UNIVERSITÉ FRANÇOIS – RABELAIS DE TOURS

ÉCOLE DOCTORALE : SSBCV

Plasticité Génomique et Expression Phénotypique

THÈSE présentée par :

Sébastien Guizard

soutenue le : **1er juillet 2016**

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

Discipline/ Spécialité : Sciences de la Vie

Étude de l'organisation du génome de poulet à travers les séquences répétées

THÈSE dirigée par :

Dr. Yves Bigot	Directeur de Recherches CNRS, INRA Tours
Benoît Piégu	Ingénieur de Recherches CNRS, INRA Tours

RAPPORTEURS :

Dr. Richard Cordeaux	Directeur de Recherches CNRS, Université de Poitiers
Dr. Emmanuelle Lerat	Chargée de Recherches CNRS, Université Claude Bernard - Lyon 1

JURY :

Dr. Yves Bigot	Directeur de Recherches CNRS, INRA Tours
Dr. Florian Maumus	Chargée de Recherches INRA, INRA Versailles
Dr. Elisabeth Le Bihan-Duval	Directrice de Recherches INRA, INRA Tours
Pr. Alain Goudeau	Professeur, Faculté de Médecine Tours
Dr. Richard Cordeaux	Directeur de Recherches CNRS, Université de Poitiers
Dr. Emmanuelle Lerat	Chargée de Recherches CNRS, Université Claude Bernard - Lyon 1

Remerciements

Je voudrais commencer par remercier le Département PHASE (Physiologie Animale et Systèmes d'Élevage) et la Région Centre pour le financement de ces années de thèse. Je remercie également Florian Guillou et Thierry Magalon de m'avoir accueilli au sein de l'Unité Physiologie de la Reproduction et des Comportements de Tours.

Je remercie les membres du jury qui m'ont fait l'honneur de consacrer du temps à l'évaluation de mon travail de thèse ; Emmanuelle Lerat et Richard Cordeaux d'avoir accepté d'en être les rapporteurs et Florian Maumus, Elisabeth Le Bihan-Duval et Alain Goudeau d'avoir bien voulu en être les examinateurs.

Je remercie les membres de mon comité de thèse, Jérôme Salse, Claire Lemaitre et Hadi Quesnville, pour leurs critiques et conseils avisés sur mon travail.

Je souhaite remercier ensuite Yves Bigot, mon directeur de thèse, de m'avoir offert la possibilité de réaliser cette thèse et de m'avoir fait confiance malgré les conditions particulières de l'oral pour l'attribution de la bourse thèse. Grâce à toi, j'ai réalisé des rêves qui me semblaient impossibles il y a encore quelques années. Merci pour toutes les discussions, que ce soit sur les éléments transposables, la biologie, les sciences, la recherche, la politique, le rugby, le basket, etc., qui ont toujours été plus enrichissantes. Surtout, je te remercie pour la dose quotidienne de jeux de mots capillotractés qui m'ont donné plus d'un fou rire. Encore merci de m'avoir poussé toujours plus loin dans mon travail et de m'avoir montré ce qu'est être un chercheur.

Je remercie l'inventeur de la bambinolâtrie, Benoît Piégu, de m'avoir appris à devenir un vrai barbu de l'informatique. Merci de m'avoir appris comment installer, gérer les machines, comment dompter Linux, de m'avoir fait passer de « Intermediate » à « Expert » en Perl, m'avoir montré la parallélisation dans tous ses états, de m'avoir aidé à battre R… mais aussi d'avoir partagé tes connaissances en Bio en débattant sur le « fumier » pendant des heures. Par contre, je ne te remercie pas de m'avoir laissé gérer le serveur Galaxy pendant tes vacances alors que deux biologistes déchaînés s'acharnaient dessus et, surtout, j'attends toujours mon cluster de calculs que je te réclame depuis mes premiers pas dans le bureau ! :p J'ai vraiment passé de bons moments et je t'en remercie. Et j'espère que tu me donneras mon dernier « *Achievement* » ! ;) .

Je souhaite aussi remercier les autres membres de l’équipe PGEP, qu’ils soient permanents, Florian Guillou, Linda Beauclair, Isabelle Gibert, ou non, Peter Arensburger et Sassan Asgari, ainsi que toutes les personnes que j’ai pu côtoyer à la PRC.

Je tiens à remercier aussi Hadi Quesneville et l’équipe de développement de REPET, Véronique Jamilloux, Olivier Inizan, Florian Maumus, Timothée Chaumier, Mark Moissette, Joëlle Amselem, Mikaël Loaec, Isabelle Luyten, pour leur formidable accueil dans leur labo pendant trois semaines, et surtout pour toute l’aide et l’expertise qu’ils m’ont apportées dans l’utilisation et l’optimisation de REPET, et aussi pour l’initiation à la méthode AGILE tant redoutée depuis mon premier passage à l’URGI avec Florian Murat :p

Je remercie mon médecin traitant de m’avoir prescrit des wagons d’antibio et de Derinox pour tous les rhumes et autres maladies que j’ai pu attraper en trois ans !

Je remercie aussi toute ma famille et ma belle-famille pour leur soutien inconditionnel quand la santé et le moral n’étaient pas au rendez-vous. Mais mes plus grands remerciements vont à ma femme qui a accepté de quitter sa région d’Auvergne pour me suivre en Touraine et qui m’a supporté et encouragé toutes ces années, ainsi qu’à ma fille qui, depuis le 1er avril 2014, m’apporte ma dose de bonheur quotidienne.

Résumé

Les génomes des espèces aviaires ont des caractéristiques particulières comme la structure des chromosomes et le contenu en séquences répétées. En effet, alors que dans les génomes vertébrés, la proportion de répétitions dans le génome varie de 30 à 55 %, dans les espèces aviaires, cette proportion est plus faible et varie de 8 à 10 %.

L'annotation du contenu répété est le plus souvent réalisée avec le programme RepeatMasker qui s'appuie généralement sur la banque de séquences répétées Repbase. Ce genre de méthode repose uniquement sur la séquence des éléments transposables connus. De fait, ce programme n'est pas en mesure de détecter de nouvelles séquences répétées, et la qualité de l'annotation sera donc dépendante de la banque de séquences d'éléments transposables utilisée.

De plus en plus d'études montrent que les éléments transposables jouent un rôle dans le fonctionnement du génome et peuvent influer sur l'expression des gènes. Il est donc primordial que l'annotation de ces séquences soit la plus complète possible. Il existe de nombreux programmes employant des méthodes permettant la détection *de novo* des éléments transposables, soit en recherchant des structures caractéristiques, soit en comparant le génome contre lui-même. Cependant, aucune stratégie globale d'annotation standard des séquences répétées n'a jusqu'à présent été définie.

Au cours de ma thèse a été mise en place une stratégie d'annotation des séquences répétées que nous avons élaborée et appliquée à un génome de grande taille, celui de la poule rouge de jungle. L'annotation ainsi obtenue m'a permis d'étudier l'organisation du génome de cette espèce au travers de ses séquences répétées et éléments transposables.

Résumé en anglais

The genomes of avian species have special features such as the structure of chromosomes or their content in repeated sequences. Indeed, compared to vertebrate genomes in which the amount of repetitions varies from 30 to 55%, it is lower in avian species and varies from 8 to 10%.

The annotation of repeated content is most often done with the RepeatMasker program that is generally use the Repbase database of repeated sequences. This kind of approach is based solely on the sequence of already known transposable elements. In fact, this program is not able to detect new repeats and in consequence produced annotations with a quality that depends on the sequences of transposable elements used.

More and more studies show that transposable elements play a role in the functioning of the genome and can influence gene expression. It is therefore essential that the annotation of these sequences is as complete as possible.

There are many programs using methods for detecting *de novo* transposable elements, either by searching for characteristic structures, or by comparing the genome against itself. However, no standard strategy of annotation for repeated sequences have been defined yet.

My thesis aims to set-up a standard strategy of annotation for repeated sequences that was applied to a large genome, that of the red jungle fowl. The obtained annotation allowed me studying the genome organization in this species through its repeated sequences and transposable elements.

Table des matières

Liste des tableaux

Liste des figures

Liste des annexes

Glossaire

Modèle (génome) :	Séquence nucléique se rapprochant de la séquence génomique réelle présente dans le noyau.
Draft :	Version non finalisée d'un modèle de génome.
Lectures (Reads) :	Séquences d'ADN produites d'un séquençage. La longueur et le nombre de lectures produites varient en fonction de la technologie de séquençage.
Assemblage :	L'assemblage de génome consiste à reconstituer un génome à partir d'un ensemble de lectures.
Contig :	Assemblage de lectures de séquençage.
Scaffold :	Assemblage de contigs.
Unlocalized :	Séquences assemblées qui ont pu être affectées à un chromosome du modèle, mais qui n'ont pu être placées au sein de celui-ci.
Unplaced :	Séquences assemblées qui n'ont pu être affectées à un chromosome du modèle.
Unaffected :	Chromosome artificiel inclus dans le modèle comprenant toutes les séquences Unplaced.
HSP	High Scoring Pair : paire de séquences de longueur égale dont l'alignement local est maximal et dont le score d'alignement atteint ou dépasse un seuil minimal.

Abréviations

ADN : Acide DésoxyriboNucléique

ARN : Acide RiboNucléique

ARNr : Acide RiboNucléique ribosomique

BAC : Bacterial Artificial Clone

BED : Browser Extensible Data

cM : centiMorgan

CNV : Copy Number Variation

Da : Dalton

ENV : glycoprotéine d'enveloppe

ET : Élément transposable

FAO : Food and Agriculture Organization

GAG : Group AntiGens

Gbp : Giga paire de bases

GFF3 : General Feature Format

Go : Gigaoctet

ICGSC : International Chicken Genome Sequencing Consortium

ISB : Institut for Systems Biology

kpb : kilo paire de bases

LTR : Long Terminal Repeat

Mbp : Mega paire de bases

NGS : Next Generation Sequencing

ORF : Open Reading Frame

pb : paire de bases

PBS : Primer Binding Site

POL : polymérase

RM : RepeatMasker

sMAR : Scaffold/Matrix Attachment Region

SSR : Simple Sequence Repeat

TIR : Terminal Inverted Repeat

To : Tera octet

TRF : Tandem Repeat Finder

TSD : Target Site Duplication

UCSC : University of California, Santa Cruz

Un : Unaffected

Unl : Unlocalized

Unp : Unplaced

Avant-Propos

Tableau 1 : Technologies et coûts de séquençage
Source : Liu et al : Comparison of next-generation sequencing systems. J Biomed Biotechnol 2012, 2012:251364.

Tableau 2 : Historique des « principales » espèces dont le génome a été séquencé

Année	Espèce	Année	Espèce
1976	Bactériophage *MS2*	2002	*Mus musculus* *Oryza sativa*
1977	Phage *Φ-X174*	2003	***Homo sapiens*** [Lancement de ENCODE]
1995	*Haemophilus influenzae* *Mycoplasma genitalium*	2004	*Rattus norvegicus* *Gallus Gallus*
1990	[Lancement « Human Genome Project »]	2005	*Pan troglodytes* *Canis lupus familiaris boxer*
1996	*Saccharomyces cerevisiae* *Methanococcus jannaschii*	2006	*Ostreococcus tauri*
1997	*Escherichia coli*	2007	***Homo sapiens*** *JC Venter* *Vitis vinifera*
1998	*Mycobacterium tuberculosis* *Caenorhabditis elegans*	2008	***Homo sapiens*** *JD Watson* ***Homo sapiens*** *Nigerian male* ***Homo sapiens*** *YanHuang No.1* *Ornithorhynchus anatinus*
1999	Aeropyrum pernix K1	2009	*Sorghum bicolor* ***Homo sapiens*** *Korean male* *Equus caballus thoroughbred* *Zea mays.*
2000	*Drosophila melanogaster* *Arabidopsis thaliana*	2010	*Glycine max* *Xenopus tropicalis* *Malus x domestica cv Golden Delicious*
2001	***Homo sapiens*** (1st draft) Staphylococcus aureus aureus	2011	*Macropus eugenii* *Anolis carolinensis*
		2012	*Tursiops truncatus* *Fibroporia radiculosa* *Bos indicus*

Depuis leur découverte, les génomes n'arrêtent pas de surprendre les biologistes en révélant toujours plus d'informations sur leurs constituants et les différents mécanismes qui les régissent.

Nommé « nuclein » lors de sa première mise en évidence par Friedrich Miescher en 1869, la vision de l'ADN a évolué au fur et à mesure des observations, des découvertes, et des avancées techniques et technologiques. L'ADN a été initialement décrit comme un isolat non protéique et non lipidique riche en phosphore. Il prend le nom de chromatine lorsque Walther Flemming décrit la mitose des cellules en 1879. La structure double hélice de l'ADN composé de deux bases puriques et deux bases pyrimidiques est finalement décrite par Rosalind Franklin, Francis Crick et James Watson en 1953. Connaissant sa structure, il est alors devenu possible de décrire le code génétique décrypté en 1966 par Marshall Nirenberg. Au fil des années, il est devenu évident que pour mieux comprendre les organismes vivants, il est indispensable de connaître la séquence ADN de leur génome. La première technique de séquençage « rapide » de l'ADN fut créée en 1975 par Allan Maxam et Walter Gilbert et permettait de séquencer 100 bases par « run ». Cette nouvelle technologie a rapidement permis à Walter Fiers en 1976 de séquencer le premier génome ARN, celui du bactériophage MS2. En 1977, Frederick Sanger fut le premier à séquencer un génome ADN, celui du phage Φ-X174, en utilisant la méthode dite de terminaison de chaîne [Sanger et al 1977].

Depuis cette époque, les méthodes de séquençage n'ont cessé de s'améliorer, offrant la possibilité de séquencer de plus en plus de bases, beaucoup plus rapidement et surtout à des coûts de plus en plus réduits [Liu et al 2012] (Tableau 1).

Grâce au développement des techniques de séquençage, la communauté des biologistes a pu produire au cours des deux dernières décennies un grand nombre de séquences génomiques sur des espèces de plus en plus complexes (Tableau 2). Alors que le projet « Génome Humain » débuté en 1990 a demandé aux chercheurs 13 ans pour produire un modèle de séquence complet, il est maintenant possible de séquencer ce génome en l'espace d'une semaine pour un coût d'environ 1 000 €, sans tout à fait atteindre cependant le même degré de précision. Ces nouvelles performances ont changé la vision de l'ADN permettant de mieux décrire les différents compartiments des génomes.

Tableau 3 : Taille de génomes et nombre de gènes
Source : Krishnan J: Code in Non-coding. Proc Indian Natl Sci Acad 2015, 81:609–628.

L’annotation des gènes chez *Homo sapiens* a démontré que son génome contient seulement 23 000 gènes [Krishnan et al 2015] pour un nombre de protéines qui dépasse de loin la complexité génique, mettant ainsi à mal le dogme « une protéine = un gène » émis par George Beadle et Edward Tatum en 1941. Les gènes couvrent plus d’un tiers du génome humain (~41 % dans hg38), mais la partie codante, les exons, ne représentent que 2 %. Bien qu’ils soient les porteurs de l’information génétique, le nombre de gènes n’est pas lié à la taille du génome (Tableau 3).

La partie de l’ADN non comprise dans les exons représente donc 98 % du génome pour l’homme. Elle est appelée ADN non codant (ADNnc) ou matière noire. Longtemps dénommée « Junk DNA » [Ohno 1972], elle a été considérée comme de l’ADN parasite sans effet délétère [Orget et al 1980]. Cette matière noire est composée de séquences uniques, et de séquences répétées ayant la capacité de se déplacer dans le génome, les éléments transposables. De nos jours, ces séquences ne sont plus considérées comme de l’ADN fardeau ou poubelle, mais comme des séquences participant au fonctionnement du génome.

Présents dans la quasi-totalité des espèces, que ce soit les cellules eucaryotes et procaryotes, ou les virus, les éléments transposables ont une diversité et une représentation qui sont très variables d’un génome à un autre. Ayant une couverture de génome de 0,14 % chez *Tetraodon nigroveridis* (poisson de l'espèce tétrodon vert) à plus de 95 % chez les *Lilium* (Lys), ils représentent 45 à 50 % dans le génome humain. Le génome du poulet se démarque des autres espèces de vertébrés avec une couverture du génome par les séquences répétées de seulement 8 à 10 %.

De nombreux programmes de détection et d’annotation des éléments transposables ont été développés en s’appuyant sur différentes méthodes ayant chacune ses avantages et ses défauts. Le programme le plus utilisé est RepeatMasker (RM) qui se base sur les séquences d’éléments transposables connus. Cette méthode ayant plusieurs défauts et n’étant pas capable de détecter de nouveaux éléments transposables, nous avons mis en place une nouvelle stratégie de détection et d’annotation des séquences répétées qui s’appuie sur une série d’outils d’analyse fonctionnant *de novo*.

Dans ce manuscrit, je commencerai par présenter les espèces aviaires en évoquant les génomes d'oiseaux séquencés puis leurs caractéristiques. J'aborderai ensuite le génome du poulet en débutant par la présentation de la version 4 modèle du génome puis j'exposerai son contenu en gènes et en séquences répétées.

Je ferai une présentation des différents types de séquences répétées, dont les éléments transposables, puis j'évoquerai leurs impacts possibles sur les génomes. J'exposerai ensuite les différentes classifications existantes et ferai l'inventaire des séquences génétiques mobiles du génome du poulet. Ensuite, je ferai une revue des différentes méthodes de détection et d'analyse des différents types de séquences répétées en présentant d'abord les méthodes existantes puis les programmes disponibles. Enfin, je présenterai les objectifs de la thèse ainsi que les travaux qui en ont découlé. Je terminerai par une discussion des résultats et une conclusion.

Introduction

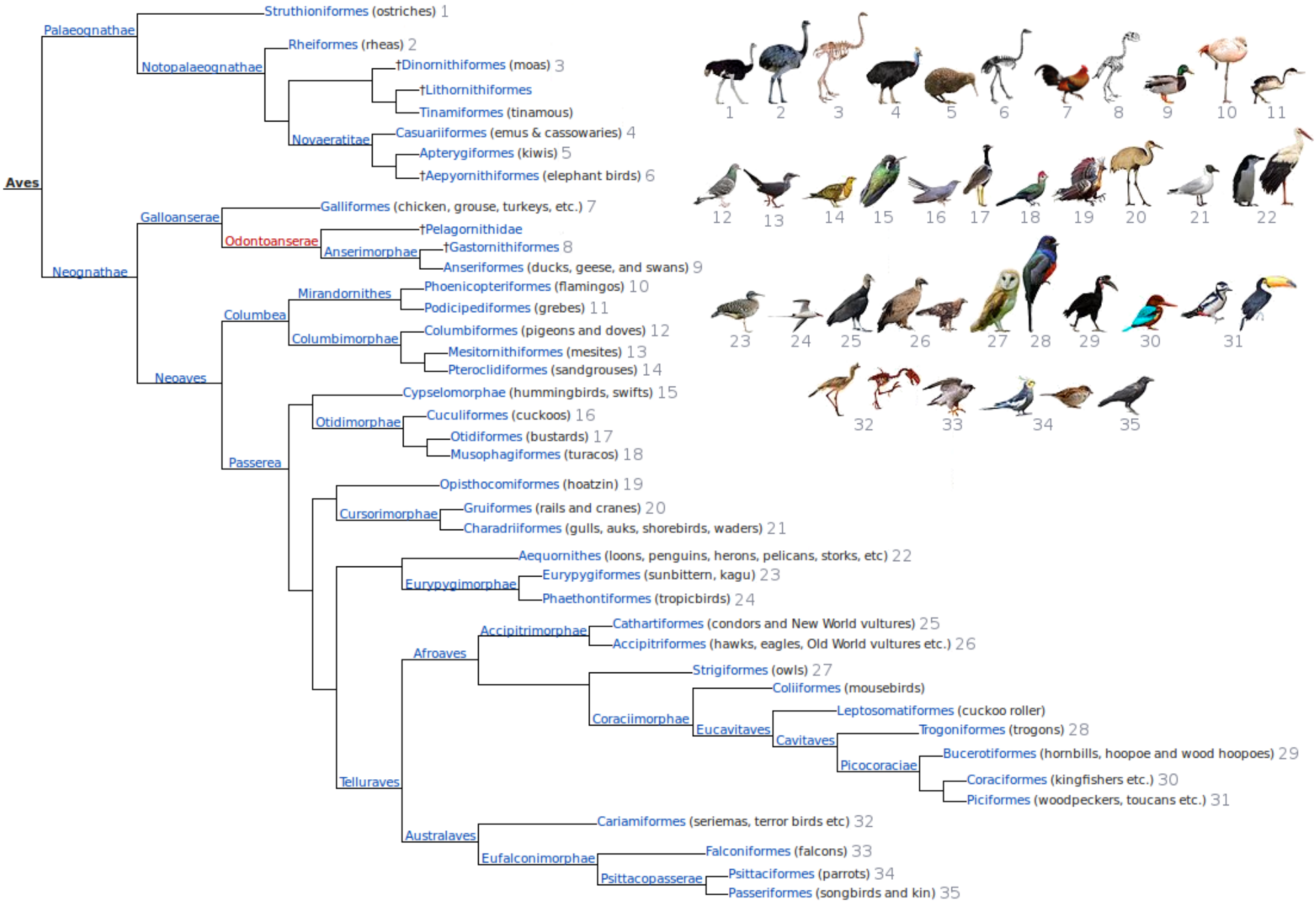


Figure 1 : Arbre phylogénétique des oiseaux
Source : Jarvis et al. (2014), Yury et al. (2013), Wikipedia

1. Les génomes aviaires

1.1. Les oiseaux

Appartenant à la classe des amniotes, vertébrés tétrapodes qui possèdent un amnios[1], les oiseaux forment le clade des *Aves*. Avec plus de 10 500 espèces, ils représentent la classe des vertébrés tétrapodes ayant la plus grande diversité d'espèces. Leur classification se découpe en deux grands groupes (Figure 1) [Jarvis et al 2014, Yuri et al 2013]. Les *Palaeognathae* incluent tous les oiseaux ayant perdu leur capacité de vol, tels que l'autruche, le kiwi, l'émeu, le nandou, ainsi que des espèces disparues comme les moas ou les aépyornithidés (oiseaux-éléphants) dont la taille pouvait atteindre 3 mètres. La branche des *Neognathae* comprend la quasi-totalité des espèces d'oiseaux vivants. Elle se décompose en deux clades. Les *Galloanserae* qui comprennent le poulet et la dinde ; espèces formant le clade G*alliforme* ; et des espèces aquatiques comme le canard, le cygne ou l'oie. Les *Neoaves* incorporent le clade *Passerea* regroupant tous les Passeriformes (Mésange, Grand corbeau…), soit plus de la moitié des espèces d'oiseaux.

Les oiseaux sont largement étudiés. En effet, ils sont un modèle de choix pour la biologie développementale grâce à la manipulation « facile » des embryons, l'étude de l'expression des gènes, mais aussi la création de lignées transgéniques par différentes méthodes telles que l'utilisation de vecteurs viraux ou transposons injectés directement dans l'œuf ou les cellules primordiales germinales [Doran et al 2016, Seidl et al 2013, Poynter et al 2013]. Ce modèle est aussi utilisé dans le domaine de la neurologie comme exemple pour l'étude des réseaux de neurones activés lors de l'apprentissage du chant [Gobes et al 2010, Balakrishnan et al 2014, Mello et al 2015, Mello 2014]. La compréhension de leur sensibilité aux pathogènes et leur immunologie est également importante pour limiter les risques sanitaires. En effet, l'élevage intensif d'espèces aviaires telles que le poulet, la dinde et le canard représente une très grande part de la production de viande mondiale (>30 %[2]).

1L'amnios, ou sac amniotique, est l'enveloppe qui se constitue autour de l'embryon et qui a pour rôle de le protéger en maintenant autour de lui un liquide amniotique.
2http://www.planetoscope.com/elevage-viande/1276-consommation-de-poulets-et-volailles-dans-le-monde.html

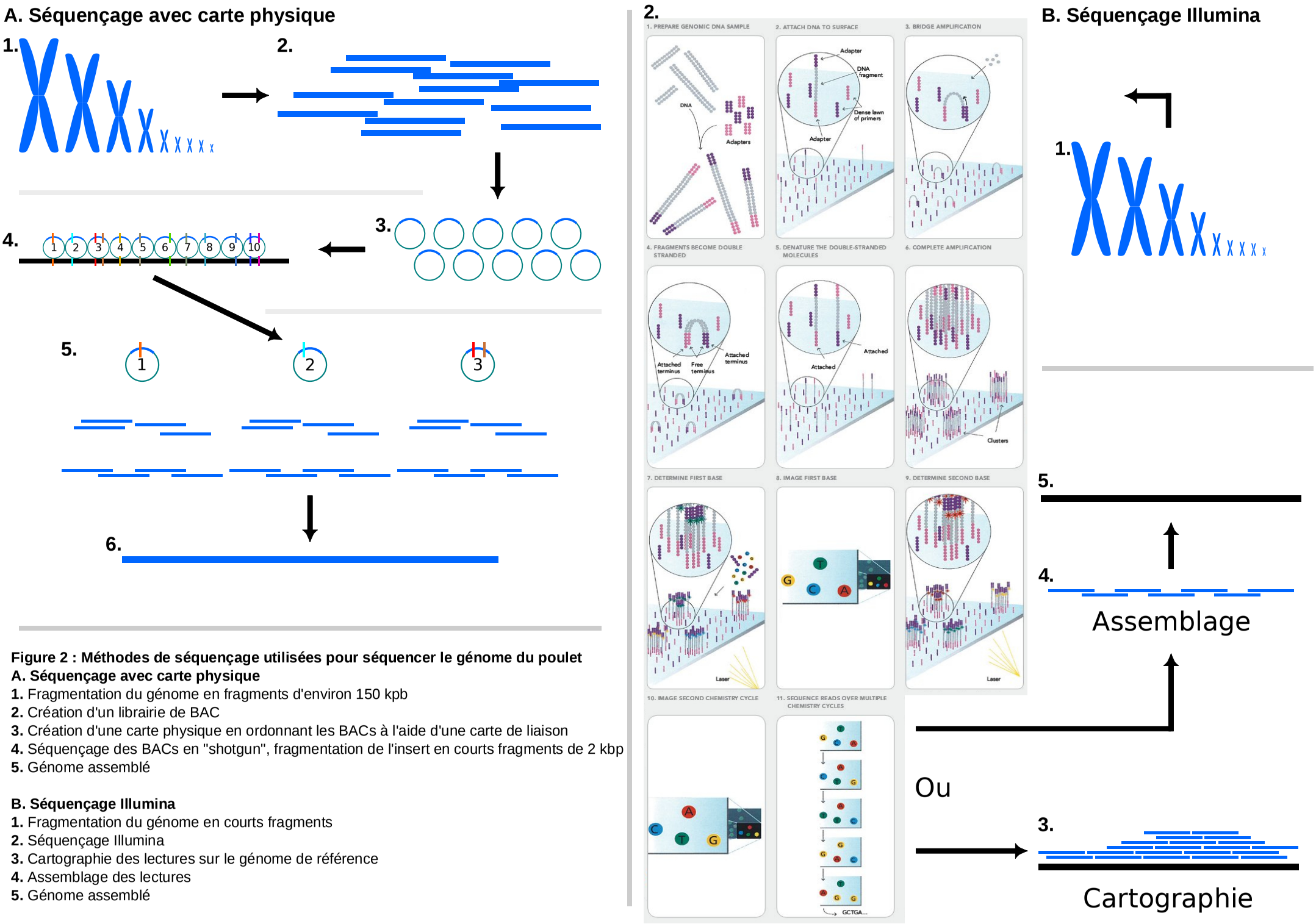


Figure 2 : Méthodes de séquençage utilisées pour séquencer le génome du poulet

A. Séquençage avec carte physique

1. Fragmentation du génome en fragments d'environ 150 kpb
2. Création d'un librairie de BAC
3. Création d'une carte physique en ordonnant les BACs à l'aide d'une carte de liaison
4. Séquençage des BACs en "shotgun", fragmentation de l'insert en courts fragments de 2 kbp
5. Génome assemblé

B. Séquençage Illumina

1. Fragmentation du génome en courts fragments
2. Séquençage Illumina
3. Cartographie des lectures sur le génome de référence
4. Assemblage des lectures
5. Génome assemblé

L'infection d'élevages présente donc un risque important pour la santé humaine. La propagation des pathogènes peut être très rapide à l'échelle mondiale, du fait de la contamination des élevages et la propagation via la migration des espèces sauvages. C'est pourquoi une grande part de la recherche est axée sur l'étude et l'amélioration de la résistance aux pathogènes de ces espèces, notamment avec la création de lignées résistantes par sélection génomique [Clement et al 2015, Liu et al 2013, Huang et 2015, Ha 2008, Nelson et al 2016].

Les enjeux et l'importance des espèces aviaires d'un point de vue agronomique, économique ou de sécurité alimentaire, ont fait que le troisième génome vertébré séquencé fut celui de la poule rouge de jungle.

1.2. Les génomes aviaires séquencés

Alors que la première version du génome humain a été publiée en 2001, le premier génome aviaire séquencé a été publié en 2004. L'espèce sélectionnée fut la poule rouge de jungle, Gallus gallus, car elle est considérée parmi les 4 espèces asiatiques (*G. sonnerattii*, *G. lafayetii* et *G. varius)* comme étant celle à l'origine de toutes les variétés domestiques anciennes et actuelles de poule. Le séquençage fut réalisé par le « International Chicken Genome Sequencing Consortium » [Hillier et al 2004] qui regroupe 175 chercheurs provenant de 50 laboratoires localisés en Chine, au Danemark, en France, en Allemagne, au Japon, en Pologne, à Singapour, en Espagne, en Suède, en Suisse, au Royaume-Uni et aux États-Unis. Doté d'un budget de 13 millions de dollars, il a produit un modèle de génome en combinant le séquençage de plasmides, fosmides et de chromosomes bactériens artificiels (Bacterial Artificial Chromosomes, BAC) avec une profondeur du séquençage final de 6X et un ordonnancement des séquences obtenu en utilisant une carte physique de marqueurs (Figure 2A). Il fut le seul génome aviaire disponible jusqu'au séquençage des génomes du diamant mandarin et de la dinde en 2008 et en 2010. La diversité des génomes disponibles s'est étoffée à partir de 2013 où 9 génomes d'espèces de perroquets, de canards, de passereaux, de faucons et de gallinacés ont été séquencés. En 2014, le « avian phylogenomic project »[3] [Zhang et al 2014], avec l'objectif d'étudier la diversité génomique aviaire et de résoudre la phylogénie des oiseaux, a séquencé 48 espèces d'oiseaux, 46 faisant partie des Neognathaes et 2 des Palaeognathaes. Ils ont utilisé les technologies Illumina qui permettent de produire plusieurs millions de lectures de 100 à 250 pb (Figure 2B). Il faut cependant noter

3 http://avian.genomics.cn/en/index.html

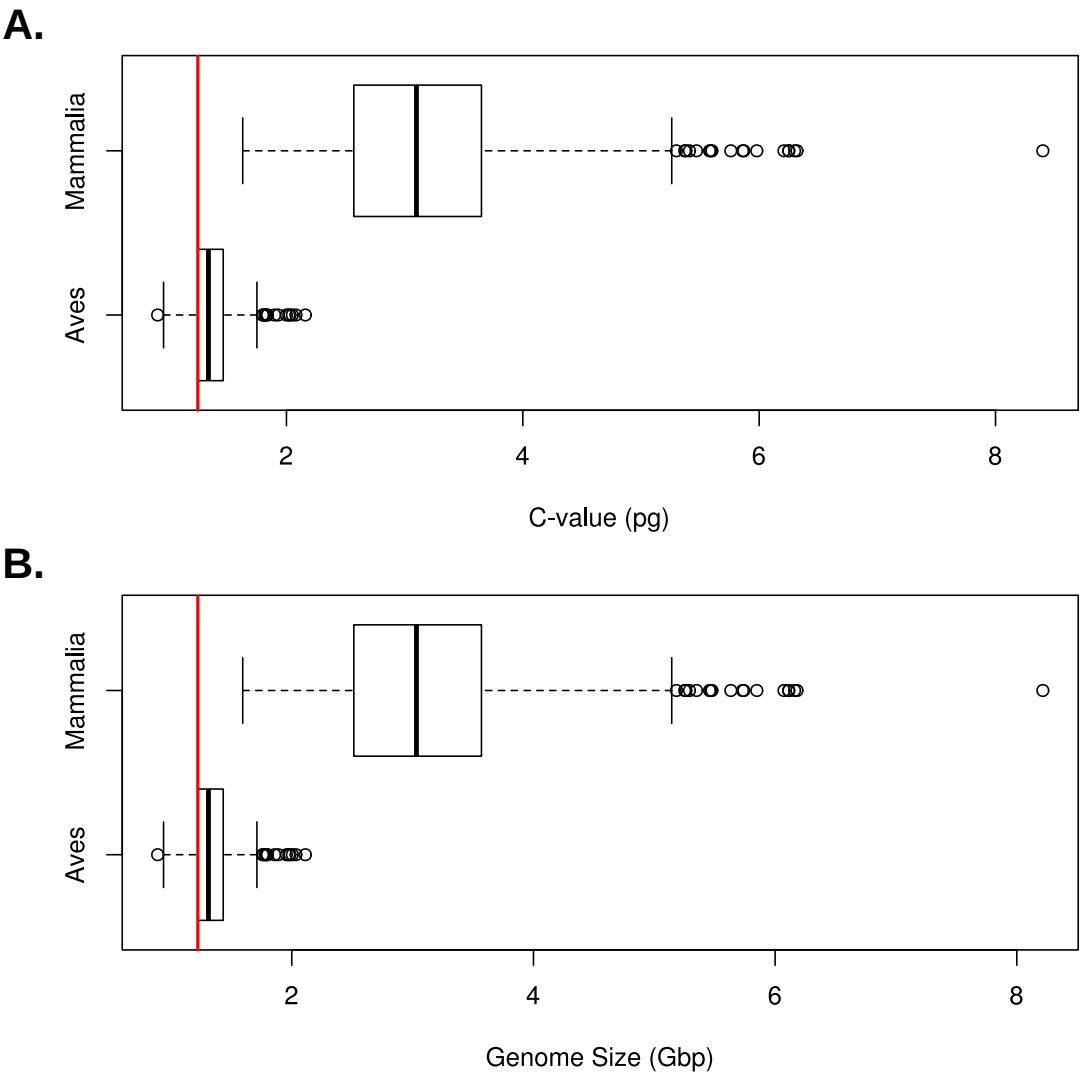


Figure 3 : Boxplots de la taille des génomes de mammifères et des espèces aviaires
La ligne de rouge indique la taille du génome de la poule rouge de jungle
A. Taille des génomes en picogrames (C-value)
B. Taille des génomes en Gpb, calculées en multiplant la C-value par 0.978 (1 pg = 978 Mpb)

que la majorité de ces génomes n'ont pu être assemblés qu'en scaffolds (Annexe 1). Mis à part celui de poule, seuls les génomes de la dinde, de la mésange charbonnière, du diamant mandarin, et de la caille japonaise ont pu être assemblés jusqu'au niveau des chromosomes en s'appuyant sur celui de la poule rouge de jungle pour les deux premiers et des cartes physiques pour les deux autres. Le modèle de la poule rouge de jungle (*Gallus gallus*) est donc la référence pour les génomes aviaires, car il est celui qui bénéficie de la carte physique la plus fiable.

1.3. Les caractéristiques des génomes aviaires

Le génome des oiseaux se distingue par plusieurs caractéristiques inhabituelles chez les vertébrés. Ainsi, alors que les chromosomes mammifères ont des tailles relativement homogènes, ceux des oiseaux, accipitridés exceptés (rapaces) [Bed'Home et al. 2003], sont organisés en macrochromosomes et en microchromosomes. Vis-à-vis des macrochromosomes, les microchromosomes présentent un taux de GC plus élevé, des densités supérieures en îlots CpG et en gènes, mais une densité en séquences répétées plus faible [Smith et al 2000, McQueen et al 1996, Andreozzi et al 2001]. Ils se démarquent également des macrochromosomes avec un taux de recombinaison plus élevé, 2,8 $cM.Mbp^{-1}$ pour les macrochromosomes et 6,4 $cM.Mbp^{-1}$ pour les microchromosomes (1 cM représentant approximativement 1 % de recombinaison) [Hans Ellegren 2005, Axelsson et al 2005]. Les microchromosomes présentent une hyperacétylation des histones H4, une caractéristique se parfaitement corrélée à leur forte densité en gènes [McQueen al 1998]. Les macrochromosomes, quant à eux, présentent une densité plus forte en ET, une densité plus faible en gènes et un taux de recombinaison plus faible.

Alors que la taille des génomes vertébrés mammifères oscille entre 1,59 Gbp pour la chauve-souris *Lophostoma carrikeri* et 8,22 Gbp pour le rat-viscache roux d'Argentine, les génomes aviaires ont une taille allant de 0,89 Gpb pour le Colibri à gorge noire à 2,11 Gpb pour l'autruche (Figure 3) [Gregory 2005]. La majorité des génomes aviaires sont plus petits que les génomes vertébrés, à l'exception de certains tétrapodes et poissons. Plusieurs hypothèses ont été émises pour expliquer cette taille réduite.

Des études comparatives ont également démontré que les introns dans les espèces dotées de la capacité de vol sont significativement plus courts [Hughes AL 1995]. En

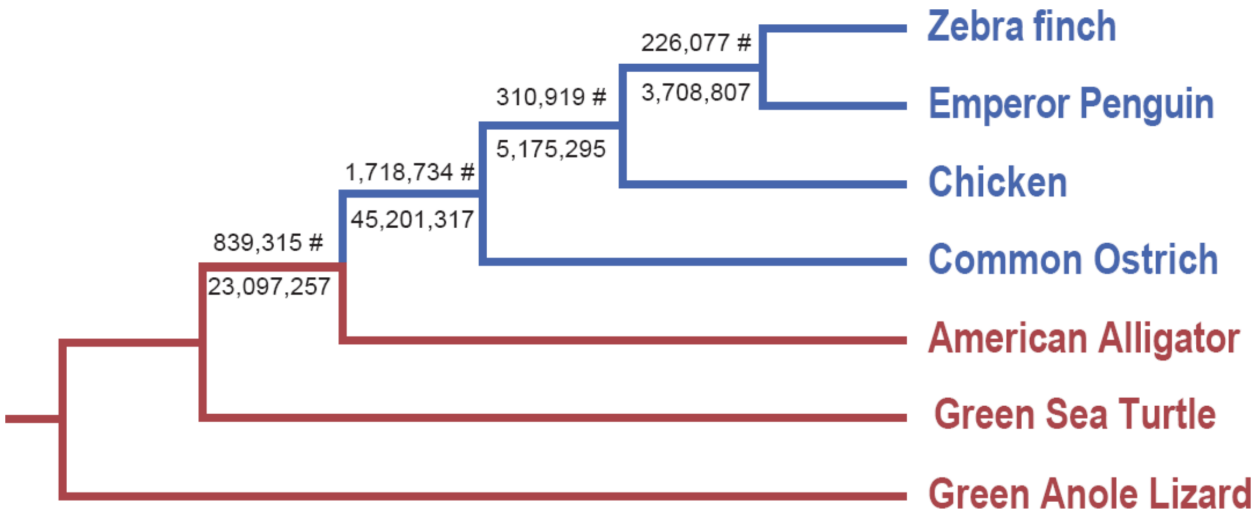


Figure 4 : Distribution des événements de petites délétions à travers l'arbre phylogénétique généré sur l'alignement d'une région génomique de 163 Mpb.
Les valeurs au-dessus des branches sont le nombre de délétions dans chaque lignée. Les valeurs sous les branches indiquent la taille totale en paire de bases des évènements de délétion.
Les espèces aviaires sont en bleu et les reptiles sont en rouge.
Les espèces aviaires ont été choisies pour être représentatives des clades majeures : Paleognathae (Autruche), Galloanserae (Poulet), Neoaves aquatiques (Pingouin), and Neoaves (Diamant mandarin).

Source: Zhang et al : Comparative genomics reveals insights into avian genome evolution and adaptation. Science (80-) 2014, 346:1311–1320.

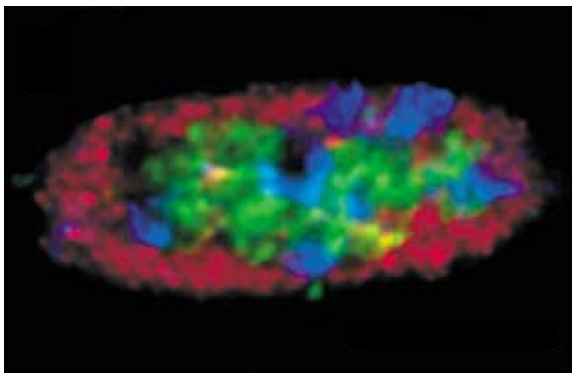


Figure 5 : Visualisation des chromosomes dans le noyau d'un fibroblaste de poulet
Les chromsomes sont visualisés par FISH. Les macrochromosomes (1 à 5 et Z) sont colorés en rouge, les chromosomes de taille moyenne (6 à 10) en bleu et les microchromosomes (11 à 28 et X) en vert.

Source : Habermann F et al, Arrangements of macro- and microchromosomes in chicken cells. Chromosome Res 2001, 9:569–84.

comparaison avec les mammifères et les reptiles, la taille des gènes codant pour des protéines est respectivement 50 % et 27 % plus petite chez les oiseaux [Zhang G et al 2014]. Une autre hypothèse propose que l'ancêtre commun aux oiseaux et reptiles a subi un très grand nombre de délétions et que ces pertes ont continué dans les espèces actuelles (Figure 4). Enfin, une troisième hypothèse s'appuie sur le fait qu'il y aurait peu d'éléments transposables. En effet, chez les mammifères, leur proportion peut varier de 34 à 52 % [Böhne A et al 2008]. Selon le « avian phylogenomic project », les transposons représentent de 4 à 10 % du génome.

Toutes ces caractéristiques (petits introns, faible quantité d'ADN répétés, délétions) pourraient être liées à une pression de sélection allant dans le sens d'une diminution de la taille du génome liée au métabolisme. Il a ainsi été proposé que les besoins métaboliques liés au vol battu auraient favorisé une réduction du génome. Sur la base de l'idée qu'avec un génome plus petit, les cellules sont de taille réduite, les échanges gazeux seraient ainsi optimisés, ce qui favoriserait l'aptitude au vol [Zhang et al 2012, Hughes et al 1995]. Cette hypothèse est appuyée notamment par les deux corrélations suivantes :

1) les chauves-souris, qui sont les seuls mammifères à réaliser un vol battu, ont aussi de petits génomes ;

2) les espèces aviaires ayant perdu leur capacité de vol comme l'autruche ont un génome de plus grande taille.

Il faut cependant rester prudent sur ce type d'explication lié au vol. En effet, le génome des nandous et des émeus, des ratites apparentés à l'autruche, est respectivement de 1,5 et 1,6 Gpb.

Au-delà de la taille des chromosomes et du génome, les génomes aviaires semblent se distinguer au travers de deux autres caractéristiques. Tout d'abord, les chromosomes aviaires ont une localisation nucléaire particulière au cours de la mitose et dans la cellule quiescente. En effet, les études cytologiques ont montré que les macrochromosomes se placent à la périphérie du noyau alors que les microchromosomes se situent au centre (Figure 5) [Habermann et al 2001]. Les génomes aviaires se démarquent aussi par la très grande conservation de leurs génomes au cours de l'évolution. L'étude comparative des génomes aviaires révèle une très forte synténie : le nombre et la structure des chromosomes sont

« presque inchangés » entre les espèces. De plus, il semblerait qu'il n'y ait aucun échange d'ADN entre les macrochromosomes et les microchromosomes, du moins vis-à-vis de ce qui est observé par exemple chez les mammifères [Nanda et al 2011, Romanov et al 2014].

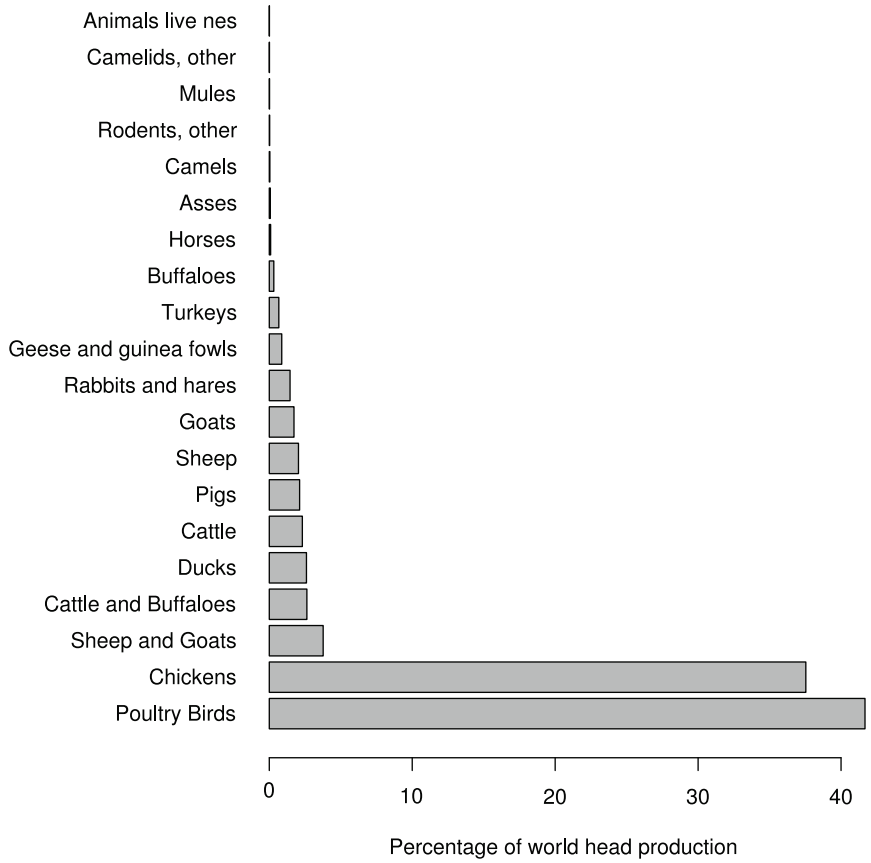


Figure 6 : Histogramme de la production animale mondiale en nombre de têtes en 2014
Source : http://faostat3.fao.org/home/E

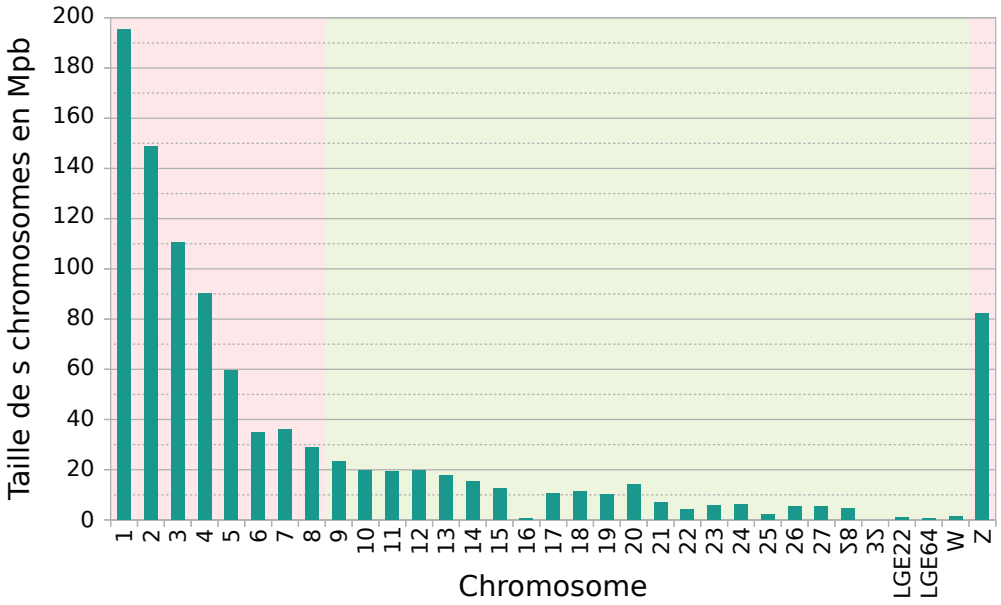


Figure 7 : Histogramme de la taille des chromosomes dans le modèle du génome de la poule rouge de jungle
Le fond de couleur rose de l'histogramme indique les macrochromosomes et le fond de couleur verte indique les microchromosomes

2. Le génome du poulet

Le choix de séquencer le génome de poule rouge de jungle a été motivé par son importance économique mondiale. En effet, selon les statistiques de la FAO (Organisation des Nations Unies pour l'alimentation et l'agriculture ; Food and Agriculture Organization), la production en 2014 de poulet et d'espèces aviaires représente en nombre d'individus plus de 80 % de la production mondiale d'animaux (Figure 6). Le poulet est une production majeure en élevage et la connaissance approfondie de son génome doit pouvoir servir à la sélection de l'espèce, par exemple pour l'amélioration de la résistance aux pathogènes ou la résilience aux stress thermiques.

2.1. Le modèle

Bien qu'il ait été le premier génome aviaire publié, la faible profondeur de séquençage (6X) a mené à plusieurs re-séquençages à l'aide des nouvelles technologies de « Next Generation Sequencing » (NGS) [Kelley et al 2010 ; Ye et al 2011 ; Zhang et al 2014]. En 2011, la version 4 du génome a été publiée à la suite d'un re-séquençage par la méthode du pyroséquençage 454, ce qui permit de produire un million de lectures de 700 pb avec très peu d'erreurs (sauf pour les homopolymères) et une profondeur de 12X. Parmi les 39 chromosomes du génome, le modèle comprend les séquences des macrochromosomes 1 à 9 dont la taille varie de 195 Mpb à 28 Mpb, celles des microchromosomes de 10 à 28 et 32 dont la taille varie de 23 Mpb à 1028 bp, celles des chromosomes sexuels Z et W dont les tailles respectives sont de 82 Mpb et 1,2 Mpb, et deux « linkage group » de 800 kpb et 965 kbp (Figure 7). Les chromosomes 29 à 31 et 33 à 39 sont absents de ce modèle. Chaque version du génome inclut les contigs qui n'ont pas été intégrés dans les chromosomes. Ils sont classés en deux catégories : les « Unlocalized » (Unl) désignent les séquences qui ont pu être affectées à un chromosome mais qui n'ont pu être placées au sein de celui-ci ; les « Unplaced » (Unp), dont le chromosome d'origine n'a pas pu être déterminé. Ces deux types de contigs sont concaténés dans un chromosome artificiel appelé « Unaffected » (Un). La version 4 comprend 1 805 séquences Unl pour 6,8 Mpb et 14 093 séquences Unp pour 352,7 Mpb.

2.2. Taille du génome

Afin d'évaluer la qualité du modèle de génome construit *in silico*, il est intéressant de comparer la taille cumulée de sa séquence à la taille réelle du génome au sein du noyau. La mesure de la C-value consiste à mesurer la masse d'ADN contenue dans le noyau d'une cellule puis d'en déduire la taille du génome. En effet, l'ADN est composé de 4 bases azotées, A, C, G, T (Adénine, Cytosine, Guanine, Thymine), dont les masses moléculaires sous forme de désoxyribonucleotide sont de 331,2213 Da, 307,1966 Da, 347,2207 Da et 322,2079 Da. Si l'on considère qu'il y a dans le génome 50 % de GC, on peut estimer que la masse moléculaire moyenne d'une paire de bases est de 615,8771 Da. La masse d'une paire de bases peut ensuite être déduite en multipliant cette valeur par l'unité de masse atomique, $1{,}660539 \times 10^{-27}$ kg. Par conséquent, la masse d'un dinucléotide est de $1{,}023\times 10^{-9}$ pg par paire de bases, et donc la densité moyenne de l'ADN est de 978 Mbp/pg [Doležel et al 2003]. La première étude visant à déterminer cette C-value l'a estimée à l'aide de cinétiques de ré-association à 1,1 à 1,4 pg sur le génome d'une poule femelle (Chromosomes sexuel = ZW) [Eden et al 1978, Olofsson et al 1983]. Récemment, cette valeur a été ré-évaluée par cytométrie de flux à 1,25 ± 0,6 pg [Tiersch et al 1991, Organ et al 2007, Mendonça et al 2010, Gregory 2015] soit une estimation de la taille de 1,223 ± 0,058 Gpb.

La somme cumulée de la taille des chromosomes, des Unp et des Unl du modèle est de 1,047 Gbp. Il y a donc une différence 0,176 Gpb entre le modèle et la taille réelle du génome. De plus, les séquences contiennent des « N » qui servent à combler les « trous » de l'assemblage quand la distance entre deux contigs est connue. Le modèle contient 14 Mpb de Ns. L'explication de cette différence entre la taille du génome nucléaire et la taille du modèle a plusieurs origines.

La première est le fait que les extrémités internes et externes des bras des chromosomes du poulet ont une caractéristique spécifique des génomes aviaires. Contrairement aux autres génomes, ils possèdent des méga-télomères [Nanda et al 1994, Delany et al. 2000, Delany et al 2007] et des méga-centromères [Shang et al 2010] dont la taille peut aller jusqu'à 2Mb et qui représentent de 4-8 % du génome. Les télomères sont facilement localisables à l'aide de leur unité de répétition, TTAGGG. Cependant, le modèle de la version 4 du modèle ne contient que quelques kpb de cette séquence.

Le deuxième élément est que les gènes codant l'ARN ribosomal (ARNr) sont manquants dans le modèle du génome. Ces séquences sont absentes même des génomes modèles les plus aboutis comme le génome humain. Normalement, dans le génome du poulet, les gènes codant les ARNr 18S, 5.2S, 28S sont présents sous forme d'environ 400 copies et l'ARNr 5S est présent sous forme de 100 copies. Cet ensemble représente environ 1 % de la taille du génome.

Enfin, la taille des chromosomes évaluée dans les caryotypes diffère de celle de séquences du modèle. Cette comparaison démontre que certains chromosomes du modèle sont plus petits et donc qu'une grande partie de leur séquence est manquante. Cela est d'autant plus évident pour le chromosome 16 dont la taille dans le modèle est de 0,5 Mpb alors que sa taille réelle est estimée à 11 Mpb. De même, pour le chromosome 32 dont la taille dans le modèle est anecdotique, 1028 pb.

En janvier 2016, une nouvelle version (5) qui complète la version 4 a été publiée. Elle inclut la séquence du chromosome 33 et un « linkage group » a été supprimé. Ce nouvel assemblage a une taille cumulée de 1,232 Gpb, ce qui est très proche de la taille estimée du génome biologique. Les séquences nouvellement intégrées ont été principalement affectées aux Unl et Unp. Dans la version 4 du modèle, 90 % des séquences sont classées comme Unp alors que dans la version 5, elles ne représentent que 35 % des séquences.

2.3. Le contenu du génome

Dans cette partie, je commencerai par décrire le contenu en gènes dans des versions 4 puis 5. Pour les séquences répétées, je ferai une description des différents types, les répétitions en tandem et les éléments transposables, que je présenterai en détail en exposant leur origine, leur impact sur les génomes. Je finirai par la présentation des éléments transposables du poulet.

2.3.1. Les gènes

Dans la version 4 du modèle du génome du poulet sont décrits 21 635 gènes qui couvrent 43,96 % du génome. 19 119 gènes sont présents dans les chromosomes, 424 dans les Unl et 2 092 dans les Unp. La partie codante, les exons, couvre 4,93 % du génome. Le contenu en gène fait débat, car il semblerait que plusieurs blocs de gènes aient disparu des

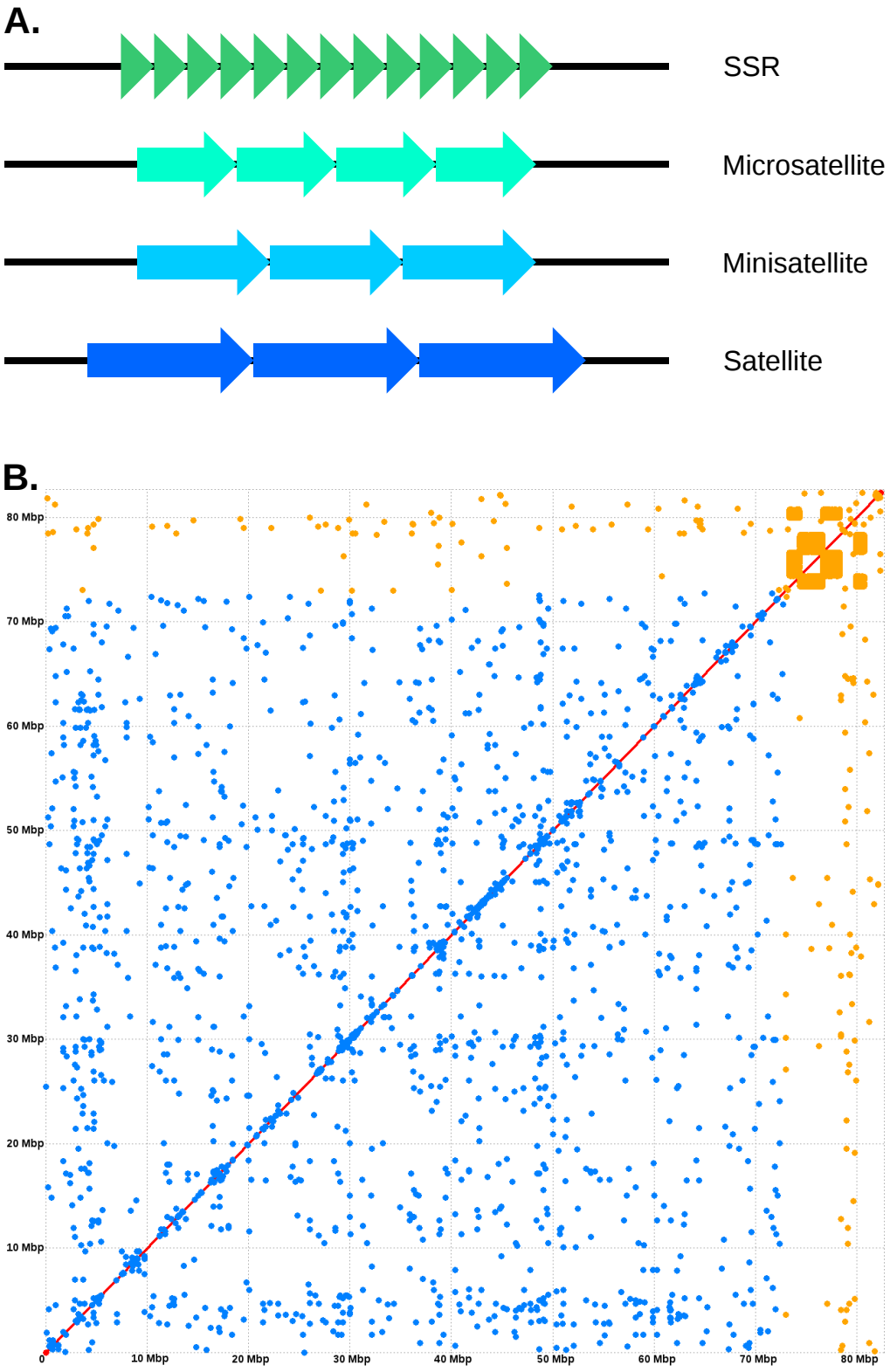


Figure 8 : Les répétitions en tandem
A. Les différents types de répétitions
Les répétitions en tandem correspondent à la répétition successive d'une séquence. L'unité de répétition est représentée par une flèche. Les répétitions en tandem sont caractérisées par la taille de l'unité de répétition.
SSR : 1 pb
Microsatellite : 2 à 10 pb
Minisatellite : 11 à 60 pb
Satellite : > 60 pb

B. Dot plot du chromosome Z : le Macrosatellite du génome du poulet
Dot plot représentant l'alignement du chromosome Z avec lui même. Les régions répétées sont représentées par les points bleus, exceptées celles faisant parti du macrosatellite qui sont représentées par les points jaunes. L'unité de répétition est d'environ 24 kpb et est répétée environ 830 fois.

génomes aviaires. Une étude a estimé que près de 6 000 gènes n'ont pas d'orthologues dans la version 4 du modèle [Schmid et al 2015]. Ces conclusions ont été remises en question. Ces gènes pourraient être non séquencés, car ils seraient constitués de séquences riches en GC [Friedman-Einat et al 2014, Lovell et al 2014, Hron et al 2015]. En effet, les méthodes de séquençage ne sont pas capables de lire les séquences riches en homopolymères et dont le GC % est élevé. La version 5 du modèle intègre 5005 gènes supplémentaires, soit 1 792 sur les chromosomes, 2 381 sur les Unl et 832 sur les Unp. Les gènes couvrent ainsi dans la version 5 49,96 % du génome et leur partie codante 6,63 %.

2.3.2. Les séquences répétées

Les séquences répétées sont les plus abondantes au sein des génomes eucaryotes. Leur quantité peut varier de 25 % dans le génome de *Drosophila melanogaster*, à 50 à 69 % chez l'homme et peut atteindre 90 % chez *Zea mays*. Il est possible de distinguer deux types de répétition dans l'ADN non codant en se basant sur leur structure, leur dispersion dans le génome et leur fréquence de répétition : les séquences répétées en tandem et les séquences moyennement répétées et dispersées.

2.3.2.1. Les séquences répétées en tandem

Les répétitions en tandem consistent en des répétitions consécutives d'un motif de taille variable (l'unité de répétition). Le nombre de répétitions peut être compris entre une dizaine à une centaine, voire plusieurs milliers. On distingue 5 types de répétitions en tandem que l'on retrouve dans différentes régions des chromosomes.

En se basant sur les caractéristiques de l'unité de répétition, il est possible de distinguer 5 catégories (Figure 8) :

- Les SSR ou répétitions en tandem de faible complexité dont l'unité de répétition correspond à une seule base, A ou T et G ou C ;
- Les microsatellites dont le motif de répétition a une longueur comprise entre 2 à 10 pb ;
- Les minisatellites dont le motif de répétition a une longueur comprise entre 11 à 60 pb [Brandström et al 2008] ;

- Les satellites dont le motif de répétition a une longueur supérieure à 60 pb ;

- Un autre type de répétitions en tandem résulte de la duplication en tandem de régions chromosomiques contenant ou non des gènes, comme celles par exemple codant pour les ARN ribosomiques, les immunoglobulines ou les récepteurs FCGRs. Le dernier type de répétition en tandem correspond à des séquences dont le nombre de répétitions est polymorphe et qu'on retrouve dans la littérature sous le nom de « copy number variations » (CNV) [Völker et al 2010]. Ces CNVs couvrent chez le poulet 9,4 % du génome et se répartissent sur 8 480 loci dont la taille varie de 1,1 à 268,8 kpb, avec une taille moyenne de 11,1 kpb [Yi G et al 2014].

Ces séquences hautement répétées sont majoritairement présentes dans les télomères et les centromères des chromosomes. Ainsi, chez le poulet, les répétitions $(TTAGGG)_n$ sont le principal élément constitutif des télomères [O'Hare et al 2009] et des ADN satellites leur sont souvent accolés. Une ou plusieurs familles d'ADN satellites peuvent être présentes dans les génomes eucaryotes dont les unités de répétition et les quantités diffèrent entre espèces.

On peut noter deux particularités concernant les ADN satellites dans le génome du poulet. Tout d'abord, une étude des micro-satellites [Primmer et 1997] a démontré que leur quantité et leur diversité semblent être plus faibles que celles des génomes de mammifères. La deuxième concerne le chromosome Z qui contient un satellite dont la taille de l'unité de répétition est unique. En effet, il est porteur d'un macro-satellite s'étendant sur 9 Mbp, dont la taille de l'unité de répétition est d'environ de 24 kpb et qui serait répétée 830 fois [Hori et 1996]. Cette région du chromosome Z est transcrite en ARN au cours de la phase de « lampbrush » qui est spécifique de l'ovogenèse [Krasikova et al 2012, Andraszek et al 2011].

Tableau 4 : Proportion d'éléments transposables dans les génomes eucaryotes
Source : Biémont C et al : What transposable elements tell us about genome organization and evolution: the case of Drosophila. Cytogenet Genome Res 2005, 110:25–34.

2.3.2.2. Les séquences répétées dispersées

2.3.2.2.1. Présentation des éléments transposables

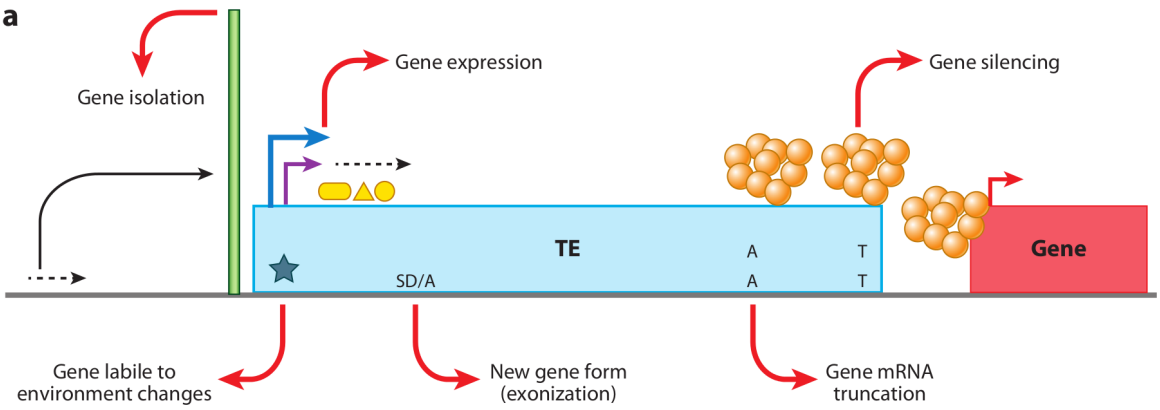
Les séquences répétées dispersées sont des segments d'ADN discrets, c'est-à-dire dont on identifie facilement les extrémités, qui sont dispersés au sein des génomes. La première personne à les avoir caractérisées fut Barbara McClintock en 1950. En étudiant l'expression différentielle de certains gènes chez le maïs, elle avait émis l'hypothèse que ces gènes étaient capables de se déplacer d'un locus à un autre et étaient sous l'influence de différents éléments régulateurs [McClintock 1950]. Ces séquences identifiées par Barbara McClintok étaient des éléments transposables (ET) qui par la suite ont été identifiés comme étant des transposons de type Ac/Ds et En/Spm.

Pierre Capy définit un ET comme étant « une séquence ADN capable de se multiplier au sein du génome d'une cellule » [Capy 2004], et dont la multiplication se fait via des intermédiaires ADN ou ARN et des enzymes de transposition de type transposase, intégrase et endonucléase. Il aura fallu près de 50 ans pour que la communauté des biologistes mesure l'importance des éléments transposables au sein des génomes. Présents dans tous les organismes en quantité très variable, y compris dans ceux des bactéries et les virus, leur part peut aller de 0,14 % chez *Tetraodon nigroviridis* à 99 % chez *Lilium speciosum* (Tableau 4) [Biémont et al 2005]. Au cours des deux dernières décennies, de nombreux travaux se sont accumulés et mettent en évidence qu'ils ont de nombreux impacts sur les génomes [Bire et al 2012].

2.3.2.2.2. L'impact des éléments transposables sur le génome

Les ETs sont des acteurs majeurs dans la variation de la taille des génomes chez les eucaryotes. Par exemple, la rapide augmentation de taille du génome de *Oryza australiensis*, une espèce de riz sauvage, résulte de la multiplication d'un élément transposable (ET) qui s'est accumulé (90 000 copies), provoquant ainsi un doublement de la quantité d'ADN chromosomique [Piegu et al 2006]. Ce phénomène a également été observé dans les espèces du genre *Gossypium*, dont le génome a triplé de taille à cause de l'amplification et de l'accumulation d'un ET [Hawkins et al 2011].

a



b

	POL II	POL III	Enhancer	Transcription factor	Environment response elements	Insulators	Heterochromatin	Poly(A) sites	Termination site (POL III)	Splice sites
								A A	T T	SD/A
LTR	✓		✓	✓	✓	✓	✓	✓		✓
LINE	✓		✓	✓				✓	✓	✓
SINE	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
DNA	✓		✓	✓	✓			✓		✓

Figure 9 : Possibilités pour que l'insertion d'un ET modifie l'expression d'un gène
a Impact possible d'un élément transposable sur l'expression d'un gène hôte. Les flêches rouges indique les effets putatifs
b Éléments régulateurs présents dans d'un chaque type d'élément transposable. Les éléments régulateurs sont décrits dans au moins une famille d'élément transposable de chaque type (LTR, LINE, SINE, DNA)
LTR : Long Terminal Repeat
LINE : Long Interpersed Element
SINE : Small Interpersed Element
DNA : DNA transposon

Source : Rebollo et al (2011). Transposable Elements: An Abundant and Natural Source of Regulatory Sequences for Host Genes. Annual Review of Genetics, 46(1), 120913153128008. http://doi.org/10.1146/annurev-genet-110711-155621

Comme suggéré par Keith Oliver et Wayne Greene dans leur théorie intitulée « TE-Thrust » [Oliver et al 2011], les ETs seraient également des moteurs de la diversité génétique permettant de créer de nouvelles espèces lors de phases de forte activité de multiplication et de transposition en liaison avec un relâchement du silençage épigénétique [Rebollo et al 2010].

Notre compréhension de l'impact de la mobilité des ETs sur les génomes qui les contiennent dépend de l'échelle de temps dans laquelle se situent les observations. La mobilité des ETs peut produire des mutations délétères et 75 maladies humaines ont été identifiées comme résultant de leur insertion dans un gène. Les ETs ont aussi été identifiés comme pouvant être le facteur déclenchant de tumeurs malignes [Callinan et al 2006, Belancio et al 2010]. Réciproquement, certaines mutations liées à leur insertion peuvent avoir un effet positif à l'échelle de l'évolution de l'hôte dans son environnement. Les ETs insérés à « proximité » de gènes peuvent agir sur l'expression de ces gènes en intégrant de nouveaux éléments régulateurs (promoteurs, enhancer, fixation de facteur de transcription, site de liaison pour des petits ARNs interférents…) dont ils sont porteurs (Figure 9) [Bourque et al 2008, Rebollo et al 2011]. Leur impact dépend aussi du contexte biologique et un type d'ET peut influencer son environnement génique de différentes façons. Par exemple, l'élément Alu, qui est présent sous forme d'environ un million de copies dans le génome humain, est capable de réguler l'expression de certains gènes par l'intermédiaire de certains de ses transcrits ARN ou de par ses capacités à modifier localement l'épissage. En effet, quand il est transcrit de façon « libre » (c.à.d. à partir de son propre promoteur), l'ARN de cet ET est capable d'interférer avec les transcrits de certains gènes de ménage lors d'un stress thermique. Dans d'autres situations, lorsque ce TE est inséré au sein d'un gène, il est source d'épissage alternatif, car il contient dans sa séquence 23 sites potentiels d'épissage (9 en 5' et 14 en 3') [Peaston et al 2004, Häsler et 2007, Ponicsan et al 2010, Walters et al 2009]. L'intérêt de la mobilité pourrait d'ailleurs être pour certains ETs un mécanisme faisant partie intégrante de la différenciation de certains tissus. Ainsi, les ETs Alu, L1 et SVA2 transposent activement au cours des dernières étapes de la différenciation des neurones chez les mammifères. Cette mobilité pourrait jouer un rôle important au cours de la neurogenèse en créant de la diversité génomique entre les neurones et ainsi participer à la complexité neuronale [Faulkner 2011].

Parfois, quelques copies d'ETs peuvent être cooptées par le génome hôte, « c.à.d. en quelque sorte domestiquées », et apporter de nouvelles fonctions. L'exemple le plus connu de domestication est le cas de la recombinaison V(D)J qui permet aux populations de lymphocytes B et T dans leur ensemble de produire des immunoglobulines capables de reconnaître n'importe quel antigène étranger. Le processus et les enzymes (RAG1 et RAG2) mis en jeu lors de cette recombinaison sont très proches de ceux de la transposition et dérivent d'un transposon *transib* [Kapitonov et al 2005]. Des cas de domestication d'ETs ont été localisés dans toutes les espèces et les systèmes génétiques qui en résultent peuvent avoir des formes très différentes qui demeurent fréquemment énigmatiques au niveau de leur fonction. C'est le cas par exemple du gène OVEX1 chez le poulet qui est dérivé d'un ancien rétrotransposon ayant perdu sa mobilité (perte des longues répétitions terminales aux extrémités). La fonction de cette protéine est encore inconnue, mais elle semble jouer un rôle important dans la physiologie des ovaires chez l'oiseau [Carré-Eusèbe et al 2009]. Des ETs domestiqués ont été également impliqués dans la structure et le fonctionnement des télomères et des centromères chez les eucaryotes [Pardue et al 2003, Wong et al 2004].

Les nombreuses implications des ETs dans les fonctions de l'hôte font désormais de ces éléments un sujet de recherche nécessaire à la compréhension du fonctionnement du génome, mais aussi de la physiologie et du développement de l'hôte. Cependant, il faut se préserver de toute exagération fonctionnaliste dans notre compréhension de l'impact des ETs sur leurs génomes hôtes. De nombreux ETs sont probablement insérés là où ils sont, participent aux polymorphismes intra et inter-spécifiques et n'ont aucune valeur sélective dans un contexte naturel. Dans certains cas, la domestication a profité de certains de ces polymorphismes a priori neutres. Par exemple, un ET à LTR (Hopscotch ; retrovirus endogène) inséré dans la région régulatrice du gène branched1 de la téosinte dans le génome du maïs agit comme un enhancer et provoque un changement de la régulation de l'expression du gène, ce qui modifie l'architecture florale de la plante [Studer et al 2011]. Au cours de la domestication du maïs, un polymorphisme portant sur l'architecture florale résultant de cet allèle muté du gène branched1 a été exploité pour inverser la position des fleurs mâles et femelles sur les pieds de maïs.

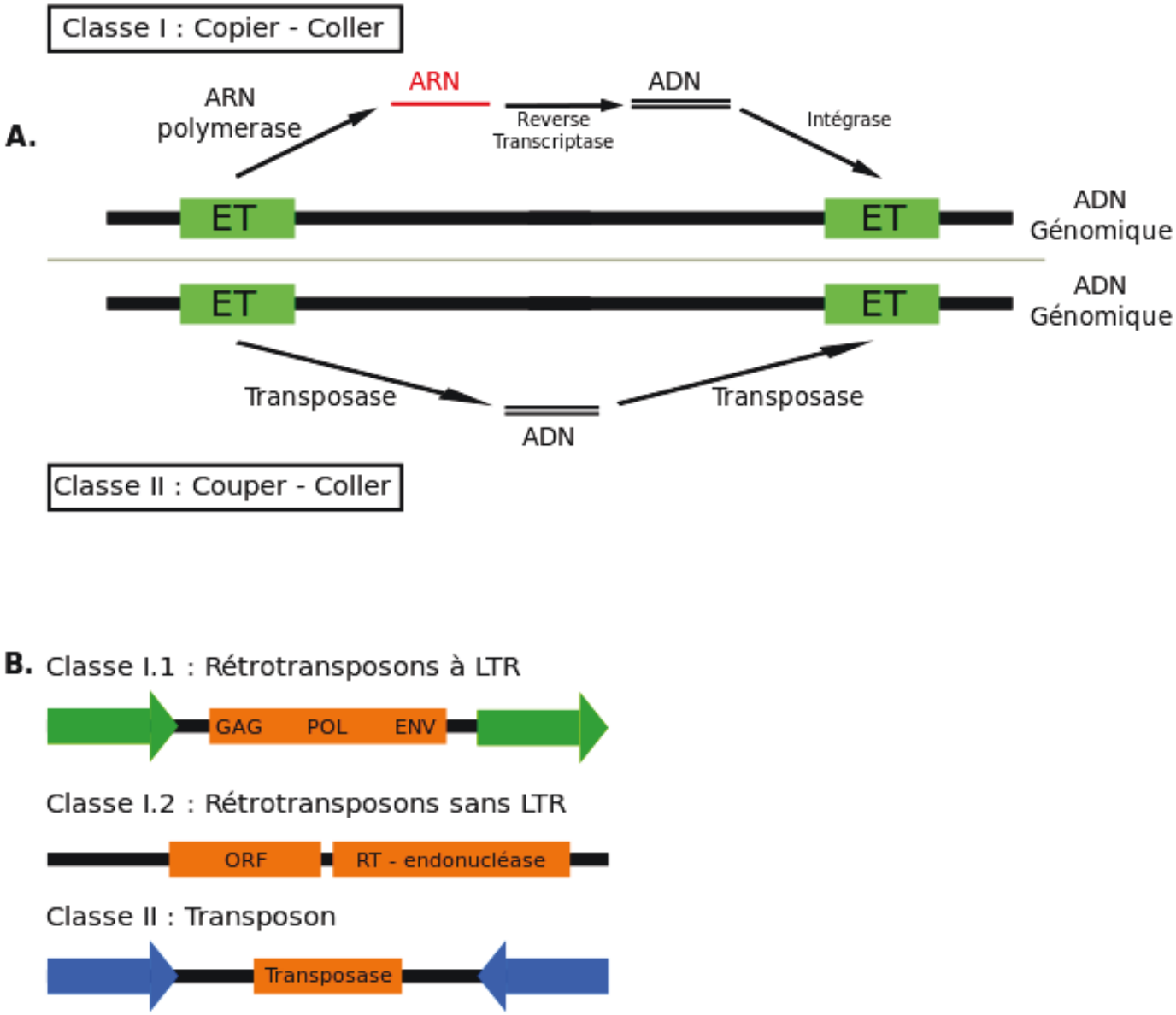


Figure 10 : Classification des éléments transposables par Finnegan

A. Classes d'éléments transposables de la classification de Finnegan
Classe I : La séquence nucléique de l'élément est en transcrit en un intermédiaire ARN, qui est rétro-transcrit en ADN double brin par une reverse transcriptase. L'ADN double brin est ensuite intégré à un nouveau locus du génome à l'aide d'une intégrase.
Classe II : La séquence nucléique de l'élément transposable est excisée par une transposase produisant un intermédiaire ADN. Cet intermédiaire de l'élément est intégré à un nouveau locus du génome par cette même transposas

B. Sous-classes de la classification de Finnegan
Classe I.1 : Les éléments de classe I.1 ont une structure proche des rétrovirus. Ils possèdent deux LTR à leurs extrémité La séquence interne code pour GAG, POL et parfois ENV.
Classe I.2 : Les éléments de classe I.2 sont dépourvus de LTR. Ils possèdent deux cadres ouverts de lecture. Le premier code pour une reverse transcriptase et une endonucléase, la fonction du deuxième n'est pas connue
Classe II : La classification de Finnegan n'instaure pas de sous-classe pour la classe II. Leurs extrémités possèdent des T et la partie interne code pour une transposase.

2.3.2.2.3. Les classifications

En plus d'être présents dans tous les organismes, les ETs affichent une très grande diversité de structure et de mécanismes de mobilité dans les génomes. En se basant principalement sur ces deux caractéristiques, plusieurs classifications ont été mises en place depuis 1989.

Classification dite de Finnegan

La première classification fut proposée en 1989 par Finnegan [Finnegan 1989]. Elle se fonde sur les mécanismes de transposition des ETs et répartit les éléments en deux classes (Figure 10). Les éléments de classe I utilisent un système de « copier-coller » en employant un intermédiaire de transposition ARN :

- l'ET est transcrit en ARN par une ARN polymérase qui est ensuite transcrite en un ADN double brin via l'action d'une reverse-transcriptase et d'une RnaseH ;
- Le matériel génétique est ensuite inséré dans le génome à l'aide d'une intégrase [Coffin et al 1997].

La classe I se divise en 2 sous-classes, la classe I.1 pour les éléments transposables du type rétrotransposon LTR et la classe I.2 pour les séquences de type rétrotransposon sans LTR. Bien que Finnegan suggère l'existence des SINEs en 1992 [Finnegan 1992], ils ne sont pas décrits dans cette première classification.

Les éléments de classe II utilisent un système de « couper-coller » en employant un intermédiaire ADN :

- L'ET est excisé du génome par la coupure d'un ou deux brins par une transposase ;
- Il est ensuite réintégré au sein du génome par cette même transposase.

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

Structural features

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

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

Protein coding domains

AP, Aspartic proteinase — APE, Apurinic endonuclease — ATP, Packaging ATPase — C-INT, C-integrase — CYP, Cysteine protease — EN, Endonuclease

ENV, Envelope protein — GAG, Capsid protein — HEL, Helicase — INT, Integrase — ORF, Open reading frame of unknown function

POL B, DNA polymerase B — RH, RNase H — RPA, Replication protein A (found only in plants) — RT, Reverse transcriptase

Tase, Transposase (* with DDE motif) — YR, Tyrosine recombinase — Y2, YR with YY motif

Species groups

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

Figure 11 : Classification des éléments transposables par Wicker et al

Source : Wicker T et al: A unified classification system for eukaryotic transposable elements. Nat Rev Genet 2007, 8:973–82.

Les éléments transposables de la classe II sont constitués de répétitions inversées terminales (« Terminal Inverted Repeat », TIR) et contiennent en leur centre au moins un gène qui code une enzyme de type transposase nécessaire à la mobilité.

La découverte de nouveaux transposons et la mise en évidence de leur très grande diversité ont obligé les chercheurs à mettre à jour la classification de Finnegan. En 2007 et 2008, deux nouvelles classifications sont proposées par Thomas Wicker [Wicker et al 2007] et les propriétaires de « Repbase », Vladimir Kapitonov et Jerzy Juka [Kapitonov et al 2008]. Ces deux initiatives tentent de réaliser une classification unifiée des éléments transposables.

Classification dite de Wicker

La classification dite de Wicker se base sur celle de Finnegan en conservant le système de classe I et classe II et la complète en rajoutant plusieurs niveaux d'organisation (Figure 11). Les différents niveaux de classification se basent sur des caractéristiques telles que la nature de l'intermédiaire de transposition (ADN ou ARN ; simple brin / double brin), le nombre de brins coupés lors de la transposition, l'organisation structurale, les répétitions internes aux extrémités, la conservation de séquence et/ou leur phylogénie.

La classe I regroupe les éléments génétiques mobiles utilisant un intermédiaire ARN. Elle est composée de 5 ordres qui sont établis en fonction de la structure des éléments au niveau des répétitions internes et les enzymes codés par chaque élément :

- LTR : ces éléments ont une structure similaire aux rétrovirus. Ils possèdent de longues répétitions terminales (LTR) et contiennent au moins trois types de gènes qui codent pour :
    - une GAG (Group AntiGens) qui est une polyprotéine structurale du virus, contenant les domaines de matrice (MA), capside (CA) et nucléocapside (NC),
    - une POL qui est une polyprotéine comprenant une « Aspartic Protease », une intégrase, une Reverse Transcriptase et une Rnase H. L'intégrase est de type DDE,

c’est-à-dire une enzyme dont le site catalytique comporte deux acides aspartiques (D) et un acide glutamique (E),

- ENV pour certains, qui est une glycoprotéine d’enveloppe,

- DIRS (*Dictyostelium* Transposable Element Sequence) : ces éléments codent des protéines similaires aux LTR, ENV excepté. Ils se différencient des LTR par l’intégrase qui est du type recombinase à tyrosine. Leurs extrémités peuvent soit être des répétitions terminales inversées (TIR) ou des répétitions directes sans pour autant que celles-ci puissent être considérées comme des LTRs ;

- LINE (Long INterpersed Elements): ces éléments peuvent avoir un ou deux cadres de lecture ouverts. L’un d’entre eux code pour une protéine contenant une endonucléase et une reverse transcriptase. Leurs extrémités n’ont pas de répétitions de type TIR ou LTR ;

- PLE (Penelope-Like Elements) : ces éléments sont structurellement similaires aux LINEs, mais ils codent pour une endonucléase et une reverse transcriptase qui ont des origines différentes de celles des LINEs. Leurs extrémités sont porteuses de structures répétées que certains auteurs assimilent « maladroitement » à des LTR ;

- SINE (Small INterpersed Elements) : Ces éléments de petites tailles sont dépourvus de répétition terminale et ne codent pas de protéine. Ils peuvent contenir un promoteur pour l’ARN polymérase III (pol III) et dérivent évolutivement en général d’ARN structuraux du génome hôte (ARN de transfert, 7SL ARN, etc.), ou de chimères de différents ARNs. Ils utilisent la machinerie de transposition des LINEs pour se multiplier et se déplacer.

La classe II regroupe les éléments transposables utilisant un intermédiaire ADN et comporte deux sous-classes. La première intègre les transposons qui couperaient les deux brins d’ADN lors de la transposition et comporte deux ordres :

- TIR (Terminal Inverted Repeat) : ces éléments possèdent des répétitions terminales inversées et contiennent au moins un gène codant pour une transposase ;

- Cryptons : ces transposons sont similairement organisés au niveau de leur séquence mais ils contiennent un gène codant non pas une transposase, mais une recombinase à tyrosine.

La sous-classe 2 regroupe les transposons qui ne couperaient qu'un seul brin d'ADN lors de la transposition. Elle comporte deux ordres :

- Les Helitrons : ces éléments utilisent un mécanisme de réplication circulaire pour transposer [Kapitonov et al 2007] ;

- Polinton-Marverick : ces ETs codent pour une DNA polymerase B, intégrase rétrovirale, une cystéine protéase et une ATPase. Ils possèdent des répétitions terminales inversées et leur insertion induit la création de TSDs. L'originalité de ces ETs est qu'ils possèdent des introns.

L'originalité de la classification de Wicker est qu'elle essaie d'intégrer les éléments non autonomes tels que :

- les SINEs ;

- les LARDs (LArge LTR Retrotransposon Derivative) qui sont de grands rétrotransposons à LTR issus de la perte des gènes nécessaires à leur mobilité ;

- les TRIMs (Terminal-repeat Retrotransposons In Miniature) sont des équivalents des LARDs, mais de petite taille ;

- les MITEs (Miniature Inverted-repeat Transposable Element), qui proviennent de la délétion des régions internes provenant de transposons de type TIR ;

- les SNACs (Small Non-Autonomous CACTA), qui sont des MITEs dérivés d'éléments CACTA, des transposons de type TIR [Wicker et al 2003].

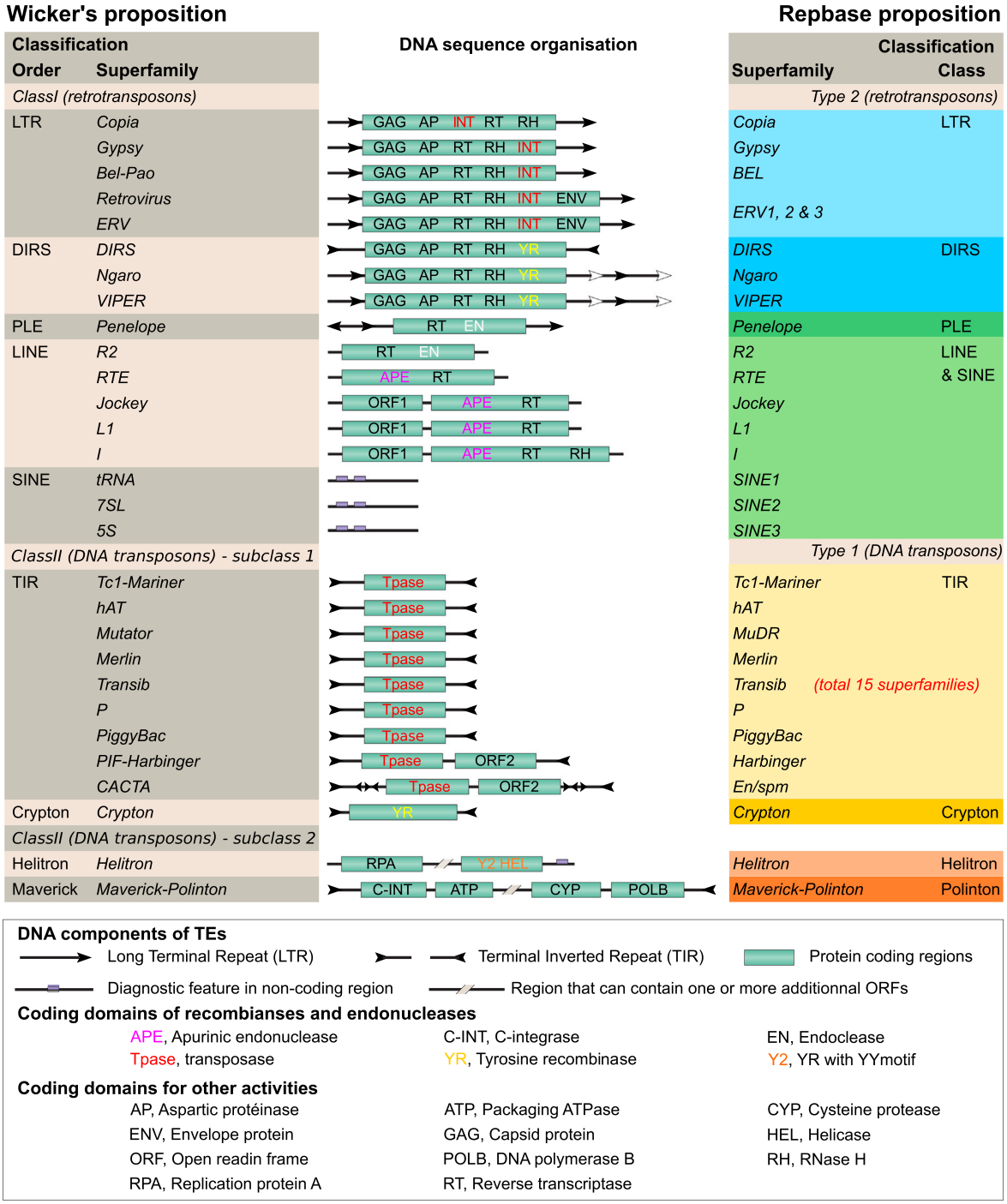


Figure 12 : Correspondance entre la classification de Wicker et de Repbase
Source : Piéguet al (2015). A survey of transposable element classification systems – A call for a fundamental update to meet the challenge of their diversity and complexity.
Molecular Phylogenetics and Evolution, 86, 90–109. http://doi.org/10.1016/j.ympev.2015.03.009

Classification dite de Repbase

Comme indiqué précédemment, Repbase est à l'origine une banque de données de séquences nucléiques d'ETs dont la première version a été publiée en 1990. Elle est constituée d'une collection de séquences consensus et fragments d'ETs. Une séquence consensus est une séquence moyenne créée à partir des copies d'un ET présentes dans un génome. Au fil des années, des séquences répétées provenant d'espèces animales et végétales ont été accumulées dans cette banque. Au milieu des années 90, Repbase est devenue gratuitement accessible aux académiques et est considérée comme la base de séquences répétées de référence pour l'annotation des génomes. La version actuelle, Repbase Update, comporte plus de 20 000 séquences. La classification de Repbase (Figure 12) utilise comme premier niveau le mécanisme de transposition et forme sept classes réparties en deux types. Le Type I regroupe les transposons qui ont un mode de transposition de type :

- « copier-coller » utilisant une transposase de type DDE ou DDD ;
- réplication circulaire ;
- « copier-coller » sans intermédiaire ADN.

Le Type II rassemble deux classes de rétrotransposons sans LTR, les PLEs et les LINEs (SINEs compris), et deux classes de rétrotransposons dits à LTR, les DIRS et les LTRs. Ces classes sont ensuite subdivisées en super familles et familles qui se distinguent en fonction des enzymes impliquées dans la transposition, leur similarité structurale et leur similarité de séquence.

Bien qu'elles soient très usitées par la communauté scientifique pour définir les ETs, ces classifications sont néanmoins incomplètes. Créées principalement par des chercheurs issus du monde des eucaryotes, elles sont axées sur les caractéristiques structurales et ne prennent pas en compte l'ensemble des mécanismes de transposition. Enfin, certains éléments mobiles ont été oubliés. En parallèle de ces classifications émerge dès 2003 une ébauche d'une classification sur la base d'une revue concernant les mécanismes de transposition [Curcio et al 2003]. Cette classification, dite de Curcio et Derbyshire, est basée sur les caractéristiques et les modes d'action des enzymes qui sont impliqués dans les différents

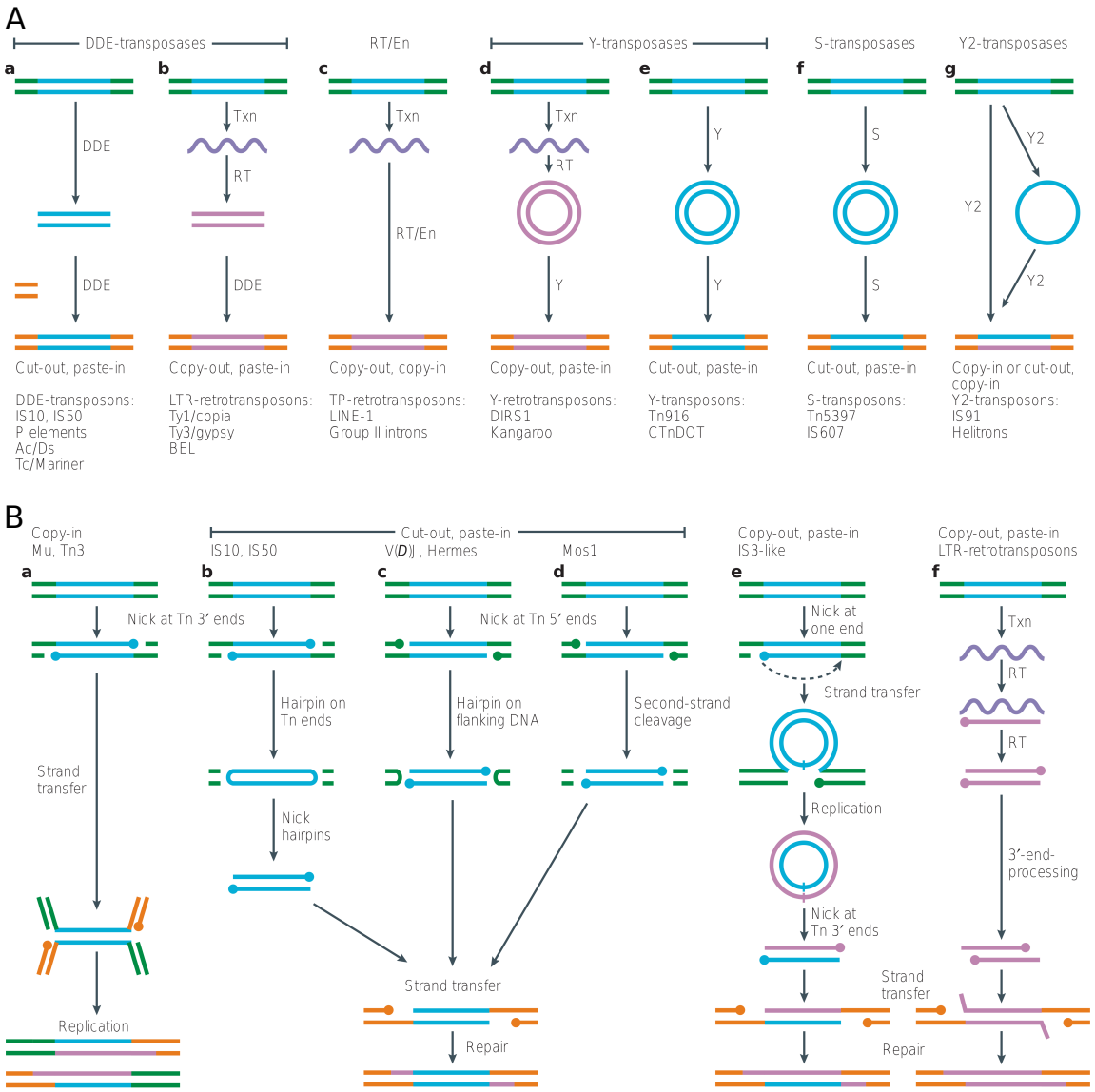


Figure 13 : Proposition de classification de Curcio et Derbyshire
A. Les transposons se déplacent de manières différentes.
Cinq familles de protéines dictent les différentes voies de transposition: DDE transposases, reverse transcriptase/endonucleases (RT/En), tyrosine (Y)-transposases, serine (S)-transposases, rolling-circle (RC)- or Y2-transposases.
Les transposons (bleus) peuvent être soit «cut-out» ou «copied-out» du site donneur (vert).
a. La plupart des transposons à tranposase-DDE peuvent s'exciser du génome afin de générer un intermédiaire linéaire, qui est le substrat pour l'intégration dans le site cible (orange).
b. Les rétrotransposons «copy-out» transposent par reverse-transcription (RT) de l'ARN (poupre) généré par la transcription (Tnx). Les rétransposons à LTR créent un ADNc double brin à partir de l'ARN (le rose représente l'ADN nouvellement répliqué) et l'intègre dans le génome en utilisant la DDE transposase.
c. Les TP-rétrotransposons utilisent une reverse-transcriptase (RT) pour copier directement leur ARN dans le site cible préalablement coupé par une endonucléase codée par le rétrotransposon (En).
d. Les Y-rétrotransposons semblent former un intermédiare ADNc circulaire par reverse-transcription. Une transposase-Y intègre l'élément au niveau du site accepteur.
e. et **f.** Les Y et S transposons codent une transposase à tyrosine ou à serine, qui médie l'excision du transposon pour former un intermédiaire circulaire. Le transposon est intégré au génome par la même série de réactions enzymatiques, mais dans l'ordre inverse.
g. Les transposons Y2 « collent » un brin du transposon dans le site cible et l'utilise comme modèle pour la réplication de l'ADN. Deux modèles ont été proposés pour la transposition des Y2.
Des représentants de chaque type de transposon sont énumérés en dessous chaque voie de transposition.

B. Les DDE transposases excisent l'ADN leurs transposons par différents mécanismes.
Les processus sont présentés en parallèle pour souligner les différentes voies de DDE transposition.
a. les transposases des transposons Mu et Tn3-like coupent leurs transposons en 3' et liguent ceux-ci au site cible. Le transposon est répliqué dans le site cible en utilisant les 3'OH pour démarrer la réplication.
b-d. Dans le processus de « cut-out, paste-in » les deux brins du site donneur sont coupés, amenant ainsi une insertion simple.
e. Les transposons IS3-like coupe à la séquence à une seule extrémité 3'. Le 3' OH résultant attaque le même brin à l'autre extrémité du transposon (ligne pointillée avec la flèche). Il semblerait que la réplication permette de régénérer l'ADN au niveau du site donneur et de libérer le transposon sous forme d'ADN double brin ciculaire. Une seconde coupure permet de linéariser l'intermédiaire ciculaire.
f. Les rétrotransposons à LTR créent une copie par une transcription (Tnx) de leur séquence suivie d'une reverse-transcription (RT). Les extrémités 3' de l'ADNc possèdent un dinucléotide CA terminal ou elles sont traitées par la DDE transposase pour généré les dinucléotides qui seront ligués au site accepteur.
Toutes les transposases DDE insérent le transposon au niveau d'un site de coupure ayant des extrémités cohésives.
La réparation de l'ADN au niveau du site d'insertion par les enzymes de la réplication à pour effet de générer des "Target Site Duplication" (TSD) à chaque extrémité (duplex rose / orange).
Les lignes bleues représentent l'ADN du transposon; les lignes vertes représentent l'ADN du site donneur; les lignes orange représentent l'ADN site cible; les lignes rose représentent l'ADN nouvellement répliqué; lignes pourpres représentent l'ARN; les cercles pleins aux extrémités de l'ADN indiquent des groupes 3'OH libres.

Source : Curcio MJ, Derbyshire KM: The outs and ins of transposition: from mu to kangaroo. Nat Rev Mol Cell Biol 2003, 4:865–77.

modes de transposition identifiés à cette époque. Ils définissent ainsi 5 classes d'éléments génétiques mobiles (Figure 13) :

- les ETs à transposases DDE ;
- les ETs à réplication circulaire ou Y2-transposases ;
- les ETs tyrosines transposases ou Y1-transposases ;
- les ETs sérines transposases ;
- les target-primed rétrotransposons.

Ils précisent également que certains ETs ayant la capacité de se déplacer dans le génome n'ont pas été inclus dans les précédentes classifications. Les introns groupe I et les introns groupe II sont capables de s'insérer ou de s'exciser de leurs gènes hôtes. Il existe aussi les intéines, des introns traduits en protéines capables de réintégrer leur séquence dans le génome. Il y a aussi les introners et les IStrons, des séquences mobiles des caractéristiques des introns.

Les lacunes de ces classifications montrent la grande complexité et diversité des ETs et amènent à s'interroger sur leur consistance. Établir un recensement des ETs est un challenge complexe et laborieux, qui n'est rien à côté du défi qui serait de faire accepter une remise en cause significative de la conception et de la classification des ETs par la communauté scientifique. Des initiatives essaient d'établir un nouveau consensus en proposant de créer une classification prenant en compte l'ensemble des éléments mobiles eucaryotes et procaryotes, introns et dérivés chimériques inclus. Les critères de classification principale reposent sur le mécanisme de transposition et les enzymes utilisées au cours de la transposition [Piégu et al 2015], et l'approche analytique essaie d'être extrêmement vigilante quant aux torsions de compréhension liées aux convergences évolutives qui peuvent avoir eu lieu sur ces mécanismes moléculaires.

Tableau 5 : Inventaire des éléments transposables dans le génome du poulet par Wicker

Auteurs	Méthode	Éléments Transposables (%)	Satellite SSR (%)
Arthur, R. R. et al 1978 Epplen, J. T. et al 1978	Cinétiques de réassociation	20	10
Wicker et al 2005	Cinétiques de réassociation + séquençage	4,3	3 – 4
Institut for System Biology 2011	RepeatMasker + Repbase	9,74	1,73

Tableau 6 : Inventaire des éléments transposables dans le génome du poulet par Wicker
Source: Wicker et al : The repetitive landscape of the chicken genome. Genome Res 2005, 15:126–36.

Name	Class[a]	Copy Number	Total bp
CR1	LINE	96,230	37,160,469 (3.10%)
Galluhop	Class 2/*Mariner*	13,729	6,140,519 (0.51%)
Birddawg	LTR/*gypsy*	7,404	2,697,928 (0.22%)
Kronos	LTR/*gypsy*	4,961	3,021,541 (0.25%)
Hitchcock	LTR?	3,324	811,951 (0.07%)
Charlie	Class 2	2,292	1,203,639 (0.10%)
Soprano	LTR	1,362	768,648 (0.06%)
Telomeric	tandem	362[b]	252,835 (>0.1%)
Total		129,351	52,057,530 (4.31%)

The copy number indicates the number of identified repeat units in the first draft of the publicly available chicken genome sequence. The number in parentheses is the total base-pair count for each repeat in percent of the whole genome.
[a]A question mark indicates that the classification is uncertain.
[b]Copy number for telomeric repeats refers to the number of identified arrays of multiple tandem-repeated units.

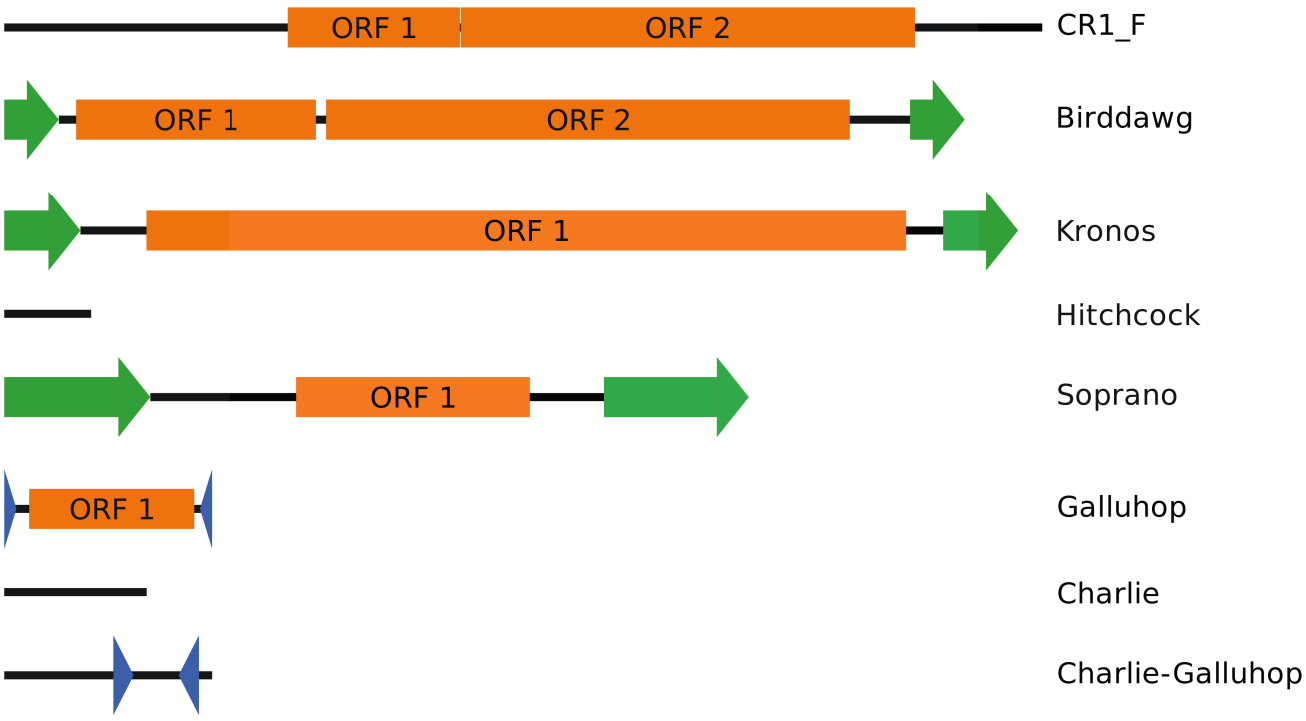


Figure 14 : Représentation des éléments transposables du poulet les plus nombreux

Terminal Inverted Repeat
(Répétition terminal inversée)

Long terminal repeat
(Longue répétition terminale)

Séquence nucléique

Open reading frame
Cadre ouvert de lecture

2.3.2.2.4. Les éléments transposables du génome du poulet

La première évaluation du contenu total en séquences répétées a été effectuée à l'aide de méthodes de cinétique de ré-association d'ADN génomique. La proportion en séquence hautement répétée (en tandem) a alors été évaluée à 10 % du génome et la proportion en séquence répétée dispersée (ETs) à 20 % du génome [Arthur et al 1977, Epplen et al 1978] (Tableau 5). Lors du premier séquençage, l'International Chicken Genome Sequencing Consortium (ICGSC) a évalué la part des ETs à 9,4 % et celle des répétitions en tandem à 0,1 %. Un an après la publication de la première version du génome, Thomas Wicker utilise des cinétiques de ré-association de l'ADN pour sélectionner des BACs riches en ETs pour ensuite les séquencer [Wicker et al 2005]. De cette manière, il propose un inventaire des ETs, et estime leur quantité à 4,3 % du génome du poulet (Tableau 6).

Les ETs les plus abondants dans ce génome sont les CR1s (Figure 14) qui sont des rétrotransposons de l'ordre des LINE selon la classification Wicker. Thomas Wicker estime leur nombre à 96 230 copies couvrant 3,10 % du génome. Ils ne comportent pas de répétitions internes et ont une taille comprise entre 4,5 kbp et 6 kpb. Ils possèdent deux ORFs, un dont la fonction reste inconnue alors que le second code pour une reverse transcriptase et une endonucléase [St John et al 2008, Kazazian et al 2004]. Selon Repbase, il y a 22 types de CR1s différents.

Thomas Wicker répertorie également 4 ETs de type rétrotransposons de type LTR : les éléments Birddawg, Kronos, Hitchcock et Soprano respectivement présents en 7 404, 4 961, 3 324 et 1 362 copies. Ce sont des éléments dont la structure se rapproche de celle des rétrovirus. Ils possèdent de longues répétitions terminales (LTR) à leurs extrémités et leur structure interne comporte trois ORFs codant respectivement une GAG, une POL et un gène ENV dans le cas des ETs de type rétrovirus. Beaucoup moins nombreux que les LINEs, ces ETs ne couvrent que 0,6 % du génome. Leurs tailles sont de 543 pb pour Hitchcock, 4,6 kpb pour Soprano (avec des LTR de 913 pb) et 6 kpb pour Birddawg (avec des LTR de 341 pb) et Kronos (avec des LTR de 476 bp) [Repbase update 19.10]. Il faut noter que Wicker classe Hitchcock comme un rétrotransposon de type LTR. Cependant, sa très petite taille indique qu'il correspondrait plutôt à un solo LTR. En effet, seul est présent un fragment de la taille d'un LTR dans lequel on ne détecte aucune trace de capacité de codage pour GAG, POL ou

ENV. Ce type de séquence serait issu d'une transposition avortée d'un élément complet ou d'une recombinaison homologue entre les deux LTR du rétrotransposon.

Les deux derniers éléments répertoriés par Thomas Wicker sont la classe des transposons de type TIR. Ils sont porteurs de répétitions terminales inversées et leur structure interne contient un ORF codant pour une transposase. Les transposons Charlie (1 298 pb) et Galluhop (2 329 pb) appartiennent respectivement aux familles *hAT* et *IS630-Tc1-mariner*. Ils sont présents en 2 292 et 13 729 copies pour une couverture cumulée du génome de 0,61 %. Thomas Wicker évoque également l'existence d'une version de ces éléments, Charlie-Galluhop, issue de l'insertion d'un transposon Galluhop dans un transposon Charlie, mais il ne précise pas le nombre de copies.

La dernière annotation en date des éléments transposables a été faite en 2011 par l'Institut for Systems Biology (ISB) en utilisant le logiciel RM [RepeatMasker Open-4.0] et sa propre banque de séquences dérivées de Repbase. RM est actuellement le programme de référence utilisé pour annoter les séquences répétées lors de la publication d'un nouveau génome animal. Dans le cas de la version 4 du génome du poulet, RM a annoté 11,47 % du génome comme étant des répétitions, dont 9,74 % d'ETs. Les LINEs sont représentés par 34 lignées de CR1s, 4 de L2, 1 RTE, 1 PLE, 2 UCON2 et 9 X_LINE. Dans cette annotation, 216 237 copies d'ETs sont présentes et couvrent 6,87 % du génome. Les rétrotransposons à LTR sont représentés par deux grandes familles, les GGLTR et les GGERV rassemblant 44 séquences consensus différentes. Cependant, ce nombre n'est pas représentatif du nombre d'espèces de LTRs présent dans ce génome. En effet, la séquence des ETs à LTR dans Repbase est fragmentée en deux séquences, l'une correspondant à la séquence des LTRs et l'autre à la séquence interne. Les rétrotransposons à LTR sont présents en 41 063 copies et couvrent 1,87 % du génome. Les transposons (TIRs) présents sont majoritairement du type Chompy, Mariner et Harbinger. Ils sont présents en 28 522 copies et couvrent 0,87 % du génome. Le premier séquençage du génome du poulet avait mis en évidence l'absence de SINE. Cependant, l'annotation RM comprend 7 SINEs présents en 6476 copies qui couvrent 0,08 % du génome. Cette annotation arbore également une série de 42 séquences répétées dont le type est inconnu. Ils sont présents en 3 332 copies et couvrent 0,05 % du génome.

Tableau 7 : Proportions d'ETs dans les génomes d'espèces aviaires évaluées par le Phylogenomic Project

Species	LINE	SINE	LTR	DNA	RC	Unknown	Other	Total
Picoides pubescens	18,2015	0,04604	0,89066	0,16695	0,00429	2,83735	0,00046	22,1473
Gallus gallus	6,0109	0,07659	1,65347	1,00501	0,0078	1,0665	0,0003	9,82058
Taeniopygia guttata	3,78953	0,05858	4,10936	0,32025	0,01772	1,38779	0,00073	9,68395
Cuculus canorus	7,83916	0,07752	0,66931	0,2746	0,00573	0,581	0,00031	9,44763
Colius striatus	6,53918	0,09797	2,19283	0,19214	0,00378	0,38859	0,00051	9,415
Melopsittacus undulatus	6,4899	0,07715	1,96803	0,20069	0,00752	0,44785	0,00016	9,19131
Chaetura pelagica	5,28185	0,10761	0,89502	0,18882	0,00352	2,56957	0,00056	9,04695
Acanthisitta chloris	6,3824	0,10215	1,45738	0,21059	0,00967	0,56122	0,00023	8,72363
Tauraco erythrolophus	2,75721	0,08774	1,80381	0,15721	0,00604	3,82672	0,00016	8,63889
Apaloderma vittatum	5,96761	0,1176	1,30969	0,22583	0,00513	0,81509	0,00021	8,44116
Antrostomus carolinensis	5,40314	0,1179	1,83793	0,32737	0,01747	0,53289	0,00032	8,23702
Geospiza fortis	3,64772	0,05835	3,3707	0,31177	0,03877	0,79969	0,00061	8,22762
Calypte anna	5,61961	0,06831	1,22855	0,2141	0,0099	0,90731	0,00032	8,0481
Meleagris gallopavo	5,39961	0,05268	1,10563	0,82249	0,00478	0,51645	0,00016	7,90179
Ophisthocomus hoazin	4,69201	0,1114	1,29602	0,15839	0,00787	1,63178	0,00022	7,89768
Merops nubicus	5,01213	0,06999	1,29546	0,13936	0,00535	1,25607	0,00028	7,77865
Mesitornis unicolor	4,61872	0,09126	1,38263	0,38166	0,00925	1,03021	0,00041	7,51414
Pelecanus crispus	3,93807	0,15337	1,87329	0,21447	0,00565	1,26833	0,00019	7,45337
Phaethon lepturus	3,91321	0,11787	1,70552	0,22142	0,0048	1,47534	0,0002	7,43835
Corvus brachyrhynchos	3,72969	0,07357	2,42901	0,21861	0,01534	0,8994	0,0004	7,36603
Podiceps cristatus	4,80367	0,10176	1,59644	0,20172	0,00562	0,60107	0,00016	7,31045
Columba livia	4,18253	0,08584	0,75945	0,34731	0,0077	1,8677	0,00035	7,25089
Charadrius vociferus	4,53428	0,12823	1,1157	0,19605	0,0065	1,0524	0,00031	7,03347
Egretta garzetta	3,91707	0,1226	1,41923	0,24157	0,00577	1,21957	0,00025	6,92605
Eurypyga helias	4,60858	0,09861	1,59535	0,14741	0,00484	0,4644	0,00017	6,91935
Haliaeetus leucocephalus	2,00515	0,1732	1,89425	0,22054	0,00317	2,59326	0,00025	6,88982
Manacus vitellinus	4,43477	0,08049	1,08186	0,24633	0,00877	0,72441	0,00033	6,57695
Nestor notabilis	4,59637	0,10459	1,3174	0,18054	0,00478	0,36856	0,00015	6,5724
Leptosomus discolor	2,93288	0,11669	1,32311	0,19283	0,00536	1,88262	0,00015	6,45363
Chlamydotis macqueenii	3,97021	0,16962	1,4033	0,23258	0,00454	0,57444	0,00024	6,35493
Phalacrocorax carbo	3,94906	0,15573	1,29058	0,20761	0,00412	0,6245	0,00026	6,23186
Nipponia nippon	3,68685	0,13068	1,22028	0,28937	0,0065	0,8274	0,00031	6,16138
Balearica gibbericeps regulorum	3,35137	0,13632	1,51228	0,23695	0,00587	0,83417	0,00017	6,07712
Pygoscelis adeliae	3,31003	0,19502	1,32086	0,25858	0,00278	0,94761	0,00025	6,03513
Buceros silvestris rhinoceros	3,61909	0,07569	1,04809	0,1602	0,00592	1,08907	0,0001	5,99816
Anas platyrhynchos domestica	4,0522	0,09728	1,09549	0,20375	0,00676	0,39224	0,00028	5,848
Pterocles gutturalis	3,46226	0,09022	1,36026	0,17044	0,0066	0,66503	0,00013	5,75494
Phoenicopterus ruber ruber	2,68686	0,14818	1,03641	0,23324	0,00541	1,49416	0,00016	5,60442
Falco peregrinus	3,0868	0,14939	1,26518	0,28248	0,00437	0,70956	0,00031	5,49809
Aptenodytes forsteri	2,40622	0,19808	1,16966	0,26165	0,00294	1,457	0,00024	5,49579
Tyto alba	2,63807	0,12622	1,78848	0,18667	0,01436	0,73733	0,00019	5,49133
Cariama cristata	3,51351	0,17871	0,90914	0,19647	0,00352	0,68527	0,00018	5,4868
Haliaeetus albicilla	2,55429	0,13944	1,70885	0,18662	0,00525	0,77347	0,00015	5,36808
Fulmarus glacialis	2,86062	0,17656	1,18551	0,21739	0,00646	0,87302	0,00016	5,31971
Gavia stellata	3,16569	0,13876	0,71228	0,21764	0,0056	0,84803	0,00014	5,08814
Cathartes aura	2,20702	0,16981	1,04973	0,19008	0,0045	0,91728	0,00011	4,53853
Struthio camelus	2,87702	0,18499	0,16645	0,35945	0,00781	0,89673	0,00016	4,49261
Tinamus guttatus	2,73101	0,08748	0,30251	0,32546	0,00995	0,65253	0,00034	4,10928

Source : Zhang G et al : Comparative genomics reveals insights into avian genome evolution and adaptation. Science (80-) 2014, 346:1311–1320.

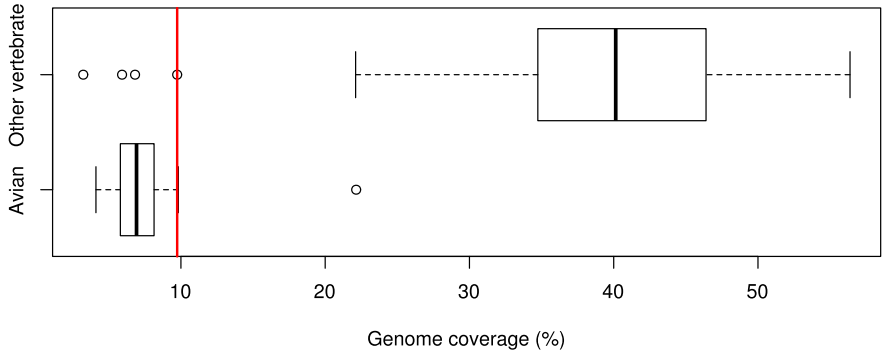


Figure 15 : Boxplot du taux de couverture des éléments transposables dans les génomes des espèces aviaires et les autres génomes vertébrés

La ligne verticale rouge indique la taille du génome du poulet

2.4. Conclusion sur les éléments transposables chez le poulet

L'évaluation de la proportion d'ETs présents dans les modèles de génome de la poule rouge de jungle a augmenté au cours des années, de 9,3 % à 9,74 % et celles du total en répétions de 9,4 % à 11,74 %. Ces proportions demeurent bien moindres que celles décrites dans les génomes vertébrés qui dépassent les 30 % (Figure 15, Tableau 7). La quantité de répétitions annotées ne correspond pas aux estimations faites directement sur le matériel biologique. En effet, si on considère les résultats des cinétiques de ré-association qui estiment le contenu total en répétition du génome du poulet à 30 %, ce pourcentage est bien supérieur à la proportion de séquences répétées dispersées dans le modèle Galgal4. Si on calcule cette proportion en tenant compte des répétitions absentes (télomères, centromères, ARNr et une partie des ADNs satellites), elle ne dépasse pas les 24 %. Ces paradoxes nous ont conduits, entre autres, à nous questionner sur la fiabilité des méthodes d'annotations utilisées pour annoter les séquences répétées dans le génome du poulet.

La procédure d'annotation standard avec RM utilise une banque d'ETs qui est généralement Repbase. Cette banque de consensus est gratuite pour les académiques et payante pour les entreprises privées. Les chercheurs académiques découvrant de nouveaux ETs sont invités à les soumettre à Repbase. Cependant, comme le Giri qui gère cette base est une entreprise à but lucratif, peu de chercheurs académiques souhaitent leur confier leurs séquences. Bien que le personnel du Giri recherche et ajoute lui-même de nouveaux consensus, il est certain que cette base n'est pas exhaustive.

Le programme RM utilise et se limite donc aux connaissances acquises sur les ETs et accumulées dans Repbase pour réaliser l'annotation. Il ne sera donc pas capable de détecter de nouveaux éléments. Appliqué à chaque nouvel assemblage, il a annoté 7,99 % de séquences répétées ETs (Repbase 8.4) dans le modèle Galgal 2 dont la taille est de 1,13 Gpb. L'annotation des versions Galgal 3 et Galgal 4 a été réalisée avec les versions de Repbase 11.02 et 16.08. Elles ont mis en évidence des couvertures en séquences répétées de 8,00 % et 9,74 % pour des tailles de modèle de 1,10 Gpb et 1,046 Gpb. La proportion en répétitions dispersées est restée quasiment identique entre les versions Galgal 2 et Galgal 3 et augmente de 1,74 % dans la version Galgal 4. Cette variation est directement liée au contenu de la banque Repbase. Alors qu'elle contenait 4 576 et 6 906 séquences dans les versions 8.4 et 11.02, en 2011 elle comprenait 26 703 séquences (Repbase 16.08). Le génome du poulet

montre que la qualité de l'annotation RM est liée à la qualité et l'exhaustivité de la base employée. Pour obtenir l'annotation optimale, il faudrait connaître la totalité des éléments répétés du génome ainsi que toutes les versions délétées. L'utilisation de RM n'est donc pas la meilleure solution. Afin de réaliser une annotation de haute qualité et de haute résolution, il est nécessaire d'utiliser des programmes faisant une recherche *de novo*.

3. Comment annoter les séquences répétées ?

Les récentes découvertes en termes d'importance et d'implication des ETs dans les génomes ont rendu indispensable le développement d'outils bio-informatiques afin de mieux les caractériser, identifier et localiser le long des chromosomes. Leur diversité de structures et de mécanismes de mobilité ainsi que la taille des populations de chaque espèce d'ETs varient d'une espèce hôte à l'autre et en conséquence, font de leur détection un vrai défi. De nombreuses approches employant des stratégies différentes visant à détecter et annoter les différents types de répétitions ont été mises en place au cours des années [Lerat 2010]. Tout d'abord, je présenterai les différentes méthodes de détection et d'annotation des séquences répétées, puis les programmes utilisés pour annoter les différents types de séquences répétées du poulet, ainsi que RepeatExplorer qui a contribué à définir une nouvelle méthode de description des ETs.

3.1. Les méthodes de détection et d'annotation

3.1.1. Méthode structurale

Une première méthode d'annotation des ETs consiste à rechercher des structures caractéristiques d'un type d'élément. Le type d'ETs se prêtant le mieux à ce genre d'analyse est celui des rétrotransposons à LTR, car ils sont bornés par deux séquences répétées de taille significative (300 à 1500 pb). C'est pourquoi de nombreux programmes ont été développés. Les signatures recherchées sont :

- L'existence de séquences répétées (les LTRs putatifs) ;
- La taille des LTRs ;
- La distance entre les LTRs ;
- La présence d'un Target Site Duplication (TSD) ;
- La présence de site de réplication (PBS, polypurine tract) ;

- Le pourcentage d'identité entre les LTR ;

- La présence de certains motifs conservés correspondant aux gènes qui codent GAG, POL, et ENV.

Les programmes LTR_STRUC [McCarthy et al 2003], LTR_PAR [Kalyanaraman et al 2006], FIND_LTR [**Rho et al 2007**], LTR_HARVERST [Ellinghaus et al 2008], LTR_FINDER [Xu et al 2007] sont tous dédiés à la recherche de rétrotransposons à LTR, mais ont comme défaut commun d'avoir un taux de faux positifs élevé [Lerat 2010] car la qualité de détection demande une optimisation des paramètres pour chaque analyse de génome. Il existe aussi quelques programmes ayant pour objectif de détecter et d'annoter des rétrotransposons non-LTR tel que TSDFINDER [Szak et al 2002], RTANALYZER [Lucier et al 2007] mais ces derniers sont spécialisés sur les LINES L1. À noter, SINEDR [Tu et al 2001], et taillés pour la détection des SINEs du génome du moustique. On peut noter aussi l'existence de programmes visant à détecter les transposons comme les MITEs ; TRANSPO [Santiago et al 2002] et MUST [Chen et al 2009] ; et les hélitrons, HelitronFinder [Du et al 2008]. La mise en œuvre de ces programmes demande de grandes connaissances sur la structure et l'organisation des éléments recherchés et limite donc la détection de nouveaux éléments. Un inconvénient majeur de ces programmes est qu'ils utilisent des méthodes spécialisées dans la détection d'un seul type d'ET et applicable parfois même à une seule espèce. Il faudrait ainsi les utiliser conjointement pour réaliser une annotation qui, à terme, serait encore et toujours extrêmement partielle.

3.1.2. Méthode de librairie

La méthode couramment utilisée est celle qui consiste à annoter les séquences répétées d'un génome en recherchant les homologies avec les séquences présentes dans une base de données de séquences d'éléments répétés connus. Pour ce faire, le programme utilise un moteur d'alignement de séquence comme BLAST [Altschul et al 1990], WU-BLAST [http://blast.wustl.edu] ou CrossMatch [http://www.phrap.org] et un post-traitement des alignements pour produire une annotation. La méthode de référence est RM [RepeatMasker Open-4.0, Tempel 2012]. Il est utilisé en routine, notamment par le NCBI, pour réaliser l'annotation des génomes nouvellement publiés et par l'ISB (d'où a été créé et entretenu RM) pour annoter les génomes des animaux. La librairie de référence est Repbase [Bao et al 2015].

C’est une librairie de séquences consensus. Cependant, la méthode utilisée pour leur création est protégée. Il est donc difficile de juger sa qualité sans refaire une annotation avec une autre méthode. Il va de soi que cette méthode est incapable de détecter de nouveaux ETs, car elle est limitée aux éléments déjà répertoriés dans la librairie utilisée.

3.1.3. Méthode *de novo*

Les méthodes précédemment présentées ne sont pas capables de détecter de nouvelles familles de séquences répétées. Les méthodes dites *de novo* ont la capacité de rechercher dans un génome assemblé ou non de nouvelles espèces d’ETs, de répétitions en tandem en tirant parti de leur présence répétée dans le génome. Pour ce faire, elles utilisent uniquement la séquence génomique et emploient différentes approches.

La première consiste à comparer le génome à lui-même à l’aide d’un programme d’alignement. Généralement, cette étape est réalisée via le programme BLAST ou WU-BLAST, comme dans les programmes Repeat Pattern Toolkit [Agarwal et al 1994], BLASTER [Quesneville et al 2005] et RECON [Bao et al 2002]. PILER [Edgar et al 2005] utilise, lui, PALS [http://www.drive5.com/pals] pour réaliser l’alignement. Les alignements permettent de localiser des segments de séquences partageant un haut pourcentage d’identité sur la majeure partie de leur longueur. Ensuite, un traitement permet de regrouper toutes les paires d’alignements homologues mais plus distants les uns des autres pour former ainsi des groupes de séquences, dits clusters, capables de s'inter-aligner les uns avec les autres. Il existe plusieurs méthodes de clustering. Le single linkage, implémenté par SILIX [Miele et al 2000], consiste à regrouper dans un même cluster des séquences similaires deux à deux. Cependant, une partie des séquences d’un cluster peuvent présenter deux à deux des taux de divergence supérieurs à celui utilisé lors de la première étape comme seuil de détection. Le « multi-linkage », implémenté par MCL [Stijn van Dongen 2000] va former des clusters en tenant compte de l’homologie entre toutes les séquences du cluster. Des séquences consensus peuvent être ensuite générées à partir de l’alignement global des clusters. Les méthodes présentées sont applicables à des génomes assemblés, mais il existe des solutions offrant la possibilité de rechercher les éléments répétés à partir des données brutes de séquençage, directement et sans assemblage préalable.

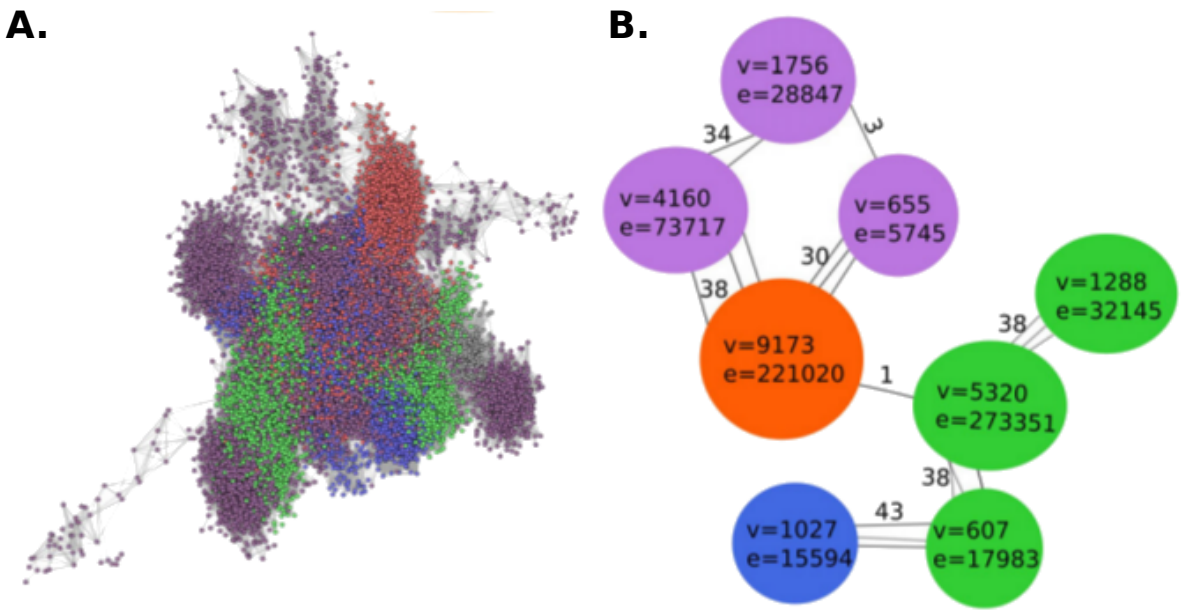


Figure 16 : Graphe créé au cours de l'analyse de RepeatExplorer
A. Représentation du graphe créé par RepeatExplorer. Les points représentent les lectures, les arrêtes représentent une homologie entre deux lectures. Les sommets de couleur représentent les différents sous-graphes correspondant aux différentes répétitions détectées.
B. Représentation schématique des sous-graphes détectés

Source : Novák P, Neumann P, Macas J: Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. BMC Bioinformatics 2010, 11:378.

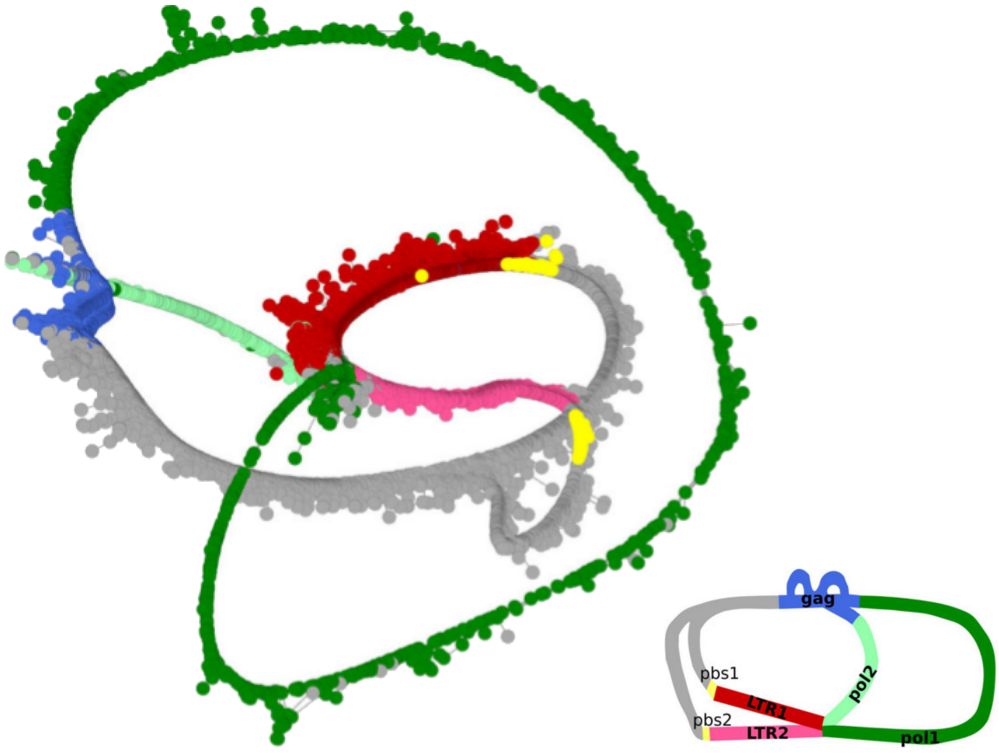


Figure 17 : Graphe du rétrotransposon à LTR gmGYPSY10 (*Glycine Max*) généré par RepeatExplorer
La comparaison des lectures contre une base protéique permet de localiser les différents domaines protéiques codés par la séquence nucléique de l'élément transposable.
les sommets bleus correspondent au domaine GAG, les sommets verts, au domaine POL, les sommets rouges aux LTR , et les sommets jaunes au Primer Binding Site.
Les différentes teintes de couleurs correspondent au différentes versions de séquence.

Source : Novák P, Neumann P, Macas J: Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. BMC Bioinformatics 2010, 11:378.

Ainsi, le pipeline RepeatExplorer [Novák et al 2010, Novák et al 2013] utilise les lectures issues d'un séquenceur 454 ou illumina et va comparer les lectures les unes aux autres. Un graphe, au sens mathématique, représentant les relations entre les lectures, est ensuite créé. Ce graphe consiste en une série de points ou sommets reliés par des arêtes, pouvant être représentés dans un espace à 2 ou 3 dimensions. Dans ce graphe, les sommets représentent les lectures. Une arête est créée entre deux lectures ayant au moins 90 % d'identité entre elles sur au moins 55 % de leurs longueurs. Une fois le graphe construit, RepeatExplorer va chercher à le subdiviser en sous-graphes correspondant aux différentes répétitions (Figure 16). Pour ce faire, il utilise l'algorithme de Louvain permettant de détecter les sommets fortement connectés.

Le type d'ET représenté par le sous-graphe peut être déduit par sa forme, mais aussi grâce à l'alignement des lectures qui le composent contre une banque de structures caractéristiques des ETs (Figure 17). Ces méthodes sont très efficaces pour la découverte de nouveaux éléments mais leur mise en œuvre est complexe. En effet, la comparaison du génome avec lui-même est une opération demandant énormément de calculs dont la quantité est proportionnelle au taux de répétition du génome étudié. C'est pour cela que RepeatExplorer ne peut traiter qu'un échantillon des séquences, 1 million de lectures, et se limite à la détection des éléments les plus répétés dans le génome.

Le temps et les capacités de calcul étant très souvent des facteurs limitants dans l'étude des séquences répétées, d'autres méthodes *de novo* demandant peu de temps de calcul ont été développées. Il s'agit des méthodes de k-mer qui consistent à considérer les séquences répétées comme des mots de taille donnée (k-mer) se répétant dans une séquence d'intérêt. L'avantage de la méthode des k-mer est qu'elle est très rapide et capable d'analyser un génome de très grande taille en peu de temps.

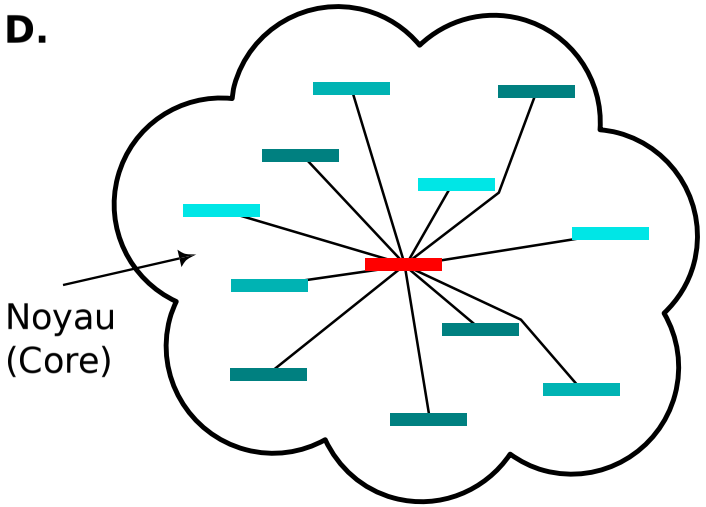
C.

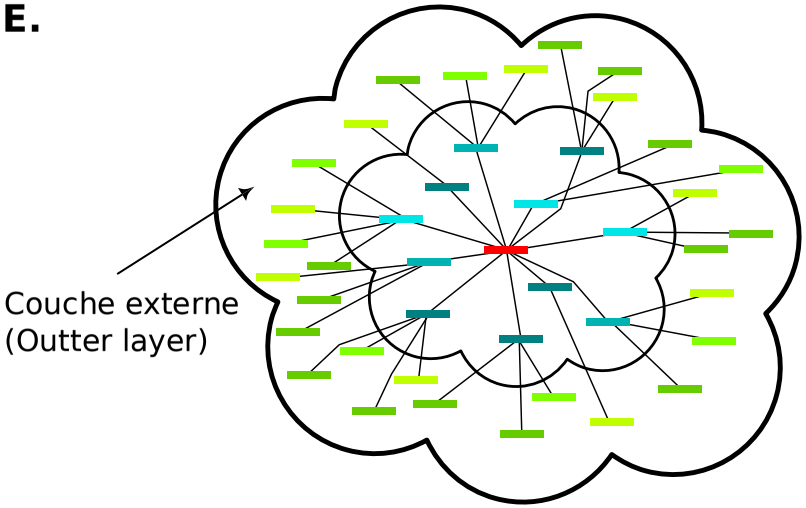
B.

Liste k-mer	Nombre dans le génome
...	...
ACGTTCATCTTCATCT	16463
ACGTTCATCTTCATCA	4
ACGTTCATCTTCATCG	57613
ACGTTCATCTTCATCC	234
ACGTTCATCTTCATCT	64643
ACGTTCATCTTCATTA	68131
ACGTTCATCTTCATTG	54318
ACGTTCATCTTCATTC	3220
ACGTTCATCTTCATTT	13
ACGTTCATCTTCATAA	3
...	...

A.

Lower cutoff :	5
Core cutoff :	15
Primary cutoff :	25
Secondary cutoff :	250
Tertiary cutoff :	2500





Seed k-mer
- Graine - k-mer le plus répété de la liste

Core k-mer
- k-mer avec 1 base de différence avec la graine
- k-mer avec 2 bases de différence avec la graine
- k-mer avec 3 bases de différence avec la graine

Outter layer k-mer
- k-mer with 1 base de difference avec son k-mer du noyau associé, le k-mer du noyau à un nombre d'apparation dans le génome supérieur au primary cuttoff
- k-mer avec 2 bases de difference avec son k-mer du noyau associé, le k-mer du noyau à un nombre d'apparation dans le génome supérieur au secondary cuttoff
- k-mer avec 3 bases de difference avec son k-mer du noyau associé, le k-mer du noyau à un nombre d'apparation dans le génome supérieur au tertiary cuttoff

Figure 18 : Fonctionnement de P-cloud
A. Paramètres de P-cloud.
Lower cutoff : nombre minimal de répétitions d'un k-mer pour qu'il soit considéré dans l'analyse.
Core cutoff : nombre minimal de répétitions d'un k-mer pour qu'il soit choisi comme graine
Primary cutoff: nombre minimal de répétitions d'un core k-mer pour qu'un k-mer ayant une base de différence soit intégré à la couche externe.
Secondary cutoff : nombre minimal de répétitions d'un core k-mer pour qu'un k-mer ayant une base de différence soit intégré à la couche externe.
Tertiary cutoff : nombre minimal de répétitions d'un core k-mer pour qu'un k-mer ayant une base de différence soit intégré à la couche externe.
B. P-clouds dresse la liste de tous les k-mers présents dans le génome et compte leur nombre de répétitions.
Les k-mers gris ne seront pas pris en compte dans l'analyse, car leur nombre de répétitions est inférieur au lower cutoff.
Le k-mer violet ne pourra pas être sélectionné comme graine, car son nombre de répétitions ne dépasse pas le core cutoff.
Le k-mer rouge est celui avec le plus de répétition dans toute la liste.
C. La construction du P-cloud débute par la sélection du k-mer le plus fréquent.
D. Construction du noyau : Tous les k-mers ayant 1, 2 ou 3 bases de différence avec la graine et ayant un nombre de répétitions supérieur au lower cutoff sont associés à la graine.
E. Construction de la couche exeterne : Pour chaque Core k-mer, les k-mers ayant une à trois bases de différence sont ajoutés à la couche externe à condition qu'il respectent les Primary,Secondary et Tertiary cutoff.

3.2. Annotation des différents types de séquences répétées

3.2.1. Sonder la proportion de répétitions d'un génome

Avant de commencer à faire une annotation détaillée des répétitions dans un génome, il est utile d'estimer leur proportion. Les programmes permettant de telles analyses se basent sur les méthodes de k-mer. Cependant, les ETs sont connus pour accumuler indépendamment des événements de mutations ponctuelles et de recombinaison qui font diverger les différentes copies d'une espèce au cours du temps. Ce phénomène de divergence est une limite à leur détection par les méthodes de k-mers, car elles se basent sur la recherche de mots exacts. Contrairement aux programmes comme Tallymer [Kurtz et al 2008] ou Jellyfish [Marçais et 2011], P-clouds [Gu et al 2008] et Red [Girgis 2015] essayent de pallier ce défaut avec deux approches différentes.

3.2.1.1. P-clouds

P-clouds tente de s'affranchir des défauts des méthodes classiques de k-mer en introduisant la possibilité de rechercher les k-mers ayant des divergences de séquences et en les regroupant dans des nuages de probabilité (Probability cloud, P-cloud) dont la construction se fait en plusieurs étapes.

La première étape est un comptage de la totalité des mots de taille k présents au sein de la séquence étudiée. La taille de ce mot est choisie pour qu'il y ait moins d'une chance de trouver par hasard un mot de cette longueur plus d'une fois. En effet, si la taille de mot est trop petite, il est possible qu'une répétition soit détectée bien qu'elle ne soit pas issue d'un mécanisme d'amplification. Cette taille de mot dépend de la taille du génome et peut être obtenue à l'aide de la formule mathématique $L = \log4(n) + 1$ où L est la taille du k-mer recherché et n la taille du génome étudié.

Pour construire le premier P-cloud, le programme va utiliser le k-mer le plus répété, comme celui d'une graine (Figure 18). Il va ensuite créer le noyau du P-cloud en associant à la graine tous les k-mers divergeant de une à trois bases. Ensuite, il va construire la couche externe du P-cloud en recherchant les k-mers ayant au maximum trois bases de différence avec au moins un des k-mers du noyau. L'éligibilité d'un k-mer pour être intégré dans l'un des composants du P-cloud dépend de 5 paramètres :

- Lower cutoff : Définit le nombre minimum de répétitions d'un k-mer pour qu'il puisse être intégré dans un P-cloud ;

- Core cutoff : Définit le nombre minimum de répétitions d'un k-mer pour qu'il puisse être utilisé comme graine de P-cloud ;

- Primary cutoff : Définit le nombre minimum de répétitions d'un k-mer du noyau pour qu'un k-mer ayant une base de différence soit intégré à la couche externe ;

- Secondary cutoff : Définit le nombre minimum de répétitions d'un k-mer du noyau pour qu'un k-mer ayant deux bases de différence soit intégré à la couche externe ;

- Tertiary cutoff : Définit le nombre minimum de répétitions d'un k-mer du noyau pour qu'un k-mer ayant trois bases de différence soit intégré à la couche externe.

Ce procédé apporte de la flexibilité à la méthode de k-mer, car il autorise de la divergence entre les mots. De plus, par son système de seuil, il prend en compte la variabilité des séquences répétées. En effet, lorsqu'une séquence est très fortement répétée dans un génome, il y a de plus fortes chances que cette séquence dérive et conduise à l'apparition de nombreuses séquences divergentes par rapport à la séquence originale. Dans les P-clouds, la graine représente la séquence répétée originale, les différentes couches correspondent aux versions ayant divergé à partir de la graine.

Quand plus aucun k-mer ne respecte les conditions d'intégration à la couche externe, les k-mers impliqués sont supprimés du comptage et un nouveau P-cloud peut être construit. Le procédé est ainsi répété jusqu'à ce qu'il n'y ait plus de k-mer ou que leur nombre de répétitions ne dépasse pas le core cutoff.

Quand tous les P-clouds ont été construits, ils sont cartographiés sur le génome analysé, ce qui permet d'avoir une estimation de la couverture totale en répétition.

Ce programme a déjà été expérimenté sur des modèles génomiques et présente des résultats prometteurs. Sur le génome humain, dont la proportion de répétition est de plus de 45 %, il a conduit à affirmer que près des deux tiers du génome sont répétés [de Koning et al 2011]. Les méthodes d'annotation par homologie comme RM détectent dans les génomes du

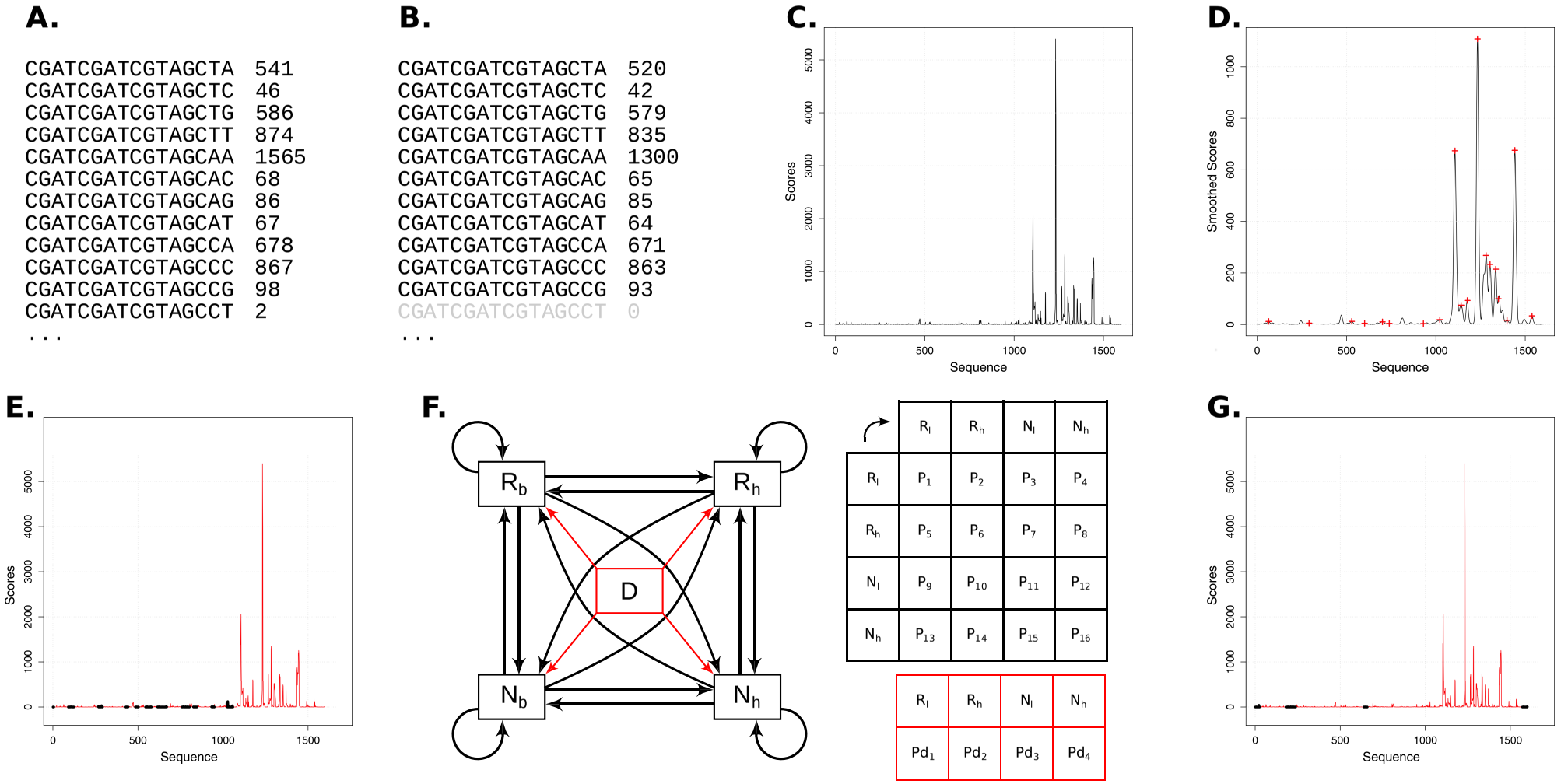


Figure 19 : Algorithme de Red

A. Comptage des k-mers présents dans le génome

B. Ajustement du comptage des k-mers. Les k-mers répétés moins de 3 fois sont considérés comme non répétés (k-mer gris). Pour les autres k-mer, le comptage ajusté est obtenu en soustrayant au nombre de k-mers comptés le nombre de k-mers atttendus estimé à l'aide d'une chaîne de Markov entrainée sur le génome.

C. Pour chaque base du génome le programme attribut un score correspondant au comptage ajusté du k-mer débutant à chacune de ces postions. Le graphique représente en abscisse la postion de chaque base dans le génome et les ordonnés les scores associés à chacune de ces bases.

D. Les scores sont lissés à l'aide d'un masque de Gauss et les maximums locaux (croix rouge) sont déterminés en calculant les dérivées primaire et secondaires des scores sur une fenêtre de 10 pb.

E. En se servant des maximums locaux, il marque les régions comme répétées (rouges) ou non répétées (noires)

F. Entrainement du modèle de Makov caché. Le jeu d'apprentissage est utilisé pour déterminer les paramètres du Modèle de Markov caché (MMC). L'exemple MMC présenté comprend 4 états: Répété avec un score haut (R_h), Répété avec un score bas (R_b), Non répété avec un score haut (N_h), Non répété avec un score bas (N_b). La tableau rouge représente les probabilités de départ de chaque état (flèches rouges). Le tableau noir représente la matrice de transition d'un état à un autre (flèches noires sur le MMC).

G. L'algorithme de Viterbi appliqué au MMC entrainé et les scores permettent déduire les régions répétées (rouges) ou non (noires).

Source (graphiques) : Girgis : Red: an intelligent, rapid, accurate tool for detecting repeats de-novo on the genomic scale. BMC Bioinformatics 2015, 16:227.

python et du mocassin à tête cuivrée (*Agkistrodon contortrix*) des taux de répétition de 21 et 45 %, alors que P-clouds détecte 55 et 40 % de séquences répétées [Castoe et al 2011]. Il a également permis, lors d'une ré-annotation du génome de *Arabidopsis Thaliana*, de révéler la présence d'un ancien centromère issu de la fusion de deux chromosomes en mettant en évidence d'anciennes répétitions en tandem [Maumus et al 2014]. P-clouds présente l'avantage d'être peu consommateur de mémoire, ce qui rend possible son utilisation sur une machine de bureau « classique ».

3.2.1.2. Red

Red est un programme récemment publié qui innove en mettant en place une méthode de détection des séquences répétées en utilisant une technique d'apprentissage automatique se basant sur un décompte de k-mers. Il s'agit d'un programme qui, à partir d'un jeu de données d'apprentissage, est capable d'intégrer des règles d'analyse pour ensuite les appliquer à un jeu de données expérimentales. L'avantage de Red est qu'il est capable d'apprendre à distinguer les séquences répétées de celles qui ne le sont pas directement à partir du génome sans avoir besoin d'une annotation de référence sur laquelle apprendre. Cependant, Red demande une grande quantité de mémoire vive pour analyser de gros génomes, contrairement à P-clouds, qui a pour objectif d'offrir la possibilité de réaliser des annotations rapides et peu consommatrices de mémoire vive.

Ce programme est constitué de 4 modules s'exécutant séquentiellement dans l'ordre suivant (Figure 19) :

- Scoring module ;
- Labeling module ;
- Training module ;
- Scanning module.

Le module de scoring a pour objectif de réaliser un comptage ajusté de tous les k-mers du génome analysé. Comme pour P-cloud, la taille du k-mer employé est calculée en fonction de la taille du génome. Le module va commencer par répertorier tous les k-mers et les stocker dans une structure de données qui s'appuie sur un algorithme d'indexation performant. Pour

limiter la détection de régions codantes et les régions dupliquées, Red commence par réaliser un ajustement du comptage des k-mers. Il va considérer les k-mers répétés seulement deux fois comme non répétés. Pour les k-mers ayant au moins 3 répétitions, leur comptage sera pondéré en soustrayant le nombre théorique de répétitions attendues au hasard dans le génome. Pour terminer, le module va associer pour chaque base du génome un score correspondant au comptage du k-mer commençant à cette base. Les scores peuvent donc être représentés sur un graphique dont l'axe des ordonnées correspond à l'emplacement des bases dans le génome et dont l'abscisse représente le score.

L'objectif du module de labeling (marquage) est de définir les régions répétées et celles qui ne le sont pas en utilisant les scores précédemment calculés. Il va tout d'abord appliquer un masque gaussien au jeu de scores afin de les lisser. Comme les régions répétées sont caractérisées par l'agrégation de hauts scores et les régions non répétées de scores faibles, le lissage limitera la détection de faux positifs. Le module va ensuite chercher les scores maximums locaux pour marquer les régions comme répétées. Il calcule la dérivée primaire et secondaire des scores sur une fenêtre de 10 pb tout au long du génome. Ce marquage des régions comme répétées ou non répétées forme le jeu d'apprentissage qui servira à calibrer le programme.

Red passe ensuite à la phase de l'apprentissage. Pour ce faire, il utilise un modèle de Markov caché, un outil statistique qui, à partir d'une suite d'états (répétés ou non), permet de déterminer une séquence d'observation (les scores). Pour fonctionner, il faut définir les 3 paramètres du modèle : les probabilités de départ, la matrice de probabilité de transition entre chaque état, et les probabilités d'émission. Il utilise le jeu d'apprentissage précédemment créé pour évaluer ces paramètres. Une fois l'apprentissage réalisé, il applique l'algorithme de Viterbi au modèle de Markov caché et les scores pour générer la série d'états inhérents à la liste de scores. La liste d'états générée permet de statuer sur le caractère répété ou non des régions de la séquence génomique.

Il n'existe pas encore d'étude ayant utilisé Red pour annoter un génome. La seule évaluation de ce programme réside dans l'étude réalisée par l'auteur, dans laquelle il compare son outil à RepeatScout [Saha et al 2008], ReCon [Bao et al 2002] et WindowMasker [Morgulis et 2006]. L'auteur estime que les résultats de son programme sont meilleurs et qu'ils ont permis de mettre en évidence de nouvelles répétitions.

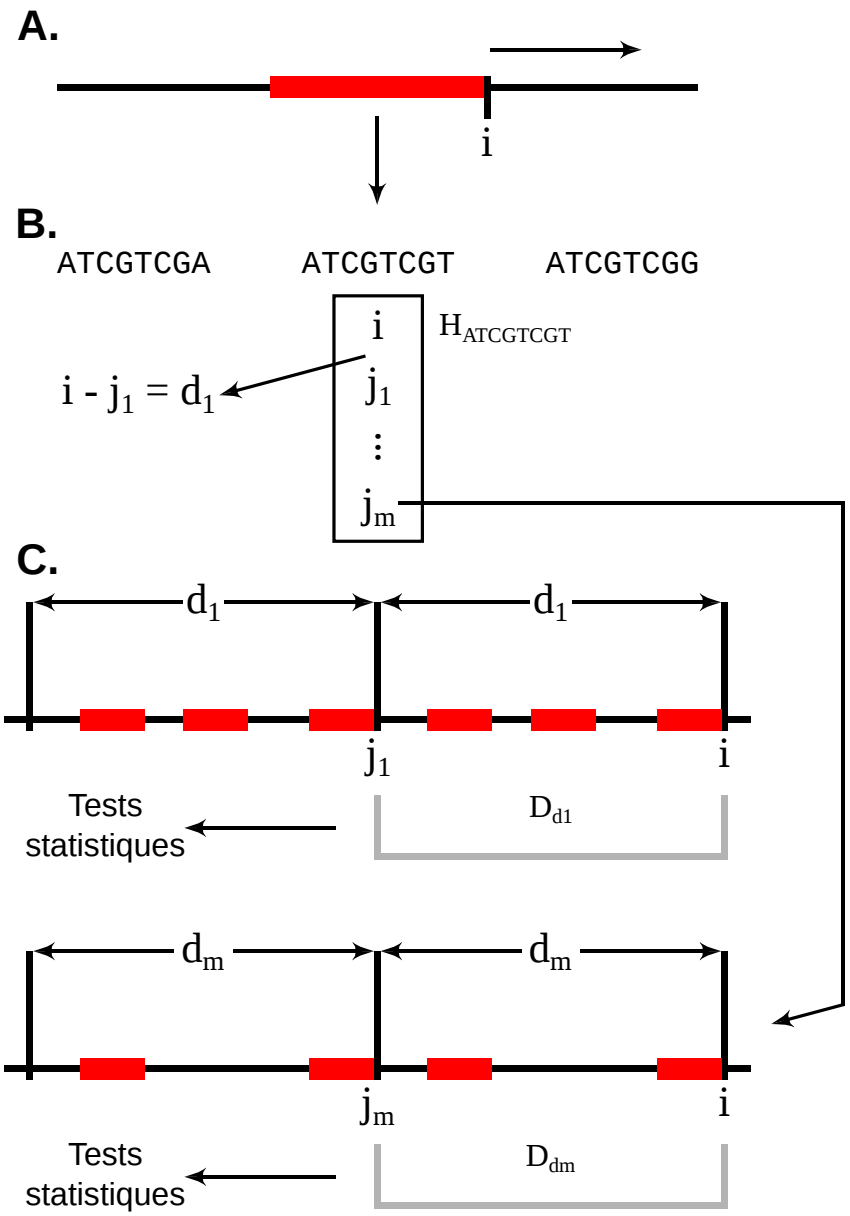


Figure 20 : Algorithme de Tandem Repeat Finder
A. Une fenêtre glissante détermine tous les k-mers pour une taille donnée
B. Pour chaque k-mer il stocke dans une liste ($H_{k\text{-}mer}$) les positions de ce k-mer dans le génome.
C. À chaque fois qu'une position est ajoutée à la liste $H_{k\text{-}mer}$, toutes les positions précédentes sont scannées. Si i est la dernière position ajoutée à la liste et j une position précédente, la distance $d=i-j$ est une unité de répétition potentielle. Afin de réaliser les tests statistiques, une liste D_d regroupe les positions des k-mers séparés d'une distance d compris entre les bornes i et j. La liste D_d est mise à jour à chaque fois qu'une nouvelle position est ajoutée. Les listes de distances proches sont également mises à jour. La variation de distance est calculée à l'aide de l'algorithme Random Walk Distribution. Les informations de la liste D_d sont ensuite validées avec deux tests statistiques. Le "sum-of-heads" qui permet de définir le nombre minimum d'identité entre les copies et le "apparent-size" qui permet de vérifier s'il s'agit d'une répétition en tandem. Si les tests sont validés, le module d'analyse va caractériser la répétition en tandem en alignant les copies de l'unité de répétition potentielle pour déterminer sa taille et créer un consensus.

3.2.2. Détecter et annoter les répétitions en tandem

Avec plus de 2 000 citations, Tandem Repeat Finder (TRF) [Benson et al 1999] est le programme de prédilection pour la recherche et l'annotation des séquences répétées en tandem. Son succès est lié à ses nombreuses qualités. Basé sur une méthode de k-mer, il est très rapide et capable d'analyser des génomes entiers en peu de temps. De plus, son algorithme adapte la taille du k-mer utilisé à la taille du génome pour limiter son temps de calcul. L'autre avantage est qu'il est capable de détecter des répétitions tandem en tenant compte des variations (insertion, délétion, mutation) possibles de l'unité de répétition.

Il fonctionne à l'aide de deux composants (Figure 20) : un module de détection et un module d'analyse. Le module de détection va analyser la séquence étudiée en utilisant une fenêtre glissante pour déterminer tous les k-mers du génome. La taille du k-mer est calculée en utilisant l'algorithme Waiting Time Distribution pour que le temps d'analyse soit raisonnable. Pour chaque k-mer, le programme tient une liste H_{kmer} répertoriant toutes ses positions dans le génome. À chaque fois qu'une position est ajoutée à la liste H_{kmer}, il scanne toutes les positions précédentes. Si on considère i comme la dernière position ajoutée à la liste et j, une position précédente, la distance d = i-j est une unité de répétition potentielle. Afin de réaliser les tests statistiques, une liste D_d regroupe les positions des k-mers séparés d'une distance d comprise entre les bornes i et j. La liste D_d est mise à jour à chaque fois qu'une nouvelle position est ajoutée. Les listes des distances proches sont également mises à jour. La variation de distance est calculée à l'aide de l'algorithme Random Walk Distribution. Les informations de la liste D_d sont ensuite validées avec deux tests statistiques. Le « sum-of-heads » permet de définir le nombre minimum d'identité entre les copies et le « apparent-size » permet de vérifier s'il s'agit d'une répétition en tandem. Si les tests sont validés, le module d'analyse va caractériser la répétition en tandem en alignant les copies de l'unité de répétition potentielle pour déterminer sa taille et créer un consensus.

Ce programme permet donc de réaliser une annotation *de novo* des répétitions en tandem et ainsi une analyse de leur quantité, leur distribution et leur diversité de séquence.

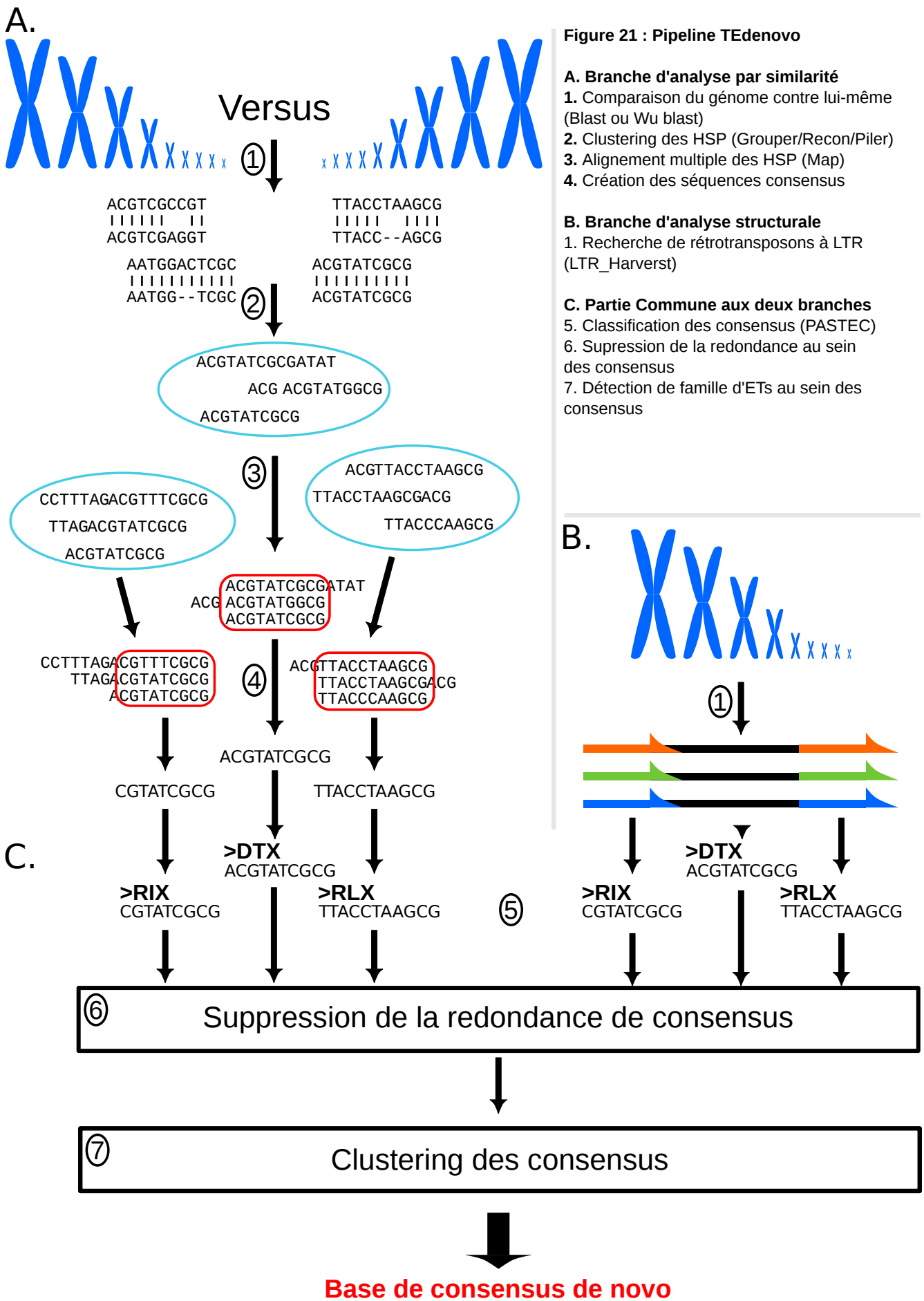


Figure 21 : Pipeline TEdenovo

A. Branche d'analyse par similarité
1. Comparaison du génome contre lui-même (Blast ou Wu blast)
2. Clustering des HSP (Grouper/Recon/Piler)
3. Alignement multiple des HSP (Map)
4. Création des séquences consensus

B. Branche d'analyse structurale
1. Recherche de rétrotransposons à LTR (LTR_Harverst)

C. Partie Commune aux deux branches
5. Classification des consensus (PASTEC)
6. Supression de la redondance au sein des consensus
7. Détection de famille d'ETs au sein des consensus

3.2.3. Détecter et annoter les éléments transposables

L'annotation des ETs est complexe. Leur grande diversité de mécanismes de transposition, de structures et la divergence rapide de leurs copies rendent leur détection difficile. Les nombreuses méthodes existantes ont toutes leurs avantages et leurs inconvénients, mais aucune ne se présente comme une solution complète de détection et d'annotation *de novo*. C'est pourquoi une méthode, dite méthode de consensus, a été mise en place à l'URGI (INRA de Versailles) par Hadi Quesneville et son équipe de recherche. Le but du pipeline REPET [Quesneville et al 2005, Flutre et al 2011] est de combiner les différentes méthodes existantes afin de tirer parti de leur diversité de résultats pour réaliser non seulement une détection *de novo* des éléments répétés, mais aussi l'annotation la plus complète possible. Il se décompose en deux grands pipelines. TEdenovo dont le but est la détection et la création de séquences consensus d'ETs et TEannot dont l'objectif est de produire une annotation fiable à partir des consensus produits par TEdenovo (méthode de librairie).

3.2.3.1. TEdenovo

Ce pipeline associe plusieurs méthodes de détection pour générer un inventaire de séquences répétées le plus complet possible. Pour ce faire, il va combiner des méthodes de détection par similarité (branche par similarité) et de détection par signature (branche structurale) (Figure 21).

La branche structurale du pipeline va utiliser le programme LTR_Harvest pour rechercher dans le génome les rétrotransposons à LTR en cherchant la présence de deux longues répétitions directes séparées d'une distance donnée sans tenir compte du contenu entre ces LTRs. Chaque région remplissant ces conditions sera considérée comme un rétrotransposon à LTR potentiel et créera une séquence consensus. La branche par recherche d'homologie consiste en une première étape de comparaison du génome à lui-même avec BLASTER. Tout d'abord, il utilise MATCHER qui, par le biais d'un algorithme de programmation dynamique, va reconnecter les copies proches. Les paires de séquences alignées obtenues sont ensuite analysées conjointement par trois programmes de clustering qui utilisent des stratégies différentes. Le premier, GROUPER, est issu de la suite BLASTER. Il commence par une étape de défragmentation des copies. Les éléments transposables

divergeant rapidement et les événements d'imbrication de séquences mobiles étant fréquents, il arrive que l'alignement d'une séquence consensus sur la séquence génomique soit interrompu localement et génère plusieurs copies. GROUPER cherche donc ces lots de copies qui correspondent à une seule et même requête en utilisant un algorithme de programmation dynamique. Une fois la défragmentation faite, il réalise un clustering de type single linkage en utilisant une contrainte de couverture très élevée (95 %) entre les séquences, ce qui permet de rechercher les variantes de l'ET. Le deuxième clustering est réalisé par RECON. Il commence par déduire une copie ancestrale d'une répétition appelée « élément ». Pour ce faire, il réalise un clustering de type single linkage à faible contrainte de couverture entre les séquences (50 %) sur les HSPs (« High-scoring Segment Pair ») provenant d'un même locus. Ensuite, il effectue un deuxième clustering à haute contrainte de couverture entre les séquences (90 %) sur les « éléments ». Le troisième programme est PILER. Il recherche des « piles » d'HSPs, des locus du génome sur lesquels plusieurs séquences requêtes ont été alignées. Il va ensuite réaliser un alignement multiple de ces piles et s'il arrive à aligner tous les HSPs sur 95 % de leur longueur, il les rassemble dans un cluster.

Les trois méthodes génèrent donc une liste de clusters qui vont être ensuite filtrés. Tout d'abord, tous les clusters comportant moins de trois séquences sont éliminés. Pour les clusters regroupant de nombreuses séquences, seules les 20 plus grandes sont conservées, car les séquences mal alignées engendrent une baisse de la qualité des consensus. Les clusters issus de GROUPER subissent une troisième sélection visant à éliminer ceux dont la taille cumulée des membres dépasse 20 kpb et qui couvre plus de 30 kpb du génome, afin de ne pas générer des consensus correspondant à des duplications segmentales.

L'étape suivante du programme est la création des séquences consensus à partir des clusters. Pour ce faire, TEdenovo va calculer l'alignement global de chaque cluster, puis en déduire une séquence consensus. Les ETs ne sont pas les séquences qui se prêtent le mieux à l'alignement multiple. En effet, les différentes copies peuvent avoir accumulé des quantités plus ou moins importantes de mutations, ce qui rend leur alignement multiple difficile avec les programmes classiques tels que Muscle [Edgar et al 2004], Clustal-W [Thompson et al 1994], MAFFT [Katoh et al 2002] ou PRANK [Loytynoja et al 2010]. C'est pour cela que le programme MAP a été intégré dans TEdenovo. MAP a été développé pour aligner des séquences comportant beaucoup de délétions, et contrairement aux autres programmes

d’alignement multiple, il autorise de grandes délétions au sein de l’alignement multiple. À partir de chaque alignement multiple, TEdenovo va générer une séquence consensus avec les règles suivantes : pour chaque position dans l’alignement, il va conserver la base majoritaire et va exclure les positions ne comportant qu’une seule base alignée.

La liste de consensus *de novo* de séquences répétées créée par l’analyse par similarité et structurale est encore de type inconnu. Une étape supplémentaire est ajoutée afin de les classer selon la classification de Wicker. Pour cela, il recherche la présence de :

- répétitions terminales ;
- répétitions en tandem ;
- cadres de lectures ;
- queues polyA.

Il peut également rechercher des informations sur les consensus en les comparant à des banques de :

- profil HMM nucléotidique et protéique d’éléments transposables ;
- ADNc de gènes de l’hôte ;
- ADN ribosomiques ;

Les informations collectées sont stockées dans une base de données qui servira à l’étape de classification.

Le programme PASTEC est en charge de la classification. Il est composé d’une série de programmes et d’agents de classification qui vont calculer chacun, pour chaque consensus, un score de correspondance avec le type d'ET qu’ils sont en charge d’évaluer. Ce score dépend des différentes informations de similarité et de structures détectées à l’étape précédente. Des agents de classifications sont également dédiés à la détection de gènes hôtes, de répétitions en tandem et d’ADNr. Un dernier programme, le « super agent », va examiner pour chaque consensus la série de scores obtenue et va attribuer le type d'ET sur la base du

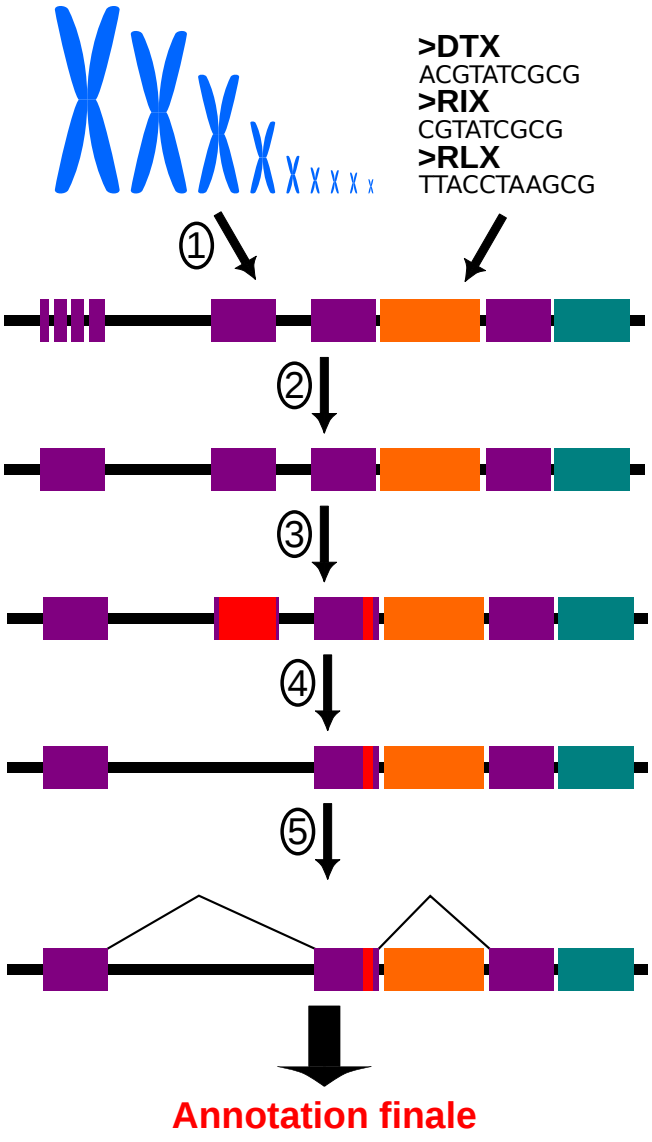


Figure 22 : Pipeline TEannot
1. Le génome est annoté à l'aide de la librairie de séquences consensus de novo par trois programmes différents : BLASTER, RepeatMasker, Censor
2. Les copies fragmentées et proches sont fusionnées à l'aide de Matcher
3. Les SSR sont annotés à l'aide de trois programmes différents : TRF, Mreps, RMSSR
4. Les copies correspondant à des SSR sont supprimées de l'annotation
5. Les copies éloignées sont reconnectées à l'aide de l'algorithme Long-join

score le plus élevé. Il va également statuer sur le caractère complet du consensus (par exemple si un consensus de rétrotransposon à LTR possède bien des répétitions terminales, une intégrase, une reverse transcriptase, une RNAase H, une GAG et un gène ENV) mais aussi sur le caractère chimérique (par exemple, un consensus portant des structures de LINE et de LTR). Les séquences consensus sont ensuite renommées en suivant une nomenclature respectant un code à trois lettres proposé dans la classification de Wicker [Wicker et al 2007] indiquant également les informations sur la complétion.

L'analyse de trois programmes de « clustering » faite sur un même jeu de données conduit à un certain niveau de redondance dans les consensus. Elle est éliminée si un consensus caractérisé comme incomplet est inclus (c.à.d. est une sous-séquence) dans un consensus considéré comme complet. Sont éliminés également les consensus qui ont été classifiés comme des répétitions en tandem, des gènes hôtes, ou comme ADNr.

Lors de la dernière étape, TEdenovo tente de créer des familles de consensus d'éléments transposables en réalisant un « clustering » avec BlastClust ou avec MCL. En sortie de TEdenovo est obtenue une série de consensus auxquels une classification a été attribuée (et quand cela est possible). Les consensus issus de la branche structurale possèdent au minimum une seule copie et ceux issus de la branche par recherche d'homologie ont au moins 3 copies dans le génome analysé.

3.2.3.2. TEannot

Le deuxième pipeline de REPET est destiné à produire une annotation des ETs du génome en combinant les résultats de plusieurs annotateurs puis en les analysant. Il nécessite en entrée le génome d'intérêt ainsi qu'une base de consensus, soit celle générée par TEdenovo, soit une base de séquences créée par l'utilisateur (Figure 22).

TEannot utilise BLASTER, RM et CENSOR [Kohany et al 2006] pour cartographier les séquences de référence sur le génome. Pour filtrer les faux positifs, il génère une version aléatoire du génome d'entrée avec le programme Shuffle de la suite HMMER et la ré-annote avec les trois programmes. Cette annotation permet de déduire pour chaque méthode un score en dessous duquel une copie annotée peut être obtenue par le hasard. Les annotations ne remplissant pas cette condition sont éliminées.

TEannot va ensuite défragmenter les copies en cherchant dans un premier temps à reconnecter les copies proches, puis dans un deuxième temps les copies les plus éloignées. TEannot procède alors à une recherche des microsatellites à l'aide de trois programmes, TRF, mreps [Kolpakov 2003] et RM (SSR). Cela permettra d'éliminer toutes les copies de l'annotation correspondant à des microsatellites. Il applique ensuite la procédure du « long join » qui sert à reconnecter les copies séparées par une longue distance. Cette étape permet de repérer les ETs dont la séquence de la copie a pu être coupée en deux par l'insertion d'un autre ET. Pour ce faire, il va évaluer l'âge (c.à.d. le niveau de dégénérescence de la séquence) de deux segments distants d'un même élément et si ceux-ci correspondent, il les rassemble dans une seule et même annotation. Pour terminer, il crée des fichiers d'annotation de type GFF3 décrivant toutes les copies d'ETs annotées.

Objectifs de la thèse

Aujourd'hui, la plupart des efforts en termes d'annotation sont concentrés sur la détection des gènes. Cependant, la partie codante d'un génome ne représentant que 2 à 4 % des génomes. Dans ce contexte, les séquences répétées ont longtemps été ignorées alors qu'elles représentent chez l'homme au minimum 45 % du génome nucléaire, voire les deux tiers pour l'estimation la plus optimiste. Pourtant, ces répétitions jouent un rôle majeur dans les fonctions d'un génome. Elles facilitent les recombinaisons ectopiques, peuvent modifier l'expression des gènes, provoquer des ré-arrangements du génome en étant capables de déplacer de larges portions d'ADN comme dans le cas des éléments ogres dont la taille peut atteindre 25 kpb [Macas et al 2007, Steinbauerová et al 2008] et de modifier l'organisation et la fonction d'un gène. Lors de la publication d'un nouvel assemblage, les séquences répétées sont généralement annotées à partir d'une simple analyse par le programme RM avec comme référence la base Repbase. Cette annotation sera donc dépendante des connaissances actuelles des ETs du génome présents dans la librairie Repbase. L'étude *de novo* des ETs étant exceptionnelle, les annotations sont généralement partielles et l'identification d'éléments nouveaux est impossible avec cette méthode. Aujourd'hui, il est donc nécessaire de revoir la stratégie d'annotation de façon à être le plus complet possible pour détecter *de novo* toute la population d'ETs du génome analysé et produire une annotation de très haute résolution et retrouvant les copies les plus anciennes.

1. Ré-annotation des séquences répétées du génome de la poule rouge de jungle

Chaque initiative d'annotation *de novo* des nouveaux génomes applique sa propre méthode en utilisant des programmes d'annotation et de détection différents. Cette absence de protocole standard a été soulignée récemment par la publication d'un article en 2015, « A call for benchmarking transposable element annotation methods » [Hoen et al 2015]. Les auteurs mettent en évidence la diversité de programmes d'analyse des ETs, mais ils déplorent l'absence de méthodes standards valisées pour guider les chercheurs non aguerris dans le domaine de l'annotation des ETs. L'annotation du génome du poulet est donc partielle. En effet, la seule analyse *de novo* des ETs faite sur ce génome a été réalisée avec CENSOR en 2004 lors de sa première publication. L'étude de Wicker en 2005 a fait usage d'une méthode expérimentale pour caractériser les répétitions. Cette méthode ne pouvant repérer que les éléments abondamment répétés, elle n'a pas permis une avancée significative des connaissances sur l'organisation du génome du poulet. Les annotations des nouvelles versions ont quant à elles été faites en utilisant RM avec Repbase comme référence. L'usage de cette approche limitée par son principe explique en partie le faible taux d'ETs identifiés à ce jour dans ce génome. Repbase est une banque visant à rassembler les consensus de séquences répétées provenant de tous les organismes, mais elle n'est pas exhaustive. De plus, l'annotation actuelle ne permet pas une étude fiable de l'organisation du génome, de la dynamique des séquences répétées et de leur histoire évolutive.

À partir de ce constat, l'objectif de cette thèse est de créer une nouvelle annotation du génome du poulet. Pour cela, notre premier objectif a été de proposer une méthode (c.à.d. une stratégie d'analyse) permettant une détection de la population la plus complète possible et une annotation du contenu en répétitions de haute résolution. La première partie de la thèse a consisté à mettre au point cette méthode. Il a d'abord fallu procéder à la sélection des programmes les plus pertinents pour l'analyse du génome de la poule rouge de jungle. En effet, il existe peu de programmes capables de prendre en charge un génome entier d'amniotes (à l'exception des génomes de faible taille tels que celui de *tétraodon*, dont la taille est de 340

Mpb). L'assemblage en chromosomes et sa petite taille relative (vis-à-vis des mammifères) font du génome du poulet un candidat de choix pour tester une méthode d'annotation approfondie des séquences répétées des vertébrés.

La méthode de choix mise en place se décompose en 5 parties ordonnées. Elle commence par l'évaluation de la quantité globale de répétitions. Elle permet de définir une valeur maximale de répétitions pouvant théoriquement être annotées. Nous avons testé deux programmes, Red et P-clouds qui, de par leurs méthodes de k-mer, offrent la possibilité de faire une analyse sur le génome complet et produisent des résultats rapidement.

La deuxième partie est dédiée à la recherche et l'annotation des répétitions hautement répétées. À l'heure actuelle, l'annotation de ces séquences est centrée sur les microsatellites. Celle-ci est généralement faite avec RM qui intègre l'algorithme de TRF et se limite à des motifs de répétitions d'une taille maximum de 10 paires de bases (5 dans le cadre du dernier assemblage galgal5). Du fait de ce choix, les mini-satellites et les satellites ne sont pas recherchés alors qu'ils peuvent représenter une grande partie du génome. Nous avons donc fait le choix d'utiliser TRF, car il permet d'analyser le génome en entier en peu de temps mais aussi de rechercher des séquences hautement répétées dont le motif de répétition peut avoir une taille allant qu'à 2 kpb, couvrant ainsi les microsatellites, les mini-satellites et les ADN satellites.

La troisième partie est dédiée à l'annotation des ETs. Il est évident que l'amélioration de l'annotation actuelle ne peut se faire que par l'utilisation d'une méthode *de novo*. Le package REPET semble être la solution la plus appropriée, car il combine une série de programmes de détection et d'annotation, intègre un post-traitement des résultats, et est livré avec une batterie d'outils permettant une analyse supplémentaire en exploitant les données collectées au cours de son processus.

La quatrième partie est consacrée à l'annotation de la matière noire. Nous avons tenté d'affiner la recherche des copies d'ETs pour retrouver les fragments les plus anciens non détectables avec les méthodes classiques. Normalement, ces études se font via des méthodes de k-mer comme P-clouds, mais celles-ci ne permettent pas de définir l'origine de cette matière noire. Nous avons donc expérimenté une méthode qui consiste en l'utilisation directe des copies d'ETs annotées par REPET. Le fondement de cette approche est d'utiliser les

copies annotées plutôt que les consensus pour annoter avec TEannot. Son avantage est qu'elle permet d'aller chercher des copies et des segments d'ETs dont l'identité est plus éloignée. Cette méthode a été élaborée et testée sur un génome de petite taille, celui d'*Arabidopsis thaliana* [Maumus et al 2014]. L'appliquer au génome de la poule est un challenge, car le nombre de copies à prendre en considération est beaucoup plus important.

La dernière partie de notre méthode concerne les CNVs qui désignent des régions du génome dupliquées. Ces CNVs engendrent selon l'espèce analysée une variation du nombre de copies d'un même gène. Nous avons intégré à nos résultats les données issues d'une précédente annotation des CNVs du génome de la poule réalisée en 2014 [Yi et al 2014].

Les outils développés au cours de la thèse auront vocation à être utilisés pour annoter les génomes nouvellement séquencés. Certains outils et les données produites sont à disposition en ligne sur le site chicken-repeats.inra.fr.

2. Redécouvrir le génome du poulet

Grâce aux outils disponibles et à ceux que nous avons développés, nous avons établi une nouvelle annotation des ETs du génome de la poule rouge de jungle. À partir de cette annotation, nous avons exploité ces nouvelles connaissances pour étudier leur distribution entre les chromosomes. Nous avons démontré que certains chromosomes étaient plus ou moins riches en ETs et que la distribution de la plupart des 34 espèces d'ETs n’était pas le fait du hasard. Enfin, nous avons étudié la distribution des éléments transposables dans les gènes en évaluant la part d’ETs insérée dans les exons, les introns et dans les régions inter-géniques.

Travaux

1. Le bon outil de calcul

L'avènement des technologies NGS a fait rentrer la biologie dans l'ère de la bio-informatique. La production et l'analyse de données à l'échelle du génome demandent de grandes puissances de calcul. L'analyse *de novo* des éléments transposables fait partie de ces études gourmandes en calcul. Les méthodes dites de « all-by-all », notamment utilisées par REPET, requièrent de comparer le génome à lui-même. Si l'on découpe un génome de 1 Gpb en fragments de 200 kpb avec un recouvrement de 10 kpb, cela fait 5 406 séquences à comparer au génome entier. Dans l'hypothèse où une comparaison s'effectue en 2 min, il faudrait plus de 7 jours pour simplement réaliser la première étape de REPET sur un ordinateur ayant un processeur possédant un seul cœur, c'est-à-dire ne pouvant exécuter qu'une seule tâche à la fois.

C'est pour cela que REPET a été conçu pour paralléliser les analyses. En effet, l'exécution de REPET ne peut se faire que sur une infrastructure regroupant plusieurs serveurs de calculs, un cluster. Dans le cadre de la thèse, nous avons eu accès à deux clusters. Le premier, Génotoul, est hébergé à l'INRA de Toulouse. Il est composé de plus de 1 600 cœurs, partageant plus de 15 To de mémoire vive et 800 To de stockage. Il aura permis de réaliser la plupart des calculs nécessaires à l'étalonnage du pipeline et de procéder à une exécution de TEdenovo et de TEannot en une semaine.

Au cours de l'année 2013, le cluster de Génotoul a instauré un nouveau règlement mettant en place des quotas de temps de calcul. Chaque utilisateur s'est vu allouer un nombre d'heures de calcul pour un an. Or, REPET utilise ce quota en seulement quelques exécutions. Heureusement, le laboratoire d'Olivier Panaud de l'université de Perpignan nous a autorisés à utiliser son cluster doté de 300 cœurs, ce qui nous a permis d'analyser le génome complet avec REPET pour la première fois.

Afin de pallier le manque de ressources de calcul en interne, nous avons étudié la possibilité d'utiliser le cloud EC2. Amazon propose parmi ses offres commerciales la location de machines virtuelles. Il est possible soit de les louer à l'année en payant des frais de mise en

Tableau 8 : Estimation du temps de calcul disponible pour un budget de 5 000 $ sur le cloud d'Amazon en fonction instances à la demande

Description des instances

Instance	Arch	vCPU	Mem	Stockage	$/h
c3.l	64	2	3,75	2*16 SDD	0,12
c3.xl	64	4	7,5	2*40 SSD	0,239
c3.2xl	64	8	15	2*80 SSD	0,478
c3.4xl	64	16	30	2*160 SSD	0,956
c3.8xl	64	32	60	2*320 SSD	1,912

Nombre de jours de calcul

Nb MV	c3.l	c3.xl	c3.2xl	c3.4xl	c3.8xl
1	1736	872	436	218	109
2	868	436	218	109	54
3	579	291	145	73	36
4	434	218	109	54	27
5	347	174	87	44	22
6	289	145	73	36	18
7	248	125	62	31	16
8	217	109	54	27	14
9	193	97	48	24	12
10	174	87	44	22	11
11	158	79	40	20	10
12	145	73	36	18	9
13	134	67	34	17	8
14	124	62	31	16	8
15	116	58	29	15	7

Nombre de processeurs

Nb MV	c3.l	c3.xl	c3.2xl	c3.4xl	c3.8xl
1	2	4	8	16	32
2	4	8	16	32	64
3	6	12	24	48	96
4	8	16	32	64	128
5	10	20	40	80	160
6	12	24	48	96	192
7	14	28	56	112	224
8	16	32	64	128	256
9	18	36	72	144	288
10	20	40	80	160	320
11	22	44	88	176	352
12	24	48	96	192	384
13	26	52	104	208	416
14	28	56	112	224	448
15	30	60	120	240	480

MV = Machine virtuelle
Arch = Arechitecture
vCPU = Nombre de processeurs
Mem = Mémoire vive

service puis les heures de calcul (instances réservées), soit de payer les heures de calcul (plus chères, instances à la demande), soit de payer à l'heure sur les machines non utilisées par Amazon (intances spot). Cette dernière option est la plus économique mais Amazon s'accorde le droit de récupérer les ressources utilisées à n'importe quel moment et sans préavis. En collaboration avec l'URGI, nous avons mis au point une version de REPET pouvant être utilisée sur un cluster virtuel chez Amazon. Pour ce faire, nous avons utilisé starCluster. Ce programme permet de déployer plusieurs machines virtuelles pour former un cluster virtuel sur Amazon, rapidement et simplement. La version de REPET consiste en une machine virtuelle sur laquelle REPET est préinstallé et configuré.

Le cloud d'Amazon présente de nombreux avantages, mais les coûts restent trop élevés. Les conditions imposées par les instances spots ne sont pas adaptées à REPET, car ce programme lance de nombreuses tâches mais ne supporte pas l'échec de l'une d'entre elles. Ce qui arrive quand Amazon récupère les ressources lorsque l'on utilise les instances spot. Les instances à la demande restent trop chères. Ainsi, un budget de 5 000 $ et une location de 480 processeurs donnent accès à 7 jours de calculs, soit environ 2 analyses REPET complètes (Tableau 8).

2. Les outils d’analyse maison

Les fichiers GFF générés par REPET sont nombreux et très denses en informations. Dans le but de traiter de manière efficace et rapide ces données, nous avons développé des outils permettant la gestion et l’analyse de ces fichiers GFF.

2.1. GFFtools

Pour chaque nouveau génome, ses différents éléments génomiques doivent être décrits en spécifiant leur emplacement dans le génome, leur nature, leurs caractéristiques. Plusieurs formats de fichiers ont été créés pour gérer ces informations.

Créé à l’Université de Californie à Santa Cruz (UCSC), le format « Browser Extensible Data » (BED) (http://genome.ucsc.edu/FAQ/FAQformat#format1) était utilisé à l’origine pour l’affichage des informations à travers un GenomeBrowser. C’est un fichier tabulé contenant au moins trois colonnes. Elles correspondent respectivement à l’identifiant de la séquence dans laquelle est présente l’annotation, la position où elle débute, et la position où elle se termine. Le format peut être étendu à 12 colonnes. La quatrième colonne contient un nom ou un identifiant de l’annotation, la cinquième colonne un score (exemple : e-value), la sixième, le brin sur lequel se situe l’élément et les dernières colonnes sont utilisées pour modifier la façon dont l’élément est visualisé dans un GenomeBrowser.

Ce format est simple, mais il n’est pas assez souple pour inclure des informations supplémentaires sur le locus annoté comme le consensus ayant permis son annotation et le pourcentage d’identité avec la séquence consensus. Le deuxième type fichier d’annotation le plus utilisé est le format GFF. C’est un fichier au format tabulé contenant 9 colonnes qui décrivent respectivement le nom de la séquence annotée, la source de l’annotation, le type, le début, la fin de l’annotation, un score, le brin sur lequel se situe l’annotation, la phase si l’annotation décrite est un cadre de lecture et les attributs. L’avantage du format GFF face au format BED réside dans la neuvième colonne qui peut contenir une liste d’attributs standards (ID, nom, alias, Parent, Target, Gap, Derives_from, Note, Dbxref, Ontology_term, Is_circular, cf http://www.sequenceontology.org/gff3.shtml), mais aussi une liste d’attributs personnalisés.

Figure 23 : Stockage d'un fichier GFF3 avec GFFtools
Cadre vert : Une ligne d'annotation, les informations sont stockées grâce Annotation.pm
Cadre orange : Annotation d'une séquence, les informations sont stockées grâce au module AnnotationCollection.pm
Cadre rouge : Le fichier GFF3 complet, contenant les séquences annotées ainsi que les séquences au format fasta,
ces informations sont stockées grâce au module GFF.pm

De nos jours, il existe de nombreux programmes pour manipuler les annotations contenues dans ces fichiers ou extraire leurs informations. Gff2sequence [Camiolo et al 2013] et GFF-Ex [Rastogi et al 2014] sont deux programmes qui sont en mesure d'extraire les séquences correspondant aux différentes annotations en fonction de leurs types. Gff2sequene est capable d'effectuer un contrôle de qualité sur les séquences extraites en sélectionnant celles qui ne contiennent pas de "N" ou "X" ; des caractères représentant une base dont la nature chimique n'a pas pu être déterminée ; la présence d'un codon d'initiation, etc.

La suite Bedtools [Quinlan et al 2010] a été mise au point pour comparer un grand ensemble de données de séquençage NGS aux données génomiques disponibles. Elle fournit plusieurs programmes pour déterminer les locus se chevauchant entre deux fichiers BED, trier ou fusionner les éléments, extraire la séquence décrite à partir d'un fichier FASTA ou calculer la profondeur de séquençage. Même si elle peut analyser les fichiers GFF, elle n'est pas en mesure de conserver l'ensemble des informations contenues dans la colonne d'attributs. Comme cette neuvième colonne est importante pour étudier la dynamique des ETs, nous avions besoin d'utiliser un autre outil.

La bibliothèque Bio::Perl propose le module Bio::Outil::GFF qui est capable de lire les fichiers GFF ligne à ligne et de fournir des méthodes pour accéder à son contenu. Pour analyser les fichiers GFF produits par REPET, nous avions besoin de travailler sur l'ensemble des annotations en même temps pour réaliser des opérations comme le tri des annotations ou la fusion des intervalles. C'est pourquoi nous avons développé GFFTools. Il est constitué de trois modules Perl qui sont utilisés pour lire un fichier GFF ligne à ligne et le stocker en mémoire dans une structure de données. Le module Annotation.pm peut stocker les informations de chaque colonne d'une ligne d'annotation d'un fichier (Figure 23, cadres verts). Il donne accès aux attributs de la neuvième colonne, qu'ils soient standards ou personnalisés. Le module AnnotationCollection.pm a été développé pour stocker les informations de la séquence annotée (Figure 23, cadres orange), et stocker une liste d'objets d'annotation. Il permet de passer les annotations une à une ainsi que de modifier la liste en rajoutant ou supprimant des annotations. Il permet ainsi de réaliser des opérations sur cette liste comme trier les annotations, les filtrer, calculer la couverture sur le chromosome, etc. La couche d'annotations produites par REPET peut contenir de mauvaises annotations comme des locus annotés par plusieurs consensus. Afin de limiter ce genre de situation, nous avons

développé un algorithme utilisant une méthode de graphe capable de sélectionner les annotations présentant le plus haut taux de couverture par rapport au consensus. Le dernier module, GFF.pm, est capable de stocker les en-têtes (version de GFF par exemple), ainsi que les séquences nucléotidiques incluses dans le fichier. Il construit une liste d'objets AnnotationCollection qui décrivent chacun une séquence annotée du fichier GFF (Figure 23, cadre rouge).

Ces modules permettent également de procéder à la réduction des intervalles, c'est-à-dire de fusionner les annotations chevauchantes. Ils permettent aussi d'extraire les séquences nucléotidiques, que ce soit celle du locus annoté ou que ce soit la partie du consensus l'ayant annoté.

Le code de ces modules est distribué sous une licence GNU GPL 3, et est téléchargeable sur github (https://github.com/sguizard/GFFtools).

2.2. DensityMap

Plusieurs outils pour la visualisation des données génomiques sont disponibles. Gbrowse, Jbrowse sont très efficaces pour visualiser des régions génomiques de faibles tailles, mais ils ne conviennent pas à une représentation à l'échelle du génome. Les programmes tels que Phénogramme et CVIT permettent de visualiser un génome entier, mais ils ne sont pas conçus pour afficher les répétitions dispersées de grande densité. C'est la raison pour laquelle nous avons développé DensityMap, un programme Perl qui peut visualiser la densité d'un ou plusieurs types d'éléments génomiques (gène, ARNnc, cpg …) le long des chromosomes pour un génome entier. DensityMap peut calculer des cartes de densité pour plusieurs types d'annotations différents, pour plusieurs chromosomes en une image. Il peut aussi calculer et tracer une carte de densité du GC % le long des chromosomes en utilisant la séquence nucléotidique. L'avantage majeur de DensityMap est qu'il utilise directement le fichier d'annotation GFF pour calculer les densités des éléments génomiques d'intérêt le long des chromosomes sans avoir besoin d'informations supplémentaires fournies par l'utilisateur. L'image produite est hautement configurable et les échelles de couleurs utilisées peuvent être personnalisées pour mieux s'adapter aux données représentées.

Guizard et al. BMC Bioinformatics (2016) 17:204
DOI 10.1186/s12859-016-1055-0

BMC Bioinformatics

SOFTWARE Open Access

DensityMap: a genome viewer for illustrating the densities of features

Sébastien Guizard, Benoît Piégu and Yves Bigot*

Abstract

Background: Several tools are available for visualizing genomic data. Some, such as Gbrowse and Jbrowse, are very efficient for small genomic regions, but they are not suitable for entire genomes. Others, like Phenogram and CViT, can be used to visualise whole genomes, but are not designed to display very dense genomic features (eg: interspersed repeats). We have therefore developed DensityMap, a lightweight Perl program that can display the densities of several features (genes, ncRNA, cpg, etc.) along chromosomes on the scale of the whole genome. A critical advantage of DensityMap is that it uses GFF annotation files directly to compute the densities of features without needing additional information from the user. The resulting picture is readily configurable, and the colour scales used can be customized for a best fit to the data plotted.

Results: DensityMap runs on Linux architecture with few requirements so that users can easily and quickly visualize the distributions and densities of genomic features for an entire genome. The input is GFF3-formated data representing chromosomes (linkage groups or pseudomolecules) and sets of features which are used to calculate representations in density maps. In practise, DensityMap uses a tilling window to compute the density of one or more features and the number of bases covered by these features along chromosomes. The densities are represented by colour scales that can be customized to highlight critical points. DensityMap can compare the distributions of features; it calculates several chromosomal density maps in a single image, each of which describes a different genomic feature. It can also use the genome nucleotide sequence to compute and plot a density map of the GC content along chromosomes.

Conclusions: DensityMap is a compact, easily-used tool for displaying the distribution and density of all types of genomic features within a genome. It is flexible enough to visualize the densities of several types of features in a single representation. The images produced are readily configurable and their SVG format ensures that they can be edited.

Keywords: Genome, Visualization, Annotation, GFF

Background

Visualizing the ever-increasing amounts of DNA sequence data for genomic purposes is becoming a great challenge [1]. One solution is to develop genome browsers. The first, and probably the most popular, was the UCSC Genome Browser, which was released in 2002 and used to display human genomic data [2]. Several others, including Gbrowse, JBrowse, Abrowse and Annot-J [3], are now available. They are ergonomically more efficient than the original and include new functions, such as collaborative annotation with web Appollo [4]. These browsers are useful for displaying discrete chromosome regions but are not suitable for visualizing whole chromosomes.

Other tools have been developed for visualizing whole chromosomes. One of the most widely used is Circos [5, 6], which represents chromosomes by arranging them on a circle. It can also be used to plot annotations, quantitative data and relationships between parts of different chromosomes or genomes [7]. However, Circos representations become dense as their complexity increases, which alters the efficacy of their visualization. Two new programs designed to simplify visualization of whole chromosome sequences were released recently. PhenoGram [8] represents chromosomes and uses ideograms, lines, and different coloured symbols to locate information like phenotypes, genes, CNVs, SNPs,

* Correspondence: yves.bigot@tours.inra.fr
UMR INRA-CNRS 7247, PRC, Centre INRA Val de Loire, 37380 Nouzilly, France

etc. While the PhenoGram web-interface is user-friendly, it requires the input files to be in a specific tabulated format rather than a standard format like Generic feature format (GFF), the most common format for annotation files. It also cannot display the density of a specific feature at a given position in a chromosome. CviT (ChromosomeVisualization Tool) [9] circumvents these limitations. It can represent chromosome contents from a GFF file, is readily configurable and the output image can be customized. CViT can also plot the densities of some features along chromosomes

Table 1 DensityMap options

Short	Long	Type	Description
Mandatory options			
-i	–input	string	GFF file name
-re	–region_file	string	A BED file describing sequence regions to plot. It allow to plot specific regions and not the whole seq. Example of file content: 2L[TAB]100000[TAB]200000 2R[TAB]300000[TAB]450000
-o	–output_img_name	string	output image name
-ty	–type_to_draw	string	Type (column 3 of GFF) to draw, strand(s) to plot and colour scale to use **Type:** Match, gene, CDS, etc. **Strand:** - - > strand – (1 Density Map (or DM)) + – > strand + (1 DM) both - > strand - and strand + (2 DM) fused - > Combination of strand - and strand + (1 DM) all - > strand - and strand + and fused (3 DM) **Format:** "Type1 = Strand = colour_scale" **i.e.:** "match = all = 7;gene = both = 4;CDS = fused = 10"
Generic options			
-for	–force	none	Automatically answers yes to picture size validation
-v	–verbose	none	Activate verbose
-h	–help	none	Print help
Density options			
-c	–colour_scale	integer	Number of the colour scale to use
-sc	–scale_factor	integer	Window length (in base pairs) to use
-a	–auto_scale_factor	integer	Maximum picture height in pixels
-ro	–rounding_method	string	Rounding densities with floor or ceiling
-gc	–gc	integer	Colour scale number for density map of the GC % of chromosome, Requires the presence of the sequence in ##FASTA section of the GFF file
Graphical options			
-ti	–title	string	Picture title
-w	–win_size	integer	Picture height in pixels
-sh	–show_scale	integer	Draw scale, with the integer indicating the maximum number of ticks to print on the scale
-str_w	–strand_width	integer	Strand width in pixels
-str_s	–strand_space	integer	Space between strands in pixels
-sp	–space_chr	Integer	Space between chromosomes
-lm	–lmargin	integer	Left margin in pixels
-rm	–rmargin	integer	Right margin in pixels
-tm	–tmargin	integer	Top margin in pixels
-bm	–bmargin	integer	Bottom margin in pixels
-ba	–background	integer	Picture background colour
-la	–label_strand_rotation	integer	Rotation (in degrees) of strand label
-ft_f	–ft_family	string	Text font
-ft_s	–ft_size	integer	Font size

using histograms placed beside the chromosome representation. This tool produces reliable images when the features are not too dense but becomes limited when the density of a feature like interspersed repeats or DNA motifs is high. CViT must also use a GFF file that contains the density of a feature for a given set of windows along a chromosome. As Cvit is not designed to compute these densities, the GFF file must be revised each time the window width is changed. We have therefore developed a program, DensityMap.pl, inspired by CviT, which can produce maps that include the densities of one or more types of features while displaying the whole genome in a chromosome.

Implementation

DensityMap is run with Perl script in the command line and uses the GD::SVG Perl package to produce SVG pictures. DensityMap computes a representation of the density of a feature on chromosomes using one GFF file (GFF2, GFF2.5 or GFF3) describing a chromosome as input. The program plots as many density maps along a chromosome as there are features specified. It can plot a density map for the plus strand, minus strand, or the plus and minus strands, combinations of plus and minus strands, or plus, minus and compiled strands, for each feature. Density is computed using a tilling window without overlap whose length is fixed by the user or automatically computed to produce an output image that fits the maximum image size. All this information can be set by the user in the command line. DensityMap also automatically calculates the density of a feature for each pixelized region of a chromosome, whatever the representation scale used. The way the density of a feature varies along a chromosome is represented using a colour scale from 0 to

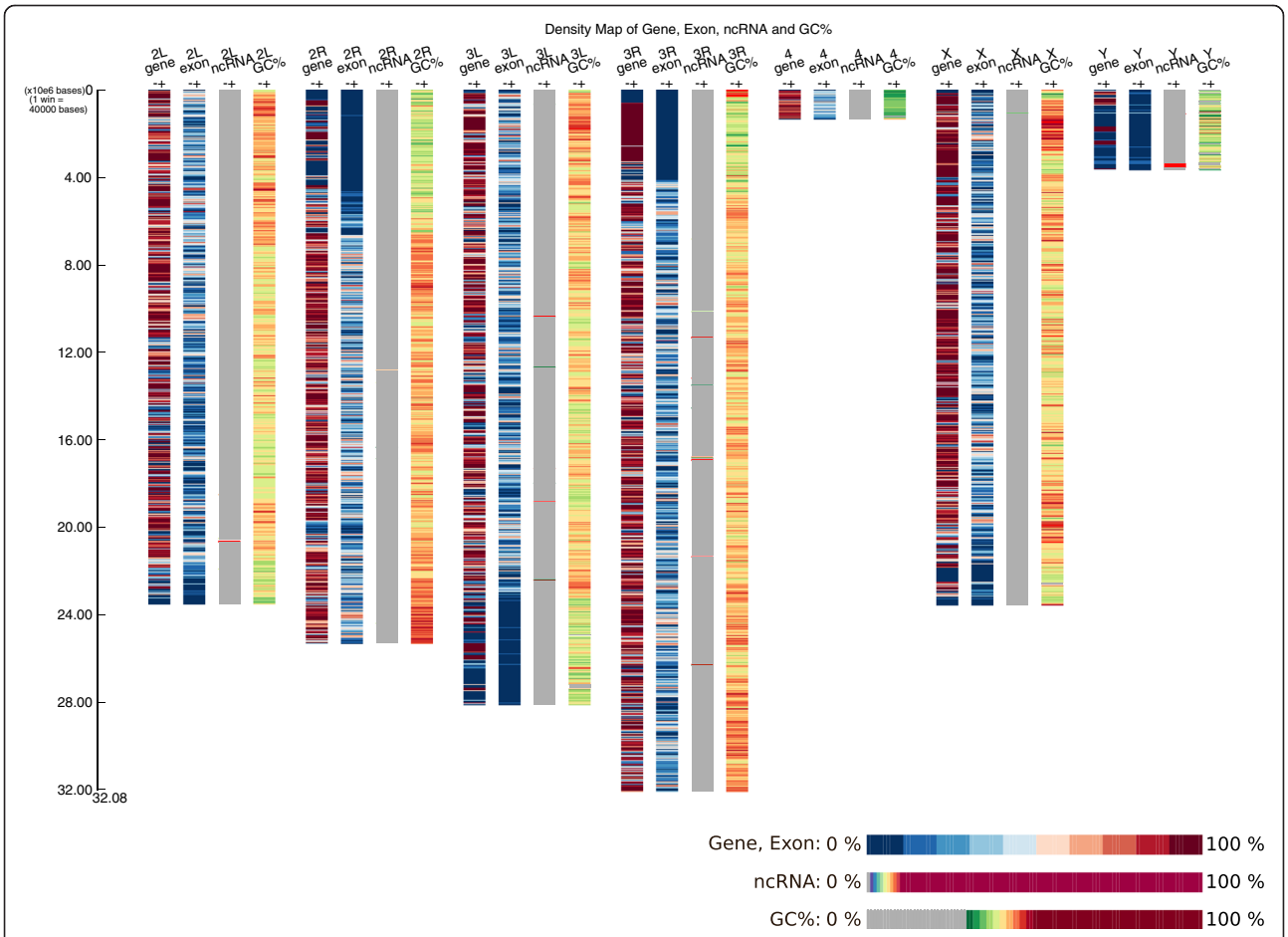


Fig. 1 Density map of genes, exons, ncRNA and GC % in chromosomes of *D. melanogater*. The command line was: DensityMap.pl -i dmel.gff3 -o egn -ty 'gene = fused;exon = fused;ncRNA = fused = 10' -gc 12 -sc 40000 -ba white -str_s 15 -str_w 25 -sp 35 -sh 50 -title "Density Map of Gene, Exon, ncRNA and GC%" -la −15 -ro ceil. The arms of chromosomes 2 and 3 are split into two annotation files 2 L, 2R and 3 L, 3R. Four density map were drawn for each chromosome, one each for genes, exons, ncRNA and GC%. Tilling windows were 40,000 bp long. Densities are represented with three colour scales. That for genes and exons is blue - red with colour tone changes for each 10 % density change. The second, for ncRNA, shows 0 % density as grey, densities of 1 to 9 % are represented by a blue—red colour gradient, and densities of 10 %l or greater by dark red. The third colour scale, for GC content, shows densities below 30 % in grey, 30–49 % as a green—red colour gradient, and densities of 50 % and above in dark red. The scale is in Mbp

100 %. A single colour scale can be used for all features investigated or each feature can have is own colour scale. Like CViT, DensityMap.pl produces visualizations that are fully configurable in a Scalable Vector Graphics (SVG) format. This makes it easy to edit high quality images for publication. The program also includes graphical options for configuring almost all elements (margins, map width, scale, etc.) of the image. The options are shown in Table 1.

The program computes the size of the output image according to the number of chromosomes (GFF files), the number of features to represent, the number of strands to plot and the window size. If the user chooses automatic scale computing, the program calculates a windows size that gives an image that lies within the maximum image size defined by the user. The program asks the user to check the output picture size before processing the data. It then builds the image by adding the various graphical elements (background, title, scale) and processes the data for plotting the chromosome strands. It sequentially opens GFF files, filter features (GFF file third column) selected by the user with the option -ty (types). The intervals are collected and sorted by their beginnings and merged to remove overlaps. Lastly, the program computes the densities - (number of bases

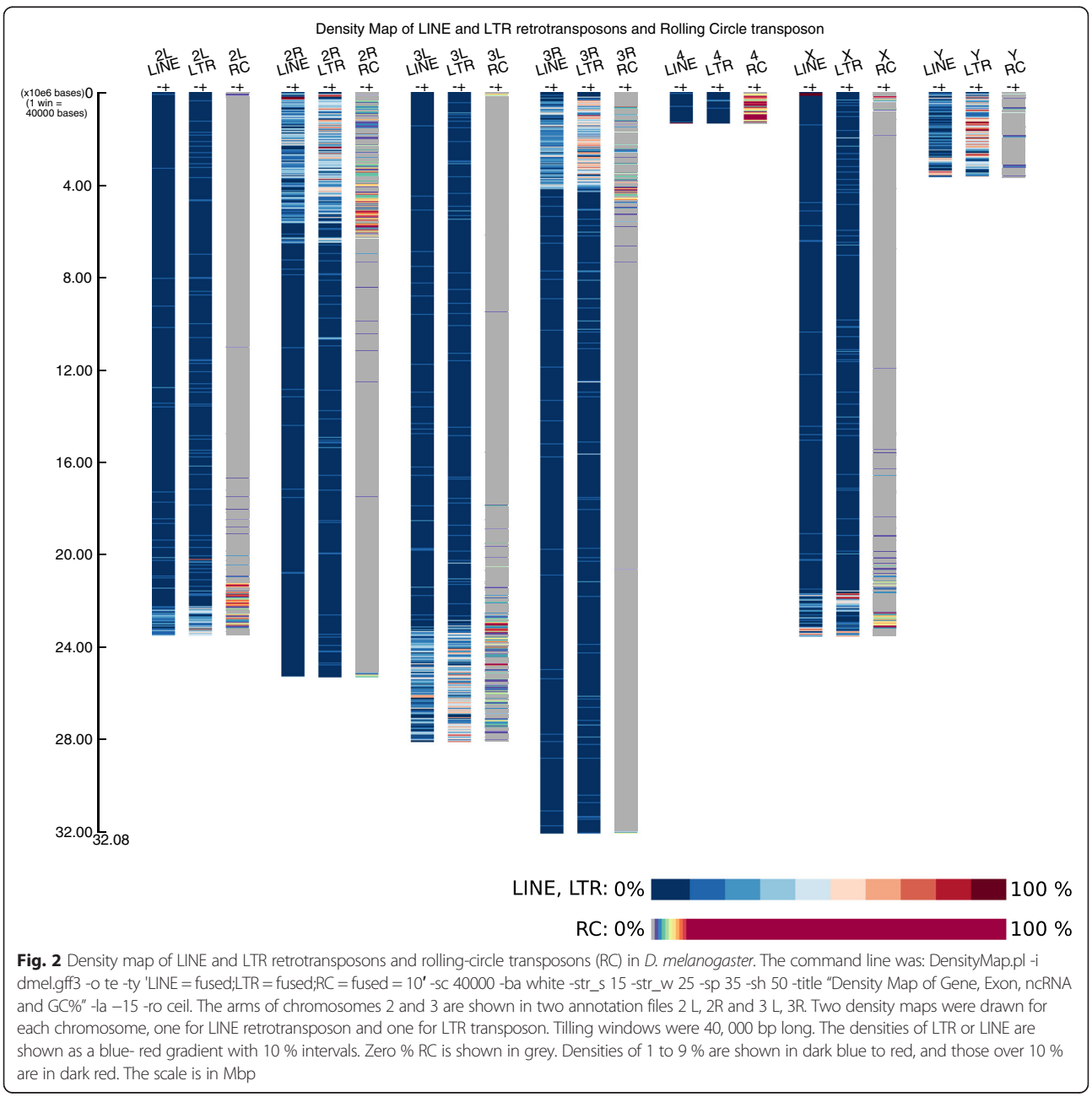


Fig. 2 Density map of LINE and LTR retrotransposons and rolling-circle transposons (RC) in *D. melanogaster*. The command line was: DensityMap.pl -i dmel.gff3 -o te -ty 'LINE = fused;LTR = fused;RC = fused = 10′ -sc 40000 -ba white -str_s 15 -str_w 25 -sp 35 -sh 50 -title "Density Map of Gene, Exon, ncRNA and GC%" -la –15 -ro ceil. The arms of chromosomes 2 and 3 are shown in two annotation files 2 L, 2R and 3 L, 3R. Two density maps were drawn for each chromosome, one for LINE retrotransposon and one for LTR transposon. Tilling windows were 40, 000 bp long. The densities of LTR or LINE are shown as a blue- red gradient with 10 % intervals. Zero % RC is shown in grey. Densities of 1 to 9 % are shown in dark blue to red, and those over 10 % are in dark red. The scale is in Mbp

covered by the feature /window size) x 100 - and then draws it within the image. A synopsis of the main algorithm and functions is supplied in Additional file 1 and a manual in Additional file 2.

Even if the main purpose of DensityMap is to plot whole genome data, it can be interesting to compare specific loci of several sequences. This can be done using the –region_file option. The user has to provide a BED file - a tabular formatted file compound of three column where the first column design the sequence, the second the region start position and the third region end position - describing the region of interest on each sequence. In addition to the density map, the program produce a CSV file - a tabular formatted file - that contain the densities computed for all features, windows and sequences.

Results

We have used DensitMap to examine two examples based on data on the genome of *Drosophila melanogaster (*available at http://flybase.org). The first (Fig. 1) illustrates the capacity of DensityMap to represent features that occur very frequently in a genome. This study is of the genes, exons, regions coding ncRNAs and the GC content of *D. melanogaster* chromosomes. The image produced shows that genes cover very large regions of the chromosomes, are absent from the centromeres and less frequent on the Y chromosome. As expected, the distribution of exons agrees with that of the genes. The representation of the GC content shows that the centromeres are GC-poor while the regions covered by genes are GC-enriched. The terminal regions are different of the rest of the X chromosome in that they are very GC-rich. The image also shows that ncRNAs are evenly distributed throughout the chromosomes, except for the centromeres and chromosome Y and a few regions where the ncRNA density is over 10 %.

The second example illustrates the ability of DensityMap to produce images describing features that occur at extreme (high or low) densities. We looked at the distributions and densities of three kinds of transposable elements (TEs): LTR and LINE retrotransposons and rolling-circle transposons. Rolling-circle transposons like helitrons are present in this genome, but they are much less abundant than LTR or LINE retrotransposons. These features were visualized with colour scales that were appropriate for features present at low density (Fig. 2). The default program setting rounds down values using a floor method that transforms values between 0 and 1 to 0. But, in this case, we selected the ceiling method, which rounds up values between 0 and 1 to 1 and are thus visualized. The densities of the LTR and LINE retrotransposons can also be visualized. Their distributions in the *D. melanogaster* genome are similar, except that LTRs are very dense in the inner regions of the Y chromosome while most LINEs are present at one end. The TEs in chromosomes 2 and 3 are clustered in the telomeres. A large intra-chromosomal region is devoid of repeated elements. Rolling circle transposons are concentrated at the ends of chromosomes 2 and 3 and the arms of the Y chromosome. The red windows seem to indicate helitron hotspots. Helitrons are also present in the inner regions of chromosomes but their densities are very low. There are two hotspots of these TEs on the X chromosome, one in each telomere; they are absent from most of the other regions. The density of helitrons in most regions of chromosome 4 is over 10 %.

Conclusion

The development of sequencing technologies has led to improvements in genome sequence models—they have become better adapted and much more varied. This, in turn, has led to the development of tools for analysing the genome models, such as genome browsers. While these tools are most useful for viewing small regions of chromosomes, very few provide an overall view of the complete genome. CViT and Phenogram provide two solutions, but they also have limitations: non-standard annotation file formats, or not designed to deal with very dense annotation files such as repeated sequences. DensityMap can automatically compute the densities of features to give a series of windows along chromosomes—and this for a complete genome. It is very flexible; it can be used to analyse not just very dense annotations but also low density annotations by applying the computing and graphical options provided. It is also very efficient for plotting density maps of total repeats – satellites, TEs, simple sequence repeats - of human genome – 5 295 850 features – in 2 min 14 second a on computer equipped of a Intel(R) Xeon(R) W3670 CPU @ 3.20GHz and 16 Go of RAM. DensityMap is very simple to install and run, and so is a good way to obtain a global view of genomic data. To make easier the usage of DensityMap to persons non initiate to linux command line, we developed a web graphical user interface for online DensityMap analysis.

Availability and requirements

- Project name: DensityMap.pl
- Project home page: https://github.com/sguizard/DensityMap
- Graphical user interface: http://chicken-repeats.inra.fr/launchDM_form.php
- Operating system(s): Linux
- Programming language: Perl
- Other requirements: Perl module GD::SVG
- License: GNU GPL v3
- Restrictions on its non-academic use: None

Additional files

Additional file 1: Synopsis of the main program and functions. (DOCX 6 kb)
Additional file 2: DensityMap Manual. (DOCX 239 kb)

Abbreviations

BED: browser extensible data; Bp: base pair; CNV: copy number variation; CPU: central processing unit; CSV: comma-separated values; DNA: deoxyribonucleic acid; GFF: generic feature format; LINE: long interspersed nuclear element; LTR: long terminal repeat; Mbp: mega base pair; RAM: random access memory; SNP: single nucleotide polymorphism; SVG: scalable vector graphics; TE: transposable element.

Competing interest

The authors declare that they have no competing interest.

Authors' contributions

SG developed the DensityMap program. BP, YB helped with program design and publication editing. All authors read and approved the final manuscript.

Authors' information

Sébastien Guizard holds a doctoral fellowship jointly funded by I.N.R.A. (PHASE department)/Région Centre, and a training grant for the Ecole doctorale "Santé, Sciences Biologiques et Chimie du Vivant" of the University PRES Centre Val de Loire.
Benoît Piégu is a C.N.R.S. engineer based at the INRA Centre, Tours.
Yves Bigot is C.N.R.S. Research Director at the INRA Centre, Tours.

Acknowledgements

We thank Jérome Salse, Hadi Quesneville and Claire Lemaitre for advice and discussions during program development. Dr Owen Parkes edited the English text.

Funding

This work was funded by the Région Centre (AviGeS Project), C.N.R.S., I.N.R.A., the Groupements de Recherche CNRS 3546 (Elements Génétiques Mobiles) and 3604 (Modèles Aviaires), and the Ministère de l'Education Nationale, de la Recherche et de la Technologie.

Received: 19 January 2016 Accepted: 13 April 2016
Published online: 06 May 2016

References

1. Batley J, Edwards D. Genome sequence data: Management, storage, and visualization. Biotechniques. 2009;46:333–5.
2. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, et al. The Human Genome Browser at UCSC. Genome Res. 2002;12:996–1006.
3. Wang J, Kong L, Gao G, Luo J. A brief introduction to web-based genome browsers. Brief Bioinformatics. 2013;14:131–43.
4. Lee E, Helt G, Reese JT, Munoz-Torres MC, Childers CP, et al. Web Apollo: a web-based genomic annotation editing platform. Genome Biol. 2013;14:R93.
5. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, et al. Circos: An information aesthetic for comparative genomics. Genome Res. 2009;19:1639–45.
6. An J, Lai J, Sajjanhar A, Batra J, Wang C, et al. J-Circos: an interactive Circos plotter. Bioinformatics. 2015;31:1463–5.
7. Pont C, Murat F, Guizard S, Flores R, Foucrier S, et al. Wheat syntenome unveils new evidences of contrasted evolutionary plasticity between paleo- and neoduplicated subgenomes. Plant J. 2013;76:1030–44.
8. Wolfe D, Dudek S, Ritchie MD, Pendergrass S. Visualizing genomic information across chromosomes with PhenoGram. BioData Mining. 2013;6:18.
9. Cannon EKS, Cannon SB. Chromosome visualization tool: a whole genome viewer. International J plant genom. 2011;2011:373875.

3. Ré-annotation et re-découverte du modèle Galgal4

La stratégie la plus largement employée pour annoter les répétitions dans les génomes animaux est d'utiliser le programme RM avec la base de données de séquences consensus d'ETs Repbase (ISB). Cette approche a permis de décrire les génomes aviaires comme ayant le plus faible contenu en répétitions (8-10 %) parmi les génomes d'espèces de vertébrés séquencés (30-55 %). À titre de comparaison, le génome de Xenopus Tropicalis, dont la taille du génome est similaire à celle du poulet (1,36 Gpb), a un taux de répétitions estimé à 32 %. L'une des limites inhérentes à l'utilisation d'une méthode de bibliothèque est que son efficacité dépend de la base de données de séquence utilisée. Cependant, d'autres méthodes existent depuis une dizaine d'années. Celles-ci font une identification *de novo* des répétitions et peuvent donc être plus appropriées. Ici, le modèle de génome Galgal4 (1,04 Gpb) a été analysé à l'aide d'une stratégie d'annotation impliquant plusieurs outils *de novo* complémentaires pour évaluer son contenu répétitif et caractériser les répétitions simples (SSR), des répétitions en tandem, et des ETs.

L'utilisation des annotateurs *de novo*, que sont REPET et TRF nous a permis de réaliser pour la première fois l'annotation *de novo* d'un génome de plus de 1 Gpb. Nous avons ainsi pu montrer que le génome de la poule rouge de jungle contient environ 19 % de SSR et d'ETs pour un contenu total en répétition de 31 à 35 % dans le génome nucléaire. Nos contrôles ont également révélé que les méthodes de bibliothèque ont tendance à surestimer la diversité en ETs, car nous avons mis en évidence la présence de seulement 34 espèces d'ETs dans Galgal4 et démontré l'absence de SINEs ou d'ET *Chompy*. Ces résultats apportent des informations importantes sur le génome de la poule rouge de jungle.

Nous avons obtenu ces résultats grâce à l'utilisation d'outils *de novo* pour annoter les répétitions dans les grands génomes animaux. Ils révèlent également des problèmes qui devront être résolus afin de développer des normes méthodologiques standards pour annoter les répétitions dans les génomes eucaryotes.

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

Sébastien Guizard[1], Benoît Piégu[1], Peter Arensburger[1,2], Florian Guillou[1], and Yves Bigot[1,3]

[1] Physiologie de la reproduction et des Comportements, UMR INRA-CNRS 7247, PRC, 37380 Nouzilly – France

[2] Biological Sciences Department, California State Polytechnic University, Pomona, CA 91768 - United States of America.

[3] Corresponding author: PRC, UMR INRA-CNRS 6175, 37380 Nouzilly, France. Tel: +33 2 47 42 75 66

e-mail addresses: sebastien.guizard@gmail.com, benoit.piegu@tours.inra.fr, arensburger@gmail.com, florian.guillou@tours.inra.fr, yves.bigot@tours.inra.fr

Abstract

Background: The program RepeatMasker and the database Repbase-ISB are the main constituents of the most widely used strategy for annotating repeats in animal genomes. They have been used to show that avian genomes have a lower repeat content (8-12%) than the sequenced genomes of many vertebrate species (30-55%). But the efficiency of such a library-based strategy is limited by on the sequence database used. Alternative, perhaps more powerful, methods that identify repeats *de novo* have existed for a least a decade. We have used an annotation strategy involving several complementary de novo tools to determine the repeat content of the model genome galGal4 (1.04 Gbp) and to identify simple sequence repeats (SSRs), tandem repeats and transposable elements (TEs).

Results: We annotated the over-1-Gbp-long galGal4 genome and showed that it contains approximately 19% SSRs and TEs and that the total repeat content of the real genome of the red jungle fowl is 31-35%. Our controls also revealed that library-based methods tend to overestimate TE diversity; we could validate the presence of only 34 TE species in galGal4. These results have a major impact on current understanding of the repeats distribution through the chromosome organization of the red jungle fowl.

Conclusions: Our results are a proof of concept of the reliability of using de novo tools to annotate repeats in large animal genomes. They have also revealed issues that will need to be resolved in order to develop gold-standard methodologies for annotating repeats in eukaryote genomes.

Keywords: satellite DNA / transposable elements / bioinformatics / benchmarking / repeat

Background

Repeated sequences are the most abundant components of many eukaryote genomes. They account for about 25% of the fruit fly, *Drosophila melanogaster,* genome [1,2], 50-69% of the human genome [3] and nearly 90% of the maize, *Zea mays,* genome [4]. The repeated sequences in eukaryote genomes vary in their structure, organization and location in chromosomes. The primary criterion is often their distribution profile in chromosomes, i.e. their organization in stretches of tandem repeats or the interspersion of copies.

The most highly repeated sequences generally lie near or within centromeres and telomeres. The tandem repeats within a chromosome segments may contain tens to several thousands of units. The two main types are: stretches of (TTAGGG)n repeats at telomere ends [5], and satellite DNAs composed of tandem repeated units of 60 to a few thousand bps. A eukaryote genome may contain one or more families of satellite DNA; the DNA sequence of the repeated unit and the abundance of each family depends on the species [6,7].

Other kinds of tandem repeats are also found in inner regions of the chromosome arms. They are known as simple sequence repeats (SSRs). The first group includes short stretches of tandem repeats with an uncomplicated sequence that are dispersed along chromosomes. This group has been subdivided into three types depending on the complexity of their repeated unit. Simple sequence repeats are stretches of A and T or C and G nucleotides. Microsatellites and minisatellites, also called variable number tandem repeats (VNTR), are 2 to 10-bp long (micro) or 11 to 60-bp long (mini) [8,9]. The other types of tandem repeats result from duplication of chromosome segments and can correspond to tandemly repeated genes such as those encoding ribosomal RNA (rRNA), immunoglobulins, chromosome regions, or even genes. These last repeats are called copy number variations (CNVs); they occur when the number of tandem repeats varies between individual alleles within a species [10,11].

The features of dispersed repeats generally result from their ability to move from one locus to another and to be amplified within chromosomes using a variety of transposition mechanisms

including “cut-and-paste” or “copy-and-paste” [12]. Although the extent of their diversity, origins and classifications are still debated, the vast majority of dispersed repeated sequences, or mobile genetic elements (MGEs), or transposable elements (TEs) in eukaryotes, have one of four phenotypes based on the organisation of their sequence (for review [12]). Three of these phenotypes include TEs that use RNA as a mobile intermediate. This RNA molecule is transcribed from a genomic copy that will later be reverse-transcribed into a DNA molecule during, or prior to, insertion at a new chromosomal site. The first of these three phenotypes corresponds to TEs that resemble retroviruses. They have long terminal repeats (LTR) and contain three open reading frames (ORFs) that encode: a group-specific antigen (Gag), a reverse transcriptase (RT), and a retroviral envelope protein (Env). These TEs are more commonly known as an LTR retrotransposon or endogenous retroviruses. The second phenotype TEs have no terminal repeats and two ORFs coding for proteins similar to Gag and RT and a poly-A 3' tail. These are known as non-LTR retrotransposons or retroposons. The third phenotype TEs are short interspersed elements (SINEs) that are derived from transcripts of host genes encoding structural RNA molecules (tRNA, 7SL RNA, 5S RNA, etc). SINEs move around using the transposition machinery of some non-LTR retrotransposons. These SINEs use a single or a double-stranded DNA molecule as a transposition intermediate. This intermediate is excised or produced by DNA replication from a genomic copy and then inserted at a new chromosomal site. These TEs are commonly called DNA transposons.

Because repeats are often abundant in eukaryotic genomes, annotating them requires considerable effort. TEs are a particular challenge because eukaryotic genomes generally contain from tens to hundreds of different “TE species” and the abundance of each one may vary considerably. Despite this diversity, only very few copies of these “TE species” are active transposed, the vast majority are inactive remnant copies with sequences that have accumulated a number of nucleotide mutations and rearrangements over time, depending on the age of each TE species in its host genome. There is currently no validated standard strategy for locating and annotating repeats in eukaryotic genomes. This problem has recently been the subject of a call to

benchmark methods for annotating transposable element in order to optimize the reporting of their efficiencies and to clarify the nature of the problems encountered [13]. The three most commonly used approaches are: library-based methods, signature-based methods, and *de novo* consensus methods (for review [14,15]). RepeatMasker (RM) is the most widely used library-based method in genome sequencing projects [16,17]; it is generally used in association with Repbase, a library source that is freely available to academics [18]. The TEs in numerous genomes have been annotated with RM and a private, inaccessible library of the Institute for System Biology (ISB) [19]. The main limitation of this approach is that the annotations depend on the quality of the reference database, including completeness and accuracy of the consensus sequences. Signature methods, in contrast, use traits unique to certain TEs or repeats. For example, LTR Finder detects the specific DNA organization and a chain of signatures (motifs) specific to retroviruses to detect LTR-retrotransposons [20]. Tandem repeat finder (TRF) is a signature method tool dedicated to the detecting all types of uncomplicated tandem repeats like simple repeats, microsatellites, minisatellites and satellite DNAs [21]. Finally, DNA *de novo* consensus methods combine a range of detection tools. The REPET package is a pipeline that uses both *de novo* and signature-based methods [22-24], and may even include a library-based step [25]. *De novo* consensus methods such as REPET have been limited until now by their need for powerful resources for calculation and storage. This has restricted their application to small eukaryotic genomes (~10 Mbp to 500 Mbp). However, the evolution of computing clusters and a recent REPET update have open the way to the treatment of bigger genomes such as those of vertebrates.

Our work has focussed on the analysis of the repeats in the smallest vertebrate genome (just over 1 Gbp): the red jungle fowl (RJF) *Gallus gallus*. The RJF and avian genomes, except for some Falconiforme species [26], are all composed of a several macrochromosomes (RJF has 9: 1 to 8, depending on their physical size, plus the Z sex chromosome), and many microchromosomes (RJF has 30 : 9 to 38, plus the W sex chromosome) [27]. The RJF genome was the third vertebrate genome to be sequenced and one of the few from a vertebrate species for which a physical map was

used to construct the first version of the genome model galGal1 [28]. This genome model was then improved in several steps [23-31] until the galGal4 release in November 2011. And this genome model is constantly being improved with regular updates to its gene annotation [32]. Hence, galGal4 must be considered to be a model of the RJF genome that still lacks some information lacking. The complexity of the RJF genome, its C-value, which reflects the amount of nuclear DNA in the haploid genome, was first evaluated by reassociation kinetics at about 1.1 to 1.4 pg for a female genome [33,34]. Several subsequent investigations using flow cytometry set its C-value at 1.25 ± 0.06 pg [35-38]. The RJF genome can be compared galGal4 by converting the C-value to an absolute number of bp by multiplying it by the unit atomic mass (1 u = 1.660539×10^{-27} kg). This gives a mean mass per nucleotide pair of 1.023×10^{-9} pg/bp and an average DNA genome density of 978 Mbp/pg [39]. The C-value, in bp, of RJF is therefore 1.223 ± 0.058 Gbp, while that of Galgal4 is 1.047 Gbp (including 14 Mbp of 'N-stretches'). Thus the difference between the size of the genome determined by flow cytometry and Galgal4 is 175 Mbp (14 %). The origins of this difference have been examined. First, there are missing sequences because nearly all the regions overlapping the megacentromeres [40] and megatelomeres [41-43], and their neighbouring satellite DNAs [44] are absent from galGal4. This lack has been estimated at about 8 % of the real RJF genome [40-44]. Similarly, RJF has no tandem repeats encoding the 18S-5.2S-28S (~400 copies) and 5S (~100 copies) rRNA [45], representing about 1% of the real genome. A third possible source of the size difference is that features that are part of AT-rich or GC-rich regions, or regions containing short motifs are not always included in Illumina libraries, so that they are lost from subsequent banking, sequencing and/or assembling in scaffolds and chromosomes [46-52]. These sequences are probably responsible for the 8 unassembled chromosomes in Galgal4 (numbers 30, 31 ,33 ,34, 35, 36, 37 and 38, two of which correspond to LGE22 and LGE64; Figure 1) [28,32]. They could also be the source of the anecdotal chromosome 32 (1028 bp) and the dwarfism of chromosome 16 (535,270 bp; its real size it estimated to be close to 11 Mbp) because of its high repeat content [53]. These limitations are also responsible for the fact that most of the other

chromosomes in galGal4 (Figure 1) are smaller than in the real genome [54], and probably for the fact that avian genomes lack at least some of the ~6000 protein-coding genes that are present in all mammals [32, 55-57].

A clear picture of the reasons for the difference in the sizes of galGal4 and the real genome is important for understanding the number and sizes of the repeats in the real RJF genome and the galGal4 assembly. Reassociation kinetics indicate that the real genome contains approximately 32% repeats [58,59]. As galGal4 lacks sequences constituting the centromeres, telomeres, the clusters encoding the rRNA and a part of the satellite DNA, the total rate of repeats in the genome model should be 22 - 24%. Successive investigations, mainly RM, have given rates that have increased over time: 9.5% in 2004 [28], 8% in 2005 [60], and 11.47% in 2011 [61] (Table 1), but are still significantly lower than those calculated from DNA reassociation kinetics. This suggested that the analysis of repeats in the galGal4 assembly needs further investigation.

We have therefore re-investigated the status of repeats in the galGal4 assembly using a mainly *de novo* annotation strategy that involves several complementary methods of detection and annotation. Our use of this strategy enabled us to detect repeats in galGal4 in abundances that are closer to those predicted by physicochemical data. Analysis of these new annotations has shed new light on the genome in terms of how its components are organised,including TE diversity, distribution, and dynamics. Finally, we discuss the benchmarking of various methods used in our investigations in the hope of stimulating debate that may lead to the definition of a gold standard for annotating repeats in assembled genome models.

Results and discussion

Evaluating the amount of repeats in galGal4 *in silico*

It is essential to be able to assess the amount of repetition for the annotation of a genome. DNA reassociation kinetics can be sued to define a minimal proportion of repeats, but the 22-24% proportion found for the RJF is only a minimal value because its calculation is limited by two parameters of the experimental procedure [62]. First the ability of this technique to detect repeats in a genome depends on the length of the fragments used (generally 200 - 250 bp). Second many of the repeated sequences in a genome like that of RJF are old [61], and some of them repeats might be recovered in the unique component of the DNA reassociation kinetics results because they are very rich in mutations. Some studies that have used more stringent reassociation conditions found an average repeat rate of 13% in the RJF genome [34]. An advantage of certain *in silico* approaches is that they can detect very short because they can be set to be extremely insensitive to the minimal size of repeated sequences and their sequence divergence. We selected two such methods based on two approaches of k-mer analysis, P-clouds [63] and Red [15] (Additional File 1).

The overall amounts of repeats in galGal4 detected with P-clouds (33%) and Red (29.9 %) were close. These *in silico* evaluations were approximately 50% higher than the minimal ones obtained with DNA reassociation kinetics. As a control we tested the reliability of both methods using two published genomes with well-known repeat contents: *Anopheles gambiae* and *Drosophila melanogaster*. Analysis of these genomes is facilitated by the fact that their "TE species" sequences are rather well-conserved. Despite the similarity of the P-clouds and Red estimations, we find that Red is the most appropriate program for calculating a reliable rate of repeats because most of the sequences that it identified as repeats were also annotated in the TE annotation (Red identified 84% of annotated TEs and P-clouds 61%, Additional file 2).

Detection and annotation of repeats in galGal4

Strategy for detecting and annotating repeats in Galgal4. We used published data and/or practical evaluations such as those described above to evaluate the efficiency of several programs (or alternate sources of information for detecting and annotating of repeats altogether or per types such as SSRs, TEs, dark matter and CNVs) in order to develop our strategy for analysing the galGal4 genome. The resulting strategy (Figure 2) was organized as five component tasks. We first used Red to approximate the total amount of repeats [15]. In the second we used TRF to analyse SSRs [21]. The third task, TE annotation, demanded investment of most resources. We used the REPET package [22-24] because it was the best-documented and had been shown to be more efficient than the RepeatScout [65] and RepeatModeler [66] packages. We decided to use REPET even though its annotation does not cover 100% of annotations calculated by the other two packages [13]. These small differences in annotation were largely circumvented by first using TRF, which we have found to be more efficient at locating SSRs than either the REPET, RepeatScout or RepeatModeler packages. Then, performing a REPET analysis in three successive detection steps (Figure 2) enabled us to go deeper than RM into the detection and annotation of fragmented repeats. The forth step consisted of annotating the dark matter (DM) as proposed by Maumus et al (2014) [25], using a library containing all repeated copies longer than 500 bp detected in step 3 and the TEannot program [67] rather than RepeatMasker (RM) [68]. Finally, we used the available annotation of CNVs in galGal4 [69].

Profiles of SSRs in galGal4 (STEP2). The total amount of SSRs in galGal4 was evaluated with RM at 1.73% [61]. We used the FASTA program of the GCG computer package [8] and *sputnik* [9] to investigate the diversity and amounts of microsatellites. Satellite DNAs have been investigated using different molecular approaches (for a review see [44]). We reinvestigated the features of both of these SSRs using TRF, which can requires only a reasonable amount of computing time to analyse a whole genome and can detect SSRs with repeated units from 1 bp to 2 kbp. We found that the assembled genome contained 3.73 % SSRs and the unassembled genome contained 12.74 %

SSRs, for a total coverage of 4.08 % in galGal4. These amounts are at least twice as large as those found with RM (2.36-folds with rates varying from 1.11 to 9.13-folds, depending on the chromosome; Table 2 and Figure 3a). We then went on to look at the features of each type of SSRs. We identified 4 SSR types based on the complexity of their repeated unit sequence: simple repeats, microsatellites, minisatellites, and tandem arrays with repeated units 60 - 2 kbp long that were selected when they were composed of at least 2 repeats. We divided these large tandem arrays into two categories: large tandem repeats (<50 repeated units) and satellite DNAs (>50 repeated units). The coverage of the various types of SSRs in chromosomes indicated that the overall densities of simple repeats and microsatellites were similar. In contrast, minisatellites and tandem arrays were more abundant in some of the galGal4 chromosomes (16, 21, 22, 23, 25, 26, 27, 28, LGE22, and LGE64) and similar in others (Figure 3b). The amounts of the various categories of SSRs are summarized in Table 3 and their features are shown in the Additional File 3.

de novo ***detection and annotation of dispersed repeats (STEP3).*** The REPET pipeline was used to detect repeats and produce annotations. It is composed of two sub-pipelines, TEdenovo, that detects repeats using a *de novo* method based on the repetition of sequences, and TEannot, that produces annotations using a combination of programs and post-processes (see Additional File 4; [22-24]). We used an iterative strategy involving three runs of the REPET pipeline to completely annotate galGal4 (Figure 2) and a version of galGal4 from which the SSRs in chromosomes and a 9 Mbp satellite DNA composed of ~22 kbp repeated units in the Z chromosome had been removed [70]. The first run (REPET1; Figure 2) reported 3926 consensuses (Library 1) corresponding to repeated sequences. These were filtered with TEannot to eliminate residual redundancy between certain consensuses and those which had no full-length copy in galGal4. The resulting 790 consensuses (Library 1f) were then used to annotate galGal4 to extract the annotated repeats and calculate a reduced version of galGal4. The second REPET run (REPET2; Figure 2) was performed with the reduced galGal4 and gave 186 consensuses (Library 2); these were filtered and 133 new

consensuses selected (Library 2f). The Libraries 1f and 2f were merged in REPET step 3 (Figure 2), and filtered manually to remove redundant sequences and sequences corresponding to tandem repeats and segmental duplications (Library 3; 613 consensuses). Lastly, these were filtered using TEannot to obtain 581 consensuses that were reduced to 499 (Library 3f) by manual curation to eliminate consensuses corresponding to pseudogenes. The final annotation of galGal4 was calculated with Library 3f using TEannot and revealed a TE coverage of 11.524 % (Figure 4a).

Detection and annotation of highly divergent repeats, mining the dark matter (DM; STEP4). Genomic dark matter may be defined as "all intergenic sequences, irrespective of functionality or expression" [71-73]. Scientific interest in dark matter was triggered by the discovery of non-coding RNAs (ncRNAs) that could regulate gene expression. Several reports have shown that dark matter is a source of ncRNA and that it can cause disease when it malfunctions [74-76]. Today's studies on dark matter are designed to annotate non-coding RNAs using RNA-Seq, cDNA sequencing, tilling arrays or to annotate cis-regulatory DNA elements using Dnase-seq [73,77]. Because genomes have undergone bursts of TE production during their evolution and because these TEs are actively repressed [78], dark matter could also be considered as a graveyard containing very different, recombined TE copies. Repeats with sequences that are rather well conserved, whose integrity is not much reduced by too many point mutations or recombination events, can be annotated using the default values of certain parameters in the REPET pipeline. We used a library containing all repeated copies of the REPET annotation and the TEannot program to access the DM, the older and/or fragmented TE segments (Figure 2, STEP 4). Our aim was to use a population of genomic copies as a probe to detect more divergent repeats (See Methods). Limitations of computational time reasons obliged us to select only TE annotations >500 bp (33757). The 33,757 annotations used at this step each originated from 222/499 consensuses calculated by REPET. Annotation of the DM increased the TE coverage in galGal4 of ~4.7 % (Figure 4b).

Finishing of the repeat annotation. A characteristic of TEannot output files is that each TE copy (i.e. all TEs corresponding to complete elements, internally deleted elements, 5' or 3' truncated elements and elements truncated at both ends) can be split into several annotations linked to different consensuses belonging to a single TE model. Therefore, we prepared the inventory of TE copies in galGal4 by processing the final annotation with GFFtools to resolve and merge stacked and juxtaposed annotations. The minimal TE copy size was set at 20 bp, 4-bp larger than that of the oligo used as a motif to study repeats in Red and P-clouds and 10-bp larger than that used in the ISB annotation.

Post processing the DM increased TE coverage by ~45%, with the [TE+DM] annotation covering 15.7% of galGal4 (Figure 4c, Additional file 5). This amount of TE must be compared to the 9.74 % in the ISB TE annotation [61]. Almost all (99.7%) of the DM annotations (4.41 % of coverage in galGal4) were new and only 0.3% of them extended existing REPET annotation (0.035 % of coverage in galGal4). The sum of SSR and [TE+DM] coverages suggests that there are at least 19.78% repeats in galGal4. But this evaluation was corrected by intersecting SSRs and [TE+DM] annotations with bedtools (Figure 5a). Because [TE+DM] includes 1% coverage by SSRs, the amount of annotated repeats was 18.78%, which was 1.64 times more dense than the ISB annotation [61]. Intersections were also calculated with the CNV annotation [69], as were those obtained with Red and P-clouds. They revealed that ~7.9% low-repeat sequences (Figure 5a, 6.26%+1.62%), corresponding to CNVs, could be added to the 18.78% of repeats, for a total of 26.7% repeated sequences in galGal4. Intersections between the Red or P-clouds annotations and other annotations ([TE+DM], SSRs and CNV) led to embarrassing results regarding the ability of these two methods to calculate reliably the total amount of repeats in a eukaryotic genome. We found that 30% of the [TE+DM] annotations (4.43/15.7% coverage in galGal4) were not identified by Red and 53% of the Red annotations (15.8/29.9% coverage in galGal4) had no counterparts among the [TE+DM], SSRs and CNV annotations (Figure 5a). The results are even more damning with P-clouds since 72% of

the annotations had no counterparts among the [TE+DM], SSRs and CNV annotations (Additional file 6).

As the fragmentation of annotated copies could lead to artefacts during the TEannot step we investigated the quality of the de novo [TE+DM] annotations. We first examined the size distribution of annotations resulting from Red for the overall amount of repeated sequences, TRF for the SSRs, REPET for the TEs, TEannot for the DM and [TE+DM] and the CNV (Figure 5b). This revealed that the range of annotation sizes calculated by Red covered the sum of those of the six other categories and that 90% of the TE copies were 20 bp to ~1100 bp. The size distributions of annotated copies for each kind of repeat were then compared to those of the ISB annotations (Figure 5c). The size distributions of LINE annotations were similar to those of the ISB, while those of the LTR, terminal inverted repeats (TIR) and SSR repeats were smaller. This was expected, since DM annotations were derived from more fragmented TE copies. We next analysed the diversity of repeats described in the [TE+DM] annotations, their common points with those in the ISB annotations and their qualities. The first thing to note is that the patterns of coverage by each TE type of 3 chromosomes (16, 32 and W) were very different from those of other chromosomes (Figure 4a and b). These different profiles are perhaps due to the size of the galGal4 chromosome 32 model (1028 bp), the greater amount of LTR-retrotransposons in chromosomes 16 and W, or even to the distribution profile of some TEs that might not be randomly distributed. This and similar issues are discussed below.

Diversity and Features of TE models in the [TE+DM] annotation

Ranking dispersed repeats within a TE "species" or repeat. Each TE or repeat "species" in libraries like Repbase or that of the ISB is defined by a consensus sequence which may be thought of as the sequence closest to an averaged sequence from a population of copies originating from a single genome. Potential protein coding capacity may also play a role in defining these consensus

sequences. The methods used to calculate these nucleic acid and protein consensus sequences are unpublished. As these consensus sequences do not represent all sequence variations they are of limited value for detecting TEs. Platforms such as Dfam [79] were developed to circumvent this issue, they use the nhmmer search [80] to annotate repeats using a library of hidden Markov models that is set up from existing populations of sequenced elements. Although Dfam improves significantly the sensitivity and takes better account of TE sequence variations, it is still of limited use for detecting the diversity of rearrangements of TEs like the non-LTR retrotransposons and, to a lesser extent, some LTR-retrotransposons and DNA transposons.

We have borrowed the concept of the TE model developed by the creators of the program RepeatExplorer [81,82] to describe a "TE species". This concept is also included in the philosophy of REPET [22,24]. It assumes that a TE model is composed of a main consensus sequence (the most complete version of the TE) plus all the consensuses detected as variants. Using this concept, our final 3f library contains 499 consensuses distributed among 34 TE models (TEs or repeat "species"; Table 4, correspondences between Repbase and ISB consensuses and the 34 TE models are shown in Additional File 7). The final clustering steps were performed manually using information from sequence databases because BLASTclust in TEdenovo cannot calculate models consistent with galGal4. Our result of 34 TE models is in striking contrast with the ISB annotation [61], which describes a TE diversity involving 317 different TE consensuses (TEs or repeated "species", from which 65 consensus sequences corresponding to repeated genes encoding structural RNA - tRNA, U RNA, 5S RNA, rRNA, etc - must be removed.). The many Repbase and ISB consensuses corresponding to non-gene repeats (252) was partly due to fragmentation of a significant number of repeats in several consensus sequences related to a single TE species. Thus, 21 TE models were split into 81 different Repbase and ISB consensuses (Additional File 7), indicating that there were 171 Repbase and ISB TE consensuses involved in the ISB annotation [61] representing 81,805 annotations covering 2 % of the genome without any equivalent among our 34 TE models. Conversely, 13 of our models have no corresponding sequence in Repbase/ISB TEs.

TE models in the [TE+DM] annotation. Our results confirm those of previous studies [60,61] showing that there are three main types of TEs in the galGal4 (Table 4), whose coverages are very different (Figure 4a and b): non-LTR retrotransposons (LINEs; 1 TE model), LTR retrotransposons (LTR; 21 TE models), and DNA transposons (TIR; 4 TE models).

The galGal4 model contained a single "species" of non-LTR retrotransposon, CR1. These were the most abundant TEs with 413,857 copies representing 66.47% of the [TE+DM] annotation (Table 4, Figure 4a and b). In the light of this new inventory, we re-investigated their diversity and found 8 sub-families (Additional file 8).

The copy number of the 33 other models of TEs and repeats varied from 22 to 67691 and together represented 33.53 % of the REPET annotation. Twenty one “species” of LTR-retrotransposons were found in the REPET annotation of galGal4 (Table 4). These were present as copies with two LTRs or solo LTRs resulting from loss of the inner part of the LTR retrotransposon by recombination between the LTRs of each inserted element (Table 4) [83], or both forms. We found no copies corresponding to complete, internally deleted, or partly truncated element of six models of solo LTRs (putative_LTR_group 4, 9, 12, 22, 28 and 30). But the REPET annotation identifies new LTR-retrotransposon “species” whose sequence indicates that they have recently diverged. This includes the retroCalimero, retroSaturnin and retroTux (Figure 6 and Additional file 9), and 4 species of old LTR-retrotransposons (Ancestral_LTR_group1 to 4; Table 4) of which only large internal fragments with damaged frames coding for Gag, RT and-or Env remain in galGal4 chromosomes. We retained the division into four TE models as previously proposed for DNA transposons [60] (Table 4), keeping in mind that they originated from only two species of DNA transposons. Galluhop was an internally deleted form of Mariner1_GG, and Charlie-Galluhop resulted from the insertion of one Galluhop element into a Charlie element before amplification of this chimeric element by Charlie-mediated transposition within chromosomes. We also found 27

consensuses within 8 TE models (Table 4) whose sequence features did not match those of one of the three types described above or with any other known eukaryotic TE [12].

These 34 TE models were completed manually using published data [84,85]. This identified four more TE "species" whose few copies in galGal4 made them undetectable using other annotation strategies (Figure 2). Two LTR elements, the Rous sarcoma virus (RSV) and the Avian myelocytomatosis virus (AMAV), were integrated as single complete copies into chromosome 1. We also found several repeated segments corresponding to an inner region of the RSV genome in chromosome 20. And three ancient LTR-retrotransposons had become domesticated in neogenes; these were found near the origin of the *ENS1, 2* and *3* genes [86], the *OVEX1* gene [87] and the *map1-like* gene (Acc N°: XP_003641886.2) on chromosomes 2, 15 and 10, 14 and 10, respectively. We also found remnant copies of DNA transposons, such as a *Polinton* TE [12,88] on chromosomes 2 and Z. These remnant sequences still contained interrupted frames coding for an RVE integrase and a *Megaviridae*-like major capsid protein on chromosome 2, and DNA polymerases B on chromosome 2 and Z (the best conserved was on the Z chromosomes). These regions are conserved in chromosome 2 and Z of the *Meleagris gallopavo* (turkey) genome; the coding frames for the DNA polymerase B on the Z chromosome are the easiest to elucidate (Additional file 11). There were also traces of a wide variety DNA transposons within 27 neogenes coding for transposase-derived proteins, all of them must have emerged before the evolutionary separation of the mammalia and sauropsida lineages (Additional file 12).

Differences between ISB and [TE+DM] annotations. As indicated above, the two annotations were not in complete agreement (Additional file 13), 171 Repbase and ISB TE consensuses involved in the ISB annotation had no equivalent in our TE models. We investigated these differences to verify how the two methods annotated loci and then to determine the quality of the ISB annotations that had no annotation by our procedure (see Additional file 14). The main conclusion was that the annotations calculated with library-based methods depend heavily on the

quality of the library used. A library that is not composed of tailored consensuses tends to force and fragment annotations.

Re-discovering the distribution profiles of TEs in galGal4 chromosomes.

TE distributions among functional elements in galGal4. The current view of the chromosome organization with respect to TEs [28,32] is that macrochromosomes display gene and TE densities that are respectively lower and higher than those of the microchromosomes. In an attempt to verify these features we investigate the depletions or the over-representations of TEs, genes, S/MAR elements and CpG islands in macrochromosomes and microchromosomes using permutation tests (see Methods). The analyses were conducted in terms of numbers of copies (Figure 7) or coverage (Additional file 15) and found that they gave overall similar results. We then used these 4 DNA elements together with chromosome size to show that there were not two, but at least three types types of chromosomes that have at least four features. The first group was chromosomes 1, 2, 3 ,4, and Z, the largest chromosomes, with more TEs and S/MARs (Figure 7a and c) and fewer genes and CpG islands than expected by chance (Figure 7b and d). The second group included chromosomes 5 ,6 ,7 ,8 and 9, with fewer TEs and CpG islands and more genes than expected (Figure 7a, b and c), but with a number of S/MARs that varied significantly from one chromosomes to another (Figure 7d). The third group contained all the smallest chromosomes, poorer in TEs and S/MARs (Figure 7a and c), and richer in genes and CpG islands than expected (Figure 7b and d). Chromosomes W and LGE64 were two notable exceptions that did not fit into these 3 chromosome types. They had features of both macrochromosomes, rich in TEs and CpG islands, and microchromosomes, chromosome size, number of genes, and SMARs (Figure 7, Additional file 15).

These features of the RJF chromosome organization were then used to investigate the distribution of TEs with reference to genes (Figure 8a,b,c), CpG islands and S/MARs (Figure 8d). We checked the TE distribution between exons, other genes and intergene regions in the regions of

galGal4 chromosomes using TE annotations resulting from STEP3 (the best-conserved TE copies; Figure 8b) and final [TE+DM] annotations (Figure 8b). TEs tended to be more abundant in intergene regions however the annotations were compared (per TE model or all models together (bars labelled ALL in Figure 10 A and B)). The exception was the repeats of the undetermined_group_1, which were abundant in exon regions once the most divergent copies (DM) were included in the calculation. Both our annotations and that of the ISB found that the abundance of TE copies in exons were similar. The [TE+DM] annotations (3.6% and 1.7% in coverage) showed that there were more TE copies (Figure 8c) in exons than in the ISB (2.1%) or the TE alone (2.3%) and 1.1% for coverage in both. This suggestsd that the rate at which ancient and more recent TE copies became recently exonized is similar to those reported for mammal genomes [89-91].

There were 21,663 CpG islands (average size: 645 bp) and 53,115 S/MAR (444 bp) in galGal4. The abundance of TEs in two kinds of elements and their 3 kbp proximal and distal regions (Figure 8d) is very like that in the rest of the genome. This is very different from the situation in mammal genomes, where regions containing S/MAR are enriched in TEs [92] and CpG islands in SINEs [93,94].

We conclude that TEs are more abundant in intergene regions of the RJF genome and are not more concentrated in CpG islands and S/MAR than in the rest of the genome. We determined the densities of all TEs and each TE species among chromosomes was investigated because the data in Figure 4 indicate that the distribution patterns of some TE species in chromosomes 16, 32 and W are quite specific.

TE distributions between and within galGal4 chromosomes. A survey of global TE density (Figure 9a) indicated that chromosomes 1, 2, 3, 4, 16, LGE64, Z and W contained more TE copies than the other chromosomes. The profiles of TE species seem to be strikingly different from one species to

another. We first found that the global density in CR1 (Figure 9b) was similar to the global TE density, except in chromosome W. The picture was very similar for each of the 8 CR1 sub-families (Additional files 16). The densities of CR1-C, CR1_F, CR1-G were greater in chromosomes 16 and W than were those of CR1-D, CR1_GG, CR1-Y and CR1_like, which were close to those of chromosomes 5 to 25. The density of CR1-H was elevated only in chromosome W.

We identified six other TE density profiles (Figure 9 and Additional file 17). The first profile contains CR1s and one other element, Hitchcock (Additional file 17B1). The TE species in the second and third profiles are found in most chromosomes; they may be super-abundant (Figure 9f, Additional file 17X to A1; Charlie, Charlie-Galluhop, Galluhop and Mariner, all of them are TIR elements), or less abundant (undetermined_group_1) relative to chromosome size. The fourth density profile included twenty LTRs and five undetermined_group_2 to 6 species; the putative_LTR_group9 is the exception. These were present in many chromosomes at low density, but were abundant in chromosome W and one or more other chromosomes 16, LGE22 and LGE94 (Figure 9c, d, Additional file 17C to N, P to W, and D1 to H1). The fifth and sixth density profiles each contained just one element, the putative_LTR_G9 (Figure 9e) is only present in half the chromosomes and Z rep elements are mostly concentrated on the Z and W chromosomes (Additional file 17I1).

We looked for TE hot spots using permutations tests (Figure 10 and Additional files 16 and 17). Global analysis of all TE models showed that the chromosomes richest in TEs (1, 2, 3, 4, 16, LGE64, Z and W; Figure 10a) are those that also contain many TE hot spots. The global profile of hot spots for CR1 elements, like the density profiles, is very similar to that of all models (Figure 10b), except for chromosome W. However, the hot spot profiles for the eight CR1 sub-families were different (Additional file 16). We found that five LTR species had *a priori* no hot spots in galGal4 (Additional file 17; ancestral_LTR_group4, EAV, putative_LTR_group9, putative_LTR_group28 and undetermined_group_5). This suggests that their distribution is driven only by certain chromosomal features, not by specific regions. The hot spots of other LTR species are generally on

chromosome W (Figure 10b,d, Additional file 17C to W), except for the putative_LTR_group4 (Figure 10c), whose hot spots were only on chromosomes 1 and 3. The hot spot profiles of the remaining TE species (TIR and undetermined) were all concentrated on the largest chromosomes (Figure 10e,f), but could be very different from one to another (Additional file 17B1 to I1).

We find that the overall distribution of TEs along the galGal4 chromosomes and their concentrations in certain chromosomal regions depends on each of the 34 TEs species. Most LTR elements were found on chromosome W. But the distributions of each TE species do not seem to reflect any preferences for insertion in the RJF genome. In fact, most RJF TEs are ancient elements that contain significant numbers of point mutations and arethus probably inactive. Therefore, their distributions are due to the insertion preference of each TE species and the ability of the RJF genome to eliminate or conserve them during evolution, depending on the region where each TE is inserted. We cannot examine this topic any further using the chicken recombination maps because these data are not available for the RJF. Calculations from domestic breeds cannot be directly used for the RJF genome since they differ from one breed to another [95], and the extent to which the sizes of the genomes and non-gene regions in between RJF and domesticated lines differ has not yet been evaluated. There is a strong correlation between GC richness and chromosome recombination rates [96], but we find no such correlation between the GC-content and local TE densities in chromosomes. The forces driving the density and hot spot profiles of each of the 34 TE models in galGal4 are therefore due to something other than ectopic recombination.

Conclusions

Our study has succeeded in its two main objectives. First, we have developed a general strategy for annotating (including quality assessment) repeated sequences in a model of an avian genome. Second, we have used this strategy to annotate the repeated sequences in galGal4 using repeat models that can readily be used directly to annotate the RJF genome.

Ins and outs of our approach to annotate repeats in eukaryotic genomes.

Before investing manpower and resources in such a task, it must be first calibrated using existing information on the sizes of real and model genomes, and estimates of repeat amounts. The size of the real genome was estimated from data on several species in various databases [38,97,98]. We used a k-mer method to calculate the genome size where data were not available, as was done recently with Cephalopoda species [99]. The reliability of this new approach needs to be tested on avian and mammal models once the programs are available. Reassociation kinetics data are particularly valuable because tools such a P-clouds and Red are unreliable for estimating repeats in galGal4. While implementation of our annotation strategy required a significant investment in time and computer resources, it enabled us to annotate the repeats in galGal4 more reliably than using RM and is available in the ISB. Our annotation strategy has shown that there are more repeats (~18.8%, rather than ~11.5% in the ISB annotation) and less TE diversity (34 rather than over 200 in the ISB annotation) in galGal4 than previously reported [28,60,61].

Our results confirm that *de novo* approaches for annotating repeats are more efficient than library-based method and are less likely to produce artefactual annotations. We found that at least some of the 8.5% ISB annotations (0.76% of the 8.87% of TE annotation in coverage) are probably artefacts. This is a fault shared by all library-based methods, which tends to force the search for sequence matches that vary greatlyin size to consensuses present in the reference sequence library. This fault can be amplified when many heterospecific TE sequences are available and the reference library contains no specific repeated sequences. Nevertheless, the published data and Repbase were useful. The many CR1 non-LTR retrotransposons in galGal4, their 5' truncation profiles, ages and their ability to insert themselves into each other,meant that REPET produced many consensuses (308/499) that would have been difficult to manage without any idea of their putative organisation in sub-families. We therefore suggest that anyone wanting to use a similar annotation strategy on

other models dhoulf perform preliminary analyses of non-LTR retrotransposons (LINEs and SINEs) before implementing the current version of the REPET pipeline.

The database of galGal5 was released in January 2016 [100], just as we were preparing the final version of our manuscript. This new version contains 1.232 Gbp, close to the C-value (1.223 ± 0.058 Gbp) and has fewer ambiguities (only 0.95% "N" in its sequence, compared to 2.39% in galGal4). Its greater size is due to the discovery of about 6400 new genes (18644 in galGal4, 25062 in galGal5). Repeat annotation with RM revealed that galGal5 has 6.98% satellite SSRs and 9.06% TEs, for a totalof 16.04% repeated sequences. These repeated sequences were annotated using the approach and models described above. Results and gff files are available at http://chicken-repeats.inra.fr/. They indicate that are 10.50 % SSRs and 10.86% TEs; our annotation gives the total amount of these repeated sequences as 21.36%. We verified the distributions of TEs in the inter-gene and intra-gene regions and found results similar to those presented here in Figure 8a and c (results are available at http://chicken-repeats.inra.fr/).

New insights provided by a deeper repeat annotation.

In addition to the number of repeats and TE diversity, our annotation update modifies the landscape of repeats in the RJF genome. First, the sum of the 4 - 8% of repeated centromere and telomere sequences to the 26.7 % of repeats found in galGal4 (SSRs + TEs + CNVs) is not far off that of the real RJF genome (31 - 35% repeated sequences, half of them TE sequences). Although the RJF genome contains fewer repeats than most mammal genomes, this repeat content is nearly a 3-times greater than previous estimates, similar to the repeats in Chiroptera (bat) genomes [101]. The distributions of repeats in avian chromosomes differ from those in other vertebrate genomes in at least two ways. First, there are many different, small families of satellite DNAs interspersed along chromosome arms in addition to repeats in megacentromers and megatelomers,, and these satellite DNAs are more abundant in small chromosomes. This distribution of satellite DNAs might

actually label each small chromosome with a sort of satellite DNA code. These labels might even be involved in chromosome recognition and influence the physical separation of small and large chromosomes that occurs during cell division in birds [102]. Second, none of the 34 TE species found in galGal4 are randomly distributed along chromosomes. Most of them are arranged in specific patterns that suggest that they were not randomly inserted into chromosomes during evolution, and conversely were not randomly eliminated from chromosomes.

This brings us to the idea that TEs are inactive in present-day birds. Recent data indicate that few TE species are active in mammals and insects and that some are involved in development and differentiation pathways [103-105]. It would therefore appear that inactive TEs are an avian characteristic, as these pathways are also present in Sauropsida species. Our annotation indicates that there are at least three active TEs in the chicken genome. The first is EAV-HP, an LTR element that was shown recently to have been active in the chicken [106,107]. The other two are in elements that were until recently considered to be neogenes coding for transposases, THAP9 and PGBD5 (Additional file 11). These two genes are present and active in every vertebrate species and were recently shown to transpose, in *trans,* non-autonomous related TIRs in the human genome [108,109].

Thus the importance of TEs in avian genomes is far from completely elucidated; the most abundant TE species may well not be the most interesting candidates for studying genome rearrangements during development.

Methods

Genome model

galGal4 was downloaded from the UCSC website (http://hgdownload.cse.ucsc.edu/downloads.html). galGal5 was downloaded from the NCBI website [100]. The file describing the annotation of CpG islands in galGal4 was also downloaded from the UCSC website. The annotation file describing the S/MAR sequences is available from Genomatix (https://www.genomatix.de/). All studies were done on both the assembled and unassembled genomes. Because our materials were only in silico data supplied by the UCSC, the NCBI and Genomatix, no ethical statement was required to achieve our works.

P-Clouds

Version 0.9 was download from the web sitehttp://www.evolutionarygenomics.com/ProgramsData/Pclouds/Pclouds.html. P-clouds does not manage correctly the 'N' residues in the sequence of genome models; it considers them to be stretches of 'A' nucleotides. The 14 Mbp of 'N' in galGal4 means that this creates a huge number of k-mer derivates from A-stretches that are false annotations. We overcame this problem by developing a wrapper for P-clouds that pretains the main program but replaces the original pre-processors and post-processors. The wrapper is a Perl script called 4pclouds.pl (P-clouds pre-post-processor) that creates an index to manage the removal of the 'Ns', then restores the scaling of the chromosomes of the model after P-clouds treatment.

P-clouds requires a set of five cutoff parameters to be launched in addition to the genome sequence to be analyzed. Parameter 1, the lower cutoff, is the minimum number of repeats of the oligo in a genome to be integrated in a P-clouds. Parameter 2, the core cutoff, is the minimum number of repeats of the oligo in a genome to be used as a seed for P-clouds. Parameters 3, 4 and 5 are the primary, secondary and tertiary cutoffs that define the smallest number of repeats required

for a core oligo to integrate to the outer layer of oligos presenting one, two or three nucleotide mismatches with it. The optimal parameters are defined by six sets of parameters c4(2, 4, 8, 80, 800), c5(2, 5, 10, 100, 1000), c8(2, 8, 16, 160, 1600), c10(2, 10, 20, 200,2000), c100(10, 100, 200, 2000, 20000), c200(20, 200, 400, 4000, 40000). Each parameter set uses a 16-nucleotide oligonucleotide (k-mer) that was calculated using the formula $l=\log_4 N + 1$, where l is the oligo size and N the genome size [63]. The final output of a P-clouds calculation is a bed file.

Red

The code of Red (in C++) and complementary information were downloaded from html.http://toolsmith.ens.utulsa.edu. Launching the compiled Red provides the genome sequence to be analyzed and an oligo (k-mer) size that is calculated using the same formula as for P-clouds (16 nucleotides). The final output of a P-clouds calculation is a bed file.

Analyses of SRRs

The TRF version 4.07b was downloaded from the tandem repeat finder website (http://tandem.bu.edu/trf/trf.download.html). The Match, Mismatch, Delta, PM, PI, Minscore, MaxPeriod parameters were set at 2, 5, 7, 80, 10, 25, and 2000. The -m option was used to obtain a masked genome and the -d option to obtain the data file output. Data file outputs were analysed using a homemade Perl script to determine the type of repeat of each annotation (simple repeat, microsatellite, minisatellite, large tandem repeats (including satellite DNA). Each annotation was then loaded into a MySQL database from which was produced a GFF file describing the features of all SSRs, each with the attribute (ninth column) containing an ID, the type of SSR, the size of the repeat unit, the repeat unit sequence, the tandem array size and the number of copies of the repeat unit. A second homemade Perl script was used to select simple sequences, microsatellites and minisatellites based on an arbitrary minimum size of 50 tandem arrays. Arrays with units over 60-bp

composed of at least 2 repeats were selected and ranked in large tandem repeats when the repeated unit was tandemly repeated fewer than 50 times and in satellite DNAs when they were repeated over this threshold.

Annotations of Dispersed repeats with REPET

Dispersed repeats were annotated in three steps using the REPET package version 2.2 (available at https://urgi.versailles.inra.fr/Tools/REPET). For the first run (REPET 1; Figure 2), the SSRs and a macro-satellite present only in the Z chromosome were removed in galGal4 and TEdenovo was used with its default parameters. TEdenovo is a pipeline that combines several programs to optimize the production of an exhaustive list of consensuses. TEdenovo was run with galGal4 after discarding three programs (Additional file 4). PILER, because it could not manage the amount of data produced during the analysis of models like galGal4. LTR_HARVEST because it produced too many false-positive consensuses. LTR_HARVEST identifies a sequence as an LTR retrotransposon as soon as it can locate two large direct repeats close enough to gather them into a pair of LTR flanking a retro-transposed DNA segment. Thus, LTR_HARVEST identified many purely artifactual LTR retrotransposons in galGal4, where copies of non-LTR retrotransposons like CR1 or the DNA transposons like Galluhop are abundant, whatever the parameter set used. We also removed BLASTclust, which intervenes at the end of the TEannot procedure because it produced aberrant clusters of consensuses under our conditions.

The output of TEdenovo, Library 1 (Figure 2), was used to produce a first annotation of galGal4 using TEannot with its default parameters. Consensuses in Library 1 were then filtered to produce the Library 1f using two programs of the REPET package. PostAnalyzeTELib.py produced statistical descriptions of each consensus used to extract the full length fragment consensuses (consensus with at least one full length copy in the genome) using GetSpecificTELibAccordingToAnnotation.py. Library 1f was then used to annotate galGal4 using

TEannot with its default parameters. The resulting annotated genome copies were then used to calculate a reduced version of galGal4. The second run (REPET 2; Figure 2) was designed to detect other repeats fragmented by nested insertion of repeats identified by REPET1. The REPET 2 run was managed by filtration similar to that used in REPET 1 to produce Library 2f. The third run (REPET 3; Figure 2) merged libraries 1f and 2f, which was filtered with TEannot to produce Library 3f. The name and classification supplied by PASTEC (Additional file 4) for each consensus in TEannot were verified and changed manually because we found 15 - 20 % errors, depending on the TE model. Library 3f was used to edit the final annotation of galGal4.

DM annotation

The TEs (>500 bp) with at least 80 % sequence similarity to their consensuses identified during the REPET procedure were extracted with GFFtools and used to detect and annotate DM. We then used TEannot with its default parameters and these TEs to mine galGal4 to locate more divergent TE segments corresponding to the DM. The resulting DM was subtracted from the annotation file with bedtools so as to remove all repeats identified in STEPs 2 and 3 of the complete annotation procedure (Figure 2).

Analysis of annotation features

Unique or intersecting annotations were computed using bedtools. The shared annotations were obtained with intersectBed and the intervals were removed using subtractBed. The coverage was computed by summing the lengths of intervals and dividing by the genome size. The results were transferred directly to R.

We developed GFFtools (Additional file 16) to analyse the TE distribution and their coverage in galGal4 chromosomes. Existing libraries like Bio::Tools::GFF in Bio::Perl can parse and analyse GFF files but none of them can readily manage the attributes column (ninth column) of a GFF file

and perform operations on features such as reducing intervals. GFFtools has two Perl objects that can store the whole GFF file in a data structure, parse features, add annotations, filter features, reduce overlapping features, and deal with overlaps. GFF files were finalized with GFFtools in order to reduce the number of overlapping features, selecting those most similar and identical between annotations, then those with the highest percentage of coverage with their annotating consensuses.

TE densities were analysed by counting the number of TE copies in each chromosome, except for long-join annotations. This was done for all models and then for each model. The REPET long-join analysis involved merging two annotations related to the same TE model and then splitting them into two or more annotations, depending on the presence of one or more inserted TEs.

Permutation tests for analysing the distribution of a repeat DNA element

We used a home-made Perl script to determine the size of chromosomes minus the coverage of a single kind of DNA element (TE, gene, S/MAR, CpG island) and thus obtain the size of the reduced genome. We next calculated the random distribution of the number of elements in the reduced genome and the number of copies of the element in each chromosome. We then calculated 100000 permutations per chromosome and fed these data into R to draw a histogram of the number of elements. This gave us the two thresholds at which there was a 1 % chance of getting a TE-rich or TE-poor distribution in each chromosome (Additional file 18).

Permutation tests for analysing the presence of TE hot spots

We used a permutation test for each kind of TE guild analysed (aTE model or a group of TE models) to determine a threshold above which a chromosome region was considered to be a TE hot spot. We first calculated, using a 50 kbp window, 1000 permutations of randomized TE distributions, and then used these distributions to determine the 1% threshold above which a 50 kbp

region in each chromosome could be a hot spot. The window size was used to take into account the coverages of TEs and Ns and so avoid overlap due to TE content and N stretches (Additional file 18).

Availability of home-made gff files and software

All home-made gff files and software (GFFtools and DensityMap.pl) are available at http://chicken-repeats.inra.fr/.

Abbreviations used

Acc N°, accession number

AMAV, Avian myelocytomatosis virus

CNV, copy number variation

DM, dark matter

Env, retroviral envelope protein

Gag, group specific antigen

ISB, institute for systems biology

IS, insertion sequence

LINE, long interspersed element

LTR, long terminal repeats

MGE, mobile genetic element

ORF, open reading frame

ppt, polypurine tract

RJF, red jungle fowl

RM, RepeatMasker

rRNA, ribosomal RNA

RSV, Rous sarcoma virus

RT, reverse transcriptase

SINE, short interspersed element

S/MAR, scaffold/matrix attachment region

SSR, simple sequence repeats

TIR, terminal inverted repeats

TRF, tandem repeat finder

TE, transposable element

VNTR, variable number tandem repeats

Competing interest

The authors declare no competing interest.

Authors' contributions

SG, BP and YB designed the research program; SG, BP and YB performed the analyses; SG, PA, FG and YB wrote the manuscript, figures, additional files and captions.

Description of additional data files

Additional file 1: Conditions of use for programs P-clouds and Red.

Additional file 2: Evaluating the efficiency of P-clouds and Red. This file describes data on the efficiencies of P-clouds, Red and REPET for evaluating the numbers of repeats in a eukaryotic genome.

Additional file 3: Features of SSRs in galGal4

Additional File 4: Diagram showing programs in both REPET TEdenovo and TEannot components.

Additional file 5: TE coverage in each galGal4 chromosomes in the ISB, REPET TE, TEannot DM and TE+DM annotations

Additional File 6: Intersections between annotation files calculated with P-clouds, TRF, and REPET.

Additional file 7: Correspondence between the names of consensus describing TEs in Repbase and ISB, and the TE models calculated with REPET

Additional File 8: Diversity of CR1 within galGal4

Additional File 9: Nucleic acid sequences of retroCalimero, retroSaturnin and retroTux.

Additional file 10: Features of 8 repeat models that cannot be assigned to a known eukaryotic TE

Additional File 11: Proteins encoded by the remnant polinton in the RFJ and turkey genomes.

Additional File 12: Characteristics of the 54 neogenes derived from DNA transposons in the human and RJF genomes.

Additional File 13: The number of RM annotations that had no equivalent in the [TE+DM] annotation.

Additional file 14: Origins of differences between ISB and [TE+DM] annotations.

Additional File 15: Graph showing the expected and observed coverages of TEs (A), genes (B), S/MAR (C) and CpG islands (D) in galGal 4 chromosomes.

Additional File 16: Histograms showing the densities of TEs and TE hot spots in galGal4 chromosomes for the 8 sub-families of CR1 elements.

Additional File 17: Histograms showing the densities of TEs (left column) and TE hot spots in galGal4 chromosomes for all TEs plus each of the 34 TE models.

Additional File 18: Graph showing thresholds calculated in permutation assays and windows calculated along chromosomes for permutation tests designed to inventory hot spots.

Authors' information

[1] Physiologie de la reproduction et des Comportement, UMR INRA-CNRS 7247, PRC, 37380 Nouzilly - France; [2] Biological Sciences Department, California State Polytechnic University, Pomona, CA 91768, United States of America; [3] Unité de Recherches en Génomique-Info, UR INRA 1164, Centre de recherche de Versailles, bat.18, RD10, Route de Saint Cyr, 78026 Versailles, France

Competing interest

None of the authors have any competing interests.

Acknowledgements

We thank Véronique Jamilloux, Timothée Chaumier, Isabelle Luyten, Joëlle Amselem, Olivier Inizan and Mark Moissette (URGI, INRA Center of Versailles, France) for training and support in the use of REPET, members of Genotoul (INRA Center of Castanet-Tolosan, France) for access to computing facilities, and Olivier Panaud (University of Perpignan, France) for access to complementary computing facilities. We thank all the speakers at the workshop "Analysis and Annotation of DNA Repeats and Dark Matter in Eukaryotic Genomes" (Tours, July 8 to 10, 2015) for fruitful discussions and suggestions in genomics and bioinformatics: Dr Davide Gabellini, Dr Jiri Macas, Dr Florian Maumus, Dr Jan Øivind Moskaug, Dr Attila Nemeth, Dr Bruno Pitard, Dr David Pollocq, Dr Sébastien Tempel, and Dr Jean Nicolas Volf. We also thank Dr Alain Vignal and Dr Frédérique Pitel (INRA Center of Castanet-Tolosan, France) and Dr Chrisrine Leterrier (INRA Center of Nouzilly, France) for kindly sharing their knowledge and experience in avian biology, genetics and genomics and for many fruitful discussions. This work was funded by the Région Centre Val de Loire (AviGeS Project), the C.N.R.S., the I.N.R.A., the Groupements de Recherche CNRS 3546 (Elements Génétiques Mobiles) and 3604 (Modèles Aviaires), and the Ministère de l'Education Nationale, de la Recherche et de la Technologie. Sébastien Guizard holds a doctoral fellowship jointly funded by I.N.R.A. (PHASE department)/Région Centre Val de Loire and a training grant for the Ecole doctorale "Santé, Sciences Biologiques et Chimie du Vivant" of the Universitary PRES Centre Val de Loire. Peter Arensburger holds a senior researcher fellowship from STUDIUM.

References

1. Schachat FH, Hogness DS. Repetitive sequences in isolated Thomas circles from Drosophila melanogaster. Cold Spring Harb Symp Quant Biol. 1974;38:371-81.

2. Manning JE, Schmid CW, Davidson N. Interspersion of repetitive and nonrepetitive DNA sequences in the Drosophila melanogaster genome. Cell. 1975;4:141-155.

3. De Koning APJ, Gu W, Castoe TA, Batzer MA, Pollock DD. Repetitive elements may comprise over two-thirds of the human genome. PLoS Genet. 2011;7:e1002384.

4. San Miguel P, Tikhonov A, Jin YK, Motchoulskaia N, Zakharov D, et al. Nested retrotransposons in the intergenic regions of the maize genome. Science. 1996;274:765-768.

5. O'Hare TH, Delany ME. Genetic variation exists for telomeric array organization within and among the genomes of normal, immortalized, and transformed chicken systems. Chromosome Res. 2009;17:947-964.

6. Beridze T. Satellite DNA.In Beridze editor.Berlin, Heidelberg, New york, London : Springer Verlag; 1986.

7. Pezer Z, Brajković J, Feliciello I, Ugarkovć D. Satellite DNA-mediated effects on genome regulation. Genome Dyn. 2012;7:153-169.

8. Primmer CR, Raudsepp T, Chowdhary BP, Møller AP, Ellegren H. Low frequency of microsatellites in the avian genome. Genome Res. 1997;7:471-482.

9. Brandström M, Ellegren H. Genome-wide analysis of microsatellite polymorphism in chicken circumventing the ascertainment bias. Genome Res. 2008;18:881-887.

10. Völker M, Backström N, Skinner BM, Langley EJ, Bunzey SK, et al. Copy number variation, chromosome rearrangement, and their association with recombination during avian evolution. Genome Res. 2010;20:503-511.

11. Yi G, Qu L, Liu J, Yan Y, Xu G, Yang N. Genome-wide patterns of copy number variation in the diversified chicken genomes using next-generation sequencing. BMC Genomics. 2014;15:962.

12. Piégu B, Bire S, Arensburger P, Bigot Y. 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. 2015;86:90-109.

13. Hoen DR, Hickey G, Bourque G, Casacuberta J, Cordaux R, et al. A call for benchmarking transposable element annotation methods. Mob DNA. 2015;6:13.

14. Lerat E. Identifying repeats and transposable elements in sequenced genomes: how to find your way through the dense forest of programs. Heredity. 2010;104:520-533.

15. Girgis HZ. Red: an intelligent, rapid, accurate tool for detecting repeats de-novo on the genomic scale. BMC Bioinformatics. 2015;16:227.

16. Tempel S. Using and understanding RepeatMasker. Methods Mol Biol. 2012;859:29-51.

17. RepeatMasker Open-4.0. http://www.repeatmasker.org (2014). Accessed 10 Sep 2015.

18. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, et al. Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res.2005;110:462-467.

19. Institute for Systems Biology: RepeatMasker Genomic Datasets. http://www.repeatmasker.org/genomicDatasets/RMGenomicDatasets.html (2014) Accessed 10 Sep 2015.

20. Xu Z, Wang H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res. 2007;35:W265-268.

21. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 1999, 27:573–80.

22. Quesneville H, Bergman CM, Andrieu O, Autard D, Nouaud D, et al. Combined evidence annotation of transposable elements in genome sequences. PLoS Comput Biol. 2005;1:166-175.

23. Flutre T1, Duprat E, Feuillet C, Quesneville H. Considering transposable element diversification in de novo annotation approaches. PLoS One. 2011;6:e16526.

24. Permal E, Flutre T, Quesneville H. Roadmap for annotating transposable elements in eukaryote genomes. Methods Mol Biol. 2012;859:53-68.

25. Maumