1997) 'Analysis of the cis-acting DNA elements required for piggyBac transposable element excision', *Molecular and General Genetics*, 255(6), pp. 605–610. doi: 10.1007/s004380050534.

Evgen'ev, M. B. and Arkhipova, I. R. (2005) 'Penelope-like elements – a new class of retroelements: distribution, function and possible evolutionary significance', *Cytogenetic and Genome Research*, 110(1–4), pp. 510–521. doi: 10.1159/000084984.

Finnegan, D. J. (1989) 'Eukaryotic transposable elements and genome evolution', *Trends in Genetics*, 5, pp. 103–107. doi: 10.1016/0168-9525(89)90039-5.

Fraser, M. J. *et al.* (1985) 'Transposon-mediated mutagenesis of a baculovirus', *Virology*, 145(2), pp. 356–361. doi: 10.1016/0042-6822(85)90172-2.

Fraser, M. J., Smith, G. E. and Summers, M. D. (1983) 'Acquisition of Host Cell DNA Sequences by Baculoviruses: Relationship Between Host DNA Insertions and FP Mutants of Autographa californica and Galleria mellonella Nuclear Polyhedrosis Viruses.', *Journal of virology*, 47(2), pp. 287–300. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16789244.

Fugmann, S. D. (2010) 'The origins of the Rag genes—From transposition to V(D)J recombination', *Seminars in Immunology*, 22(1), pp. 10–16. doi: 10.1016/j.smim.2009.11.004.

Gilbert, C. and Feschotte, C. (2018) 'Horizontal acquisition of transposable elements and viral sequences: patterns and consequences', *Current Opinion in Genetics and Development*, 49(October 2017), pp. 15–24. doi: 10.1016/j.gde.2018.02.007.

Gilbert, C., Schaack, S. and Feschotte, C. (2010) 'Quand les éléments génétiques mobiles bondissent entre espèces animales', *médecine/sciences*, 26(12), pp. 1025–1027. doi: 10.1051/medsci/201026121025.

Le Goff, G. and Hilliou, F. (2017) 'Resistance evolution in Drosophila : the case of CYP6G1', *Pest Management Science*, 73(3), pp. 493–499. doi: 10.1002/ps.4470.

Goodier, J. L. (2016) 'Restricting retrotransposons: a review', *Mobile DNA*, 7(1), p. 16. doi: 10.1186/s13100-016-0070-z.

Goodwin, T. J. D. and Poulter, R. T. M. (2001) 'The DIRS1 Group of Retrotransposons', *Molecular Biology and Evolution*, 18(11), pp. 2067–2082. doi: 10.1093/oxfordjournals.molbev.a003748.

Gray, L. T. *et al.* (2012) 'Tethering of the Conserved piggyBac Transposase Fusion Protein CSB-PGBD3 to Chromosomal AP-1 Proteins Regulates Expression of Nearby Genes in

Humans', *PLoS Genetics*, 8(9). doi: 10.1371/journal.pgen.1002972.

Guiblet, W. M. *et al.* (2018) 'Long-read sequencing technology indicates genome-wide effects of non-B DNA on polymerization speed and error rate', *Genome Research*, 28(12), pp. 1767–1778. doi: 10.1101/gr.241257.118.

Guio, L., Barrón, M. G. and González, J. (2014) 'The transposable element Bari-Jheh mediates oxidative stress response in Drosophila', *Molecular Ecology*, 23(8), pp. 2020–2030. doi: 10.1111/mec.12711.

Hancks, D. C. and Kazazian, H. H. (2010) 'SVA retrotransposons: Evolution and genetic instability', *Seminars in Cancer Biology*, 20(4), pp. 234–245. doi: 10.1016/j.semcancer.2010.04.001.

Hancks, D. C. and Kazazian, H. H. (2012) 'Active human retrotransposons: variation and disease', *Current Opinion in Genetics & Development*, 22(3), pp. 191–203. doi: 10.1016/j.gde.2012.02.006.

Handler, A. M. and McCombs, S. D. (2000) 'The piggyBac transposon mediates germ-line transformation in the Oriental fruit fly and closely related elements exist in its genome', *Insect Molecular Biology*, 9(6), pp. 605–612. doi: 10.1046/j.1365-2583.2000.00227.x.

Henssen, A. G. *et al.* (2015) 'Genomic DNA transposition induced by human PGBD5', *eLife*, 4(September), pp. 1–20. doi: 10.7554/eLife.10565.

Henssen, A. G. *et al.* (2016) 'Forward genetic screen of human transposase genomic rearrangements', *BMC Genomics*, 17(1), p. 548. doi: 10.1186/s12864-016-2877-x.

Henssen, A. G. *et al.* (2017) 'PGBD5 promotes site-specific oncogenic mutations in human tumors', *Nature Genetics*, 49(7), pp. 1005–1014. doi: 10.1038/ng.3866.

Hikosaka, A. *et al.* (2007) 'Evolution of the Xenopus piggyBac Transposon Family TxpB: Domesticated and Untamed Strategies of Transposon Subfamilies', *Molecular Biology and Evolution*, 24(12), pp. 2648–2656. doi: 10.1093/molbev/msm191.

Hood, M. E. (2005) 'Repetitive DNA in the automictic fungus Microbotryum violaceum', *Genetica*, 124(1), pp. 1–10. doi: 10.1007/s10709-004-6615-y.

Horvath, R. and Slotte, T. (2017) 'The Role of Small RNA-Based Epigenetic Silencing for Purifying Selection on Transposable Elements in Capsella grandiflora', *Genome Biology and Evolution*, 9(10), pp. 2911–2920. doi: 10.1093/gbe/evx206.

Huang, S. *et al.* (2016) 'Discovery of an Active RAG Transposon Illuminates the Origins of V(D)J Recombination', *Cell*, 166(1), pp. 102–114. doi: 10.1016/j.cell.2016.05.032.

Hurst, L. D. (2002) 'The Ka/Ks ratio: diagnosing the form of sequence evolution', *Trends in Genetics*, 18(9), pp. 486–487. doi: 10.1016/S0168-9525(02)02722-1.

Hwei-gene Heidi Wang, Fraser, M. J. and Cary, L. C. (1989) 'Transposon mutagenesis of baculoviruses: analysis of TFP 3 lepidopteran transposon insertions at the FP locus of nuclear polyhedrosis viruses', *Gene*, 81(1), pp. 97–108. doi: 10.1016/0378-1119(89)90340-5.

Ivancevic, A. M. *et al.* (2016) 'LINEs between Species: Evolutionary Dynamics of LINE-1 Retrotransposons across the Eukaryotic Tree of Life', *Genome Biology and Evolution*, 8(11), pp. 3301–3322. doi: 10.1093/gbe/evw243.

Jaurin, B. and Normark, S. (1983) 'Insertion of IS2 creates a novel ampC promoter in Escherichia coli', *Cell*, 32(3), pp. 809–816. doi: 10.1016/0092-8674(83)90067-3.

Joly-Lopez, Z. *et al.* (2016) 'Phylogenetic and Genomic Analyses Resolve the Origin of Important Plant Genes Derived from Transposable Elements', *Molecular Biology and Evolution*, 33(8), pp. 1937–1956. doi: 10.1093/molbev/msw067.

Jurka, J. (1998) 'Repeats in genomic DNA: mining and meaning', *Current Opinion in Structural Biology*, 8(3), pp. 333–337. doi: 10.1016/S0959-440X(98)80067-5.

Jurka, J. *et al.* (2005) 'Repbase Update, a database of eukaryotic repetitive elements', *Cytogenetic and Genome Research*, 110(1–4), pp. 462–467. doi: 10.1159/000084979.

Kalendar, R. *et al.* (2011) 'Analysis of plant diversity with retrotransposon-based molecular markers', *Heredity*, 106(4), pp. 520–530. doi: 10.1038/hdy.2010.93.

Kapitonov, V. V. and Jurka, J. (2001) 'Rolling-circle transposons in eukaryotes', *Proceedings of the National Academy of Sciences*, 98(15), pp. 8714–8719. doi: 10.1073/pnas.151269298.

Kapitonov, V. V. and Jurka, J. (2004) 'Harbinger Transposons and an Ancient HARBI1 Gene Derived from a Transposase', *DNA and Cell Biology*, 23(5), pp. 311–324. doi: 10.1089/104454904323090949.

Kapitonov, V. V. and Jurka, J. (2005) 'RAG1 core and V(D)J recombination signal sequences were derived from Transib transposons', *PLoS Biology*, 3(6), pp. 998–1011. doi: 10.1371/journal.pbio.0030181.

Kapitonov, V. V. and Jurka, J. (2006) 'Self-synthesizing DNA transposons in eukaryotes', *Proceedings of the National Academy of Sciences*, 103(12), pp. 4540–4545. doi: 10.1073/pnas.0600833103.

Kapitonov, V. V. and Jurka, J. (2008) 'A universal classification of eukaryotic transposable elements implemented in Repbase', *Nature Reviews Genetics*, 9(5), pp. 411–412. doi: 10.1038/nrg2165-c1.

Kaushik, M. *et al.* (2016) 'A bouquet of DNA structures: Emerging diversity', *Biochemistry and Biophysics Reports*, 5, pp. 388–395. doi: 10.1016/j.bbrep.2016.01.013.

Kazazian, H. H. (2004) 'Mobile Elements: Drivers of Genome Evolution', *Science*, 303(5664), pp. 1626–1632. doi: 10.1126/science.1089670.

Keith, J. H., Fraser, T. S. and Fraser, M. J. (2008) 'Analysis of the piggyBac transposase reveals a functional nuclear targeting signal in the 94 c-terminal residues', *BMC Molecular Biology*, 9, pp. 1–13. doi: 10.1186/1471-2199-9-72.

Kidwell, M. (1992) 'Horizontal transfer', *Current Opinion in Genetics & Development*, 2(6), pp. 868–873. doi: 10.1016/S0959-437X(05)80109-1.

Kidwell, M. G. (2002) 'Transposable elements and the evolution of genome size in eukaryotes.', *Genetica*, 115(1), pp. 49–63. doi: 10.1023/a:1016072014259.

Kidwell, M. G. and Lisch, D. (1997) 'Transposable elements as sources of variation in animals and plants', *Proceedings of the National Academy of Sciences*, 94(15), pp. 7704–7711. doi: 10.1073/pnas.94.15.7704.

Kidwell, M. G. and Lisch, D. R. (2001) 'PERSPECTIVE: TRANSPOSABLE ELEMENTS, PARASITIC DNA, AND GENOME EVOLUTION', *Evolution*, 55(1), pp. 1–24. doi: 10.1111/j.0014-3820.2001.tb01268.x.

Kimura, K. and Kidwell, M. G. (1994) 'Differences in P element population dynamics between the sibling species Drosophila melanogaster and Drosophila simulans', *Genetical*

Research, 63(1), pp. 27–38. doi: 10.1017/S0016672300032055.

Kojima, K. K. (2018) 'Human transposable elements in Repbase: genomic footprints from fish to humans', pp. 1–14. doi: 10.1186/s13100-017-0107-y.

Kojima, K. K. and Jurka, J. (2011) 'Crypton transposons: Identification of new diverse families and ancient domestication events', *Mobile DNA*, 2(1), pp. 1–17. doi: 10.1186/1759-8753-2-12.

Kokošar, J. and Kordiš, D. (2013) 'Genesis and regulatory wiring of retroelement-derived domesticated genes: A phylogenomic perspective', *Molecular Biology and Evolution*, 30(5), pp. 1015–1031. doi: 10.1093/molbev/mst014.

Kramerov, D. and Vassetzky, N. (2005) 'Short Retroposons in Eukaryotic Genomes', *International Review of Cytology*, 247, pp. 165–221. doi: 10.1016/S0074-7696(05)47004-7.

Krupovic, M. *et al.* (2014) 'Casposons: A new superfamily of self-synthesizing DNA transposons at the origin of prokaryotic CRISPR-Cas immunity', *BMC Biology*, 12(1), pp. 1–12. doi: 10.1186/1741-7007-12-36.

Krupovic, M. and Forterre, P. (2015) 'Single-stranded DNA viruses employ a variety of mechanisms for integration into host genomes', *Annals of the New York Academy of Sciences*, 1341(1), pp. 41–53. doi: 10.1111/nyas.12675.

Krupovic, M. and Koonin, E. V. (2015) 'Polintons: A hotbed of eukaryotic virus, transposon and plasmid evolution', *Nature Reviews Microbiology*, 13(2), pp. 105–115. doi: 10.1038/nrmicro3389.

Krupovic, M., Yutin, N. and Koonin, E. V. (2016) 'Fusion of a superfamily 1 helicase and an inactivated DNA polymerase is a signature of common evolutionary history of Polintons, polinton-like viruses, Tlr1 transposons and transpovirons', *Virus Evolution*, 2(1), pp. 1–9. doi: 10.1093/ve/vew019.

de la Peña, M. and Cervera, A. (2017) 'Circular RNAs with hammerhead ribozymes encoded in eukaryotic genomes: The enemy at home', *RNA Biology*, 14(8), pp. 985–991. doi: 10.1080/15476286.2017.1321730.

Lander, E. S. *et al.* (2001) 'Initial sequencing and analysis of the human genome', *Nature*, 409(6822), pp. 860–921. doi: 10.1038/35057062.

Leib-Mösch, C. *et al.* (1993) 'Genomic Distribution and Transcription of Solitary HERV-K LTRs', *Genomics*, 18(2), pp. 261–269. doi: 10.1006/geno.1993.1464.

Li, M. A. *et al.* (2011) 'Mobilization of giant piggyBac transposons in the mouse genome', *Nucleic Acids Research*, 39(22), pp. 1–9. doi: 10.1093/nar/gkr764.

Liu, C., Yang, Y. and Schatz, D. G. (2019) 'Structures of a RAG-like transposase during cut-and-paste transposition', *Nature*, 575(7783), pp. 540–544. doi: 10.1038/s41586-019-1753-7.

Lloréns, C. and Marín, I. (2001) 'A Mammalian Gene Evolved from the Integrase Domain of an LTR Retrotransposon', *Molecular Biology and Evolution*, 18(8), pp. 1597–1600. doi: 10.1093/oxfordjournals.molbev.a003947.

Lohe, A. R. *et al.* (1995) 'Horizontal transmission, vertical inactivation, and stochastic loss of mariner-like transposable elements.', *Molecular Biology and Evolution*, 12(1), pp. 62–72. doi: 10.1093/oxfordjournals.molbev.a040191.

Lohe, A. R., De Aguiar, D. and Hartl, D. L. (1997) 'Mutations in the mariner transposase: The

D,D(35)E consensus sequence is nonfunctional', *Proceedings of the National Academy of Sciences*, 94(4), pp. 1293–1297. doi: 10.1073/pnas.94.4.1293.

Lohe, A. R. and Hartl, D. L. (1996) 'Germline transformation of Drosophila virilis with the transposable element mariner.', *Genetics*, 143(1), pp. 365–74. Available at: http://www.ncbi.nlm.nih.gov/pubmed/8722788.

Luo, G.-H. H. *et al.* (2014) 'Molecular characterization of the piggyBac-like element, a candidate marker for phylogenetic research of Chilo suppressalis (Walker) in China', *BMC Molecular Biology*, 15(1), p. 28. doi: 10.1186/s12867-014-0028-y.

Makałowski, W., Mitchell, G. A. and Labuda, D. (1994) 'Alu sequences in the coding regions of mRNA: a source of protein variability', *Trends in Genetics*, 10(6), pp. 188–193. doi: 10.1016/0168-9525(94)90254-2.

Malamy, M. H., Fiandt, M. and Szybalski, W. (1972) 'Electron microscopy of polar insertions in the lac operon of Escherichia coli', *MGG Molecular & General Genetics*, 119(3), pp. 207–222. doi: 10.1007/BF00333859.

Malik, H. S., Henikoff, S. and Eickbush, T. H. (2000) 'Poised for Contagion: Evolutionary Origins of the Infectious Abilities of Invertebrate Retroviruses', *Genome Research*, 10(9), pp. 1307–1318. doi: 10.1101/gr.145000.

Mateo, L., Ullastres, A. and González, J. (2014) 'A Transposable Element Insertion Confers Xenobiotic Resistance in Drosophila', *PLoS Genetics*. Edited by C. Feschotte, 10(8), p. e1004560. doi: 10.1371/journal.pgen.1004560.

McClintock, B. (1950) 'The origin and behavior of mutable loci in maize', *Proceedings of the National Academy of Sciences*, 36(6), pp. 344–355. doi: 10.1073/pnas.36.6.344.

Miller, W. J. *et al.* (1999) 'Molecular domestication--more than a sporadic episode in evolution.', *Genetica*, 107(1–3), pp. 197–207. Available at: http://www.ncbi.nlm.nih.gov/pubmed/10952213.

Mitra, R. D. R. D. *et al.* (2013) 'Functional characterization of piggyBat from the bat Myotis lucifugus unveils an activemammalian DNA transposon', *Proccedings of the National Academy of Sciences*, 110(1), pp. 1–6. doi: 10.1073/pnas.1217548110/-/DCSupplemental.www.pnas.org/cgi/doi/10.1073/pnas.1217548110.

Mitra, R., Fain-Thornton, J. and Craig, N. L. (2008) 'piggyBac can bypass DNA synthesis during cut and paste transposition', *The EMBO Journal*, 27(7), pp. 1097–1109. doi: 10.1038/emboj.2008.41.

Montaño, S. P., Pigli, Y. Z. and Rice, P. A. (2012) 'The Mu transpososome structure sheds light on DDE recombinase evolution', *Nature*, 491(7424), pp. 413–417. doi: 10.1038/nature11602.

Morellet, N. *et al.* (2018) 'Sequence-specific DNA binding activity of the cross-brace zinc finger motif of the piggyBac transposase', *Nucleic Acids Research*, 46(5), pp. 2660–2677. doi: 10.1093/nar/gky044.

Nassif, N. *et al.* (1994) 'Efficient copying of nonhomologous sequences from ectopic sites via P-element-induced gap repair.', *Molecular and Cellular Biology*, 14(3), pp. 1613–1625. doi: 10.1128/MCB.14.3.1613.

Newman, J. C. *et al.* (2008) 'An Abundant Evolutionarily Conserved CSB-PiggyBac Fusion Protein Expressed in Cockayne Syndrome', *PLoS Genetics*. Edited by D. A. Petrov, 4(3), p.

e1000031. doi: 10.1371/journal.pgen.1000031.

Nuzhdin, S. V (1999) ‘Sure facts, speculations, and open questions about the evolution of transposable element copy number.’, *Genetica*, 107(1–3), pp. 129–37. Available at: http://www.ncbi.nlm.nih.gov/pubmed/10952206.

Oh, D. *et al.* (2020) ‘CNN-Peaks: ChIP-Seq peak detection pipeline using convolutional neural networks that imitate human visual inspection’, *Scientific Reports*, 10(1), pp. 1–12. doi: 10.1038/s41598-020-64655-4.

Ohno, S. (1972) ‘So much “junk” DNA in our genome.’, *Brookhaven symposia in biology*, 23(366), pp. 366–70. Available at: http://www.ncbi.nlm.nih.gov/pubmed/5065367.

Ono, R. *et al.* (2001) ‘A Retrotransposon-Derived Gene, PEG10, Is a Novel Imprinted Gene Located on Human Chromosome 7q21’, *Genomics*, 73(2), pp. 232–237. doi: 10.1006/geno.2001.6494.

de Parseval, N. and Heidmann, T. (2005) ‘Human endogenous retroviruses: from infectious elements to human genes’, *Cytogenetic and Genome Research*, 110(1–4), pp. 318–332. doi: 10.1159/000084964.

Pavelitz, T. *et al.* (2013) ‘PGBD5: a neural-specific intron-containing piggyBac transposase domesticated over 500 million years ago and conserved from cephalochordates to humans’, *Mobile DNA*, 4(1), p. 23. doi: 10.1186/1759-8753-4-23.

Penton, E. H., Sullender, B. W. and Crease, T. J. (2002) ‘Pokey, a New DNA Transposon in Daphnia (Cladocera: Crustacea)’, *Journal of Molecular Evolution*, 55(6), pp. 664–673. doi: 10.1007/s00239-002-2362-9.

Piégu, B. *et al.* (2015) ‘A survey of transposable element classification systems - A call for a fundamental update to meet the challenge of their diversity and complexity’, *Molecular Phylogenetics and Evolution*, 86, pp. 90–109. doi: 10.1016/j.ympev.2015.03.009.

Pritham, E. J. and Feschotte, C. (2007) ‘Massive amplification of rolling-circle transposons in the lineage of the bat Myotis lucifugus’, *Proceedings of the National Academy of Sciences*, 104(6), pp. 1895–1900. doi: 10.1073/pnas.0609601104.

Pritham, E. J., Putliwala, T. and Feschotte, C. (2007) ‘Mavericks, a novel class of giant transposable elements widespread in eukaryotes and related to DNA viruses’, *Gene*, 390(1–2), pp. 3–17. doi: 10.1016/j.gene.2006.08.008.

Proença, J. T., Barral, D. C. and Gordo, I. (2017) ‘Commensal-to-pathogen transition: One-single transposon insertion results in two pathoadaptive traits in Escherichia coli -macrophage interaction’, *Scientific Reports*, 7(1), p. 4504. doi: 10.1038/s41598-017-04081-1.

Prudhomme, S., Bonnaud, B. and Mallet, F. (2005) ‘Endogenous retroviruses and animal reproduction’, *Cytogenetic and Genome Research*, 110(1–4), pp. 353–364. doi: 10.1159/000084967.

Putnam, N. H. *et al.* (2008) ‘The amphioxus genome and the evolution of the chordate karyotype’, *Nature*, 453(7198), pp. 1064–1071. doi: 10.1038/nature06967.

Richardson, J. M. *et al.* (2009) ‘Molecular Architecture of the Mos1 Paired-End Complex: The Structural Basis of DNA Transposition in a Eukaryote’, *Cell*, 138(6), pp. 1096–1108. doi: 10.1016/j.cell.2009.07.012.

Rostant, W. G., Wedell, N. and Hosken, D. J. (2012) ‘Transposable Elements and Insecticide Resistance’, in, pp. 169–201. doi: 10.1016/B978-0-12-394394-1.00002-X.

Le Rouzic, A., Boutin, T. S. and Capy, P. (2007) 'Long-term evolution of transposable elements', *Proceedings of the National Academy of Sciences*, 104(49), pp. 19375–19380. doi: 10.1073/pnas.0705238104.

Le Rouzic, A. and Capy, P. (2005) 'The First Steps of Transposable Elements Invasion', *Genetics*, 169(2), pp. 1033–1043. doi: 10.1534/genetics.104.031211.

Rowold, D. J. and Herrera, R. J. (2000) 'Alu elements and the human genome.', *Genetica*, 108(1), pp. 57–72. doi: 10.1023/a:1004099605261.

Saedler, H. *et al.* (1974) 'IS2, a genetic element for turn-off and turn-on of gene activity in E. coli.', *Molecular & general genetics : MGG*, 132(4), pp. 265–89. doi: 10.1007/BF00268569.

Sánchez-Gracia, A., Maside, X. and Charlesworth, B. (2005) 'High rate of horizontal transfer of transposable elements in Drosophila', *Trends in Genetics*, 21(4), pp. 200–203. doi: 10.1016/j.tig.2005.02.001.

Sarkar, A. *et al.* (2003) 'Molecular evolutionary analysis of the widespread piggyBac transposon family and related "domesticated" sequences', *Molecular Genetics and Genomics*, 270(2), pp. 173–180. doi: 10.1007/s00438-003-0909-0.

Schrader, L. and Schmitz, J. (2018) 'The impact of transposable elements in adaptive evolution', *Molecular Ecology*, (July). doi: 10.1111/mec.14794.

Siguier, P. *et al.* (2006) 'ISfinder: the reference centre for bacterial insertion sequences', *Nucleic Acids Research*, 34(90001), pp. D32–D36. doi: 10.1093/nar/gkj014.

Singer, M. F. (1982) 'SINEs and LINEs: Highly repeated short and long interspersed sequences in mammalian genomes', *Cell*, 28(3), pp. 433–434. doi: 10.1016/0092-8674(82)90194-5.

Sinzelle, L., Izsvák, Z. and Ivics, Z. (2009) 'Molecular domestication of transposable elements: From detrimental parasites to useful host genes', *Cellular and Molecular Life Sciences*, 66(6), pp. 1073–1093. doi: 10.1007/s00018-009-8376-3.

Slotkin, R. K., Freeling, M. and Lisch, D. (2003) 'Mu killer causes the heritable inactivation of the Mutator family of transposable elements in Zea mays.', *Genetics*, 165(2), pp. 781–97. Available at: http://www.ncbi.nlm.nih.gov/pubmed/14573488.

Solodushko, V., Bitko, V. and Fouty, B. (2014) 'Minimal piggyBac vectors for chromatin integration', *Gene Therapy*, 21(1), pp. 1–9. doi: 10.1038/gt.2013.52.

Sookdeo, A., Hepp, C. M. and Boissinot, S. (2018) 'Contrasted patterns of evolution of the LINE-1 retrotransposon in perissodactyls: the history of a LINE-1 extinction', *Mobile DNA*, 9(1), p. 12. doi: 10.1186/s13100-018-0117-4.

Stamatakis, A. (2014) 'RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies', *Bioinformatics*, 30(9), pp. 1312–1313. doi: 10.1093/bioinformatics/btu033.

Sun, Z. C. *et al.* (2008) 'piggyBac-like elements in cotton bollworm, Helicoverpa armigera (Hübner)', *Insect Molecular Biology*, 17(1), pp. 9–18. doi: 10.1111/j.1365-2583.2008.00780.x.

Talbert, P. B. and Henikoff, S. (2020) 'What makes a centromere?', *Experimental Cell Research*, 389(2), p. 111895. doi: 10.1016/j.yexcr.2020.111895.

Tellier, M. and Chalmers, R. (2019) 'Human SETMAR is a DNA sequence-specific histone-

methylase with a broad effect on the transcriptome', *Nucleic Acids Research*, 47(1), pp. 122–133. doi: 10.1093/nar/gky937.

Troyanovsky, B. *et al.* (2016) 'The functionality of minimal PiggyBac transposons in mammalian cells', *Molecular Therapy - Nucleic Acids*, 5(10), p. e369. doi: 10.1038/mtna.2016.76.

Van't Hof, A. E. *et al.* (2016) 'The industrial melanism mutation in British peppered moths is a transposable element.', *Nature*, 534(7605), pp. 102–5. doi: 10.1038/nature17951.

Villesen, P. *et al.* (2004) 'Identification of endogenous retroviral reading frames in the human genome.', *Retrovirology*, 1, p. 32. doi: 10.1186/1742-4690-1-32.

Volff, J. N. (2006) 'Turning junk into gold: Domestication of transposable elements and the creation of new genes in eukaryotes', *BioEssays*, 28(9), pp. 913–922. doi: 10.1002/bies.20452.

Wang, J. *et al.* (2006) 'piggyBac-like elements in the tobacco budworm, Heliothis virescens (Fabricius)', *Insect Molecular Biology*, 15(4), pp. 435–443. doi: 10.1111/j.1365-2583.2006.00653.x.

Wang, J. *et al.* (2010) 'PiggyBac -like elements in the pink bollworm, Pectinophora gossypiella', *Insect Molecular Biology*, 19(2), pp. 177–184. doi: 10.1111/j.1365-2583.2009.00964.x.

Wang, W. *et al.* (2008) 'Chromosomal transposition of PiggyBac in mouse embryonic stem cells.', *Proceedings of the National Academy of Sciences*, 105(27), pp. 9290–5. doi: 10.1073/pnas.0801017105.

Wicker, T. *et al.* (2003) 'CACTA Transposons in Triticeae. A Diverse Family of High-Copy Repetitive Elements', *Plant Physiology*, 132(1), pp. 52–63. doi: 10.1104/pp.102.015743.

Wicker, T. *et al.* (2007) 'A unified classification system for eukaryotic transposable elements', *Nature Reviews Genetics*, 8(12), pp. 973–982. doi: 10.1038/nrg2165.

Wilson, M. H., Coates, C. J. and George, A. L. (2007) 'PiggyBac Transposon-mediated Gene Transfer in Human Cells', *Molecular Therapy*, 15(1), pp. 139–145. doi: 10.1038/sj.mt.6300028.

Wu, C. and Wang, S. (2014) 'PLE-wu, a new member of piggyBac transposon family from insect, is active in mammalian cells', *Journal of Bioscience and Bioengineering*, 118(4), pp. 359–366. doi: 10.1016/j.jbiosc.2014.03.010.

Wu, M. *et al.* (2008) 'An active piggyBac -like element in Macdunnoughia crassisigna', *Insect Science*, 15(6), pp. 521–528. doi: 10.1111/j.1744-7917.2008.00241.x.

Xu, H.-F. *et al.* (2006) 'Identification and characterization of piggyBac-like elements in the genome of domesticated silkworm, Bombyx mori', *Molecular Genetics and Genomics*, 276(1), pp. 31–40. doi: 10.1007/s00438-006-0124-x.

Yoder, J. A., Walsh, C. P. and Bestor, T. H. (1997) 'Cytosine methylation and the ecology of intragenomic parasites', *Trends in Genetics*, 13(8), pp. 335–340. doi: 10.1016/S0168-9525(97)01181-5.

Yoshida, J. *et al.* (2017) 'Chromatin states shape insertion profiles of the piggyBac, Tol2 and Sleeping Beauty transposons and murine leukemia virus', *Scientific Reports*, 7(1), p. 43613. doi: 10.1038/srep43613.

Yusa, K. *et al.* (2011) ‘A hyperactive piggyBac transposase for mammalian applications’, *Proceedings of the National Academy of Sciences*, 108(4), pp. 1531–1536. doi: 10.1073/pnas.1008322108.

Yusa, K. (2014) ‘piggyBac Transposon’, in *Mobile DNA III*. American Society of Microbiology, pp. 875–892. doi: 10.1128/microbiolspec.MDNA3-0028-2014.

Zylka, M. J., Simon, J. M. and Philpot, B. D. (2015) ‘Gene Length Matters in Neurons’, *Neuron*, 86(2), pp. 353–355. doi: 10.1016/j.neuron.2015.03.059.

Laura HELOU

Définition des cibles chromosomiques de PGBD5 dans les neurones pyramidaux du système nerveux central des vertébrés

Résumé

Les éléments transposables sont présents dans presque tous les génomes eucaryotes et procaryotes. Ils co-évoluent avec leurs hôtes et ont participé au façonnage de leur environnement génomique au cours de leur évolution. Ils sont en outre à l'origine de nouveaux gènes dérivés de séquences codantes de transposon par domestication. Le gène qui code la transposase domestiquée piggyBac, PGBD5, est très conservé depuis l'origine des vertébrés. Cette transposase fonctionnelle est capable de transposer un élément piggyBac d'insecte, *Ifp2*. Elle s'exprime spécifiquement dans le système nerveux central. La conservation élevée de PGBD5 chez les vertébrés suggère que sa fonction et ses cibles chromosomiques le seraient également. Au cours de ma thèse, nous avons étudié les capacités de transposition d'une transposase piggyBac sans le domaine CRD (PGBD5 ayant perdu son domaine CRD au cours de la domestication). Nous avons vérifié la nécessité de la présence d'un domaine C-terminal pour l'activité de transposase-recombinase. Puis nous avons étudier l'histoire évolutive de *pgbd5* et des *pgbd* en général. Nous avons identifié l'élément phylogénétiquement le plus proche de *pgbd5* et ses capacités à être transposé par PGBD5. Enfin, nous avons étudié les capacités de PGBD5 à exciser et réintégrer dans de nouveaux locus des *pble* humains afin d'évaluer la capacité de PGBD5 à mobiliser ces éléments présents dans le génome humain et d'affiner l'identification des cibles de la transposase.

Mots-clés : PGBD5, piggyBac, transposase

Résumé en anglais

Transposable elements are present in almost all eukaryotic and prokaryotic genomes. They co-evolve with their hosts and have participated in the shaping of their genomic environment during their evolution. In addition, they are at the origin of new genes derived from transposon coding sequences by domestication. The gene that encodes the domesticated piggyBac transposase, PGBD5, has been highly conserved since the origin of vertebrates. This transposase (specifically expressed in the central nervous system) is functional and capable of transposing an insect piggyBac element, *Ifp2*. The high conservation of PGBD5 in vertebrates suggests that its function and chromosomal targets would also be highly conserved. My thesis work aims to study the origin and activity of the domesticated PGBD5 transposase in the vertebrate genome. First, we have studied the transposal capacities of a piggyBac transposase without the CRD domain. Indeed, PGBD5 can be characterized as a transposase having lost its CRD domain during domestication. We have verified the necessity of the presence of a C-terminal domain for the transposase-recombinase activity. Then, we studied the evolutionary history of *pgbd5* and *pgbd* in general. We identified the phylogenetically closest element to pgbd5 and its ability to be transposed by PGBD5. In the last chapter, we have studied the ability of PGBD5 to excise and reintegrate into new loci of human pbles in order to evaluate the ability of PGBD5 to mobilize these elements present in the human genome and to refine the identification of transposase targets.

Keywords : PGBD5, piggyBac, transposase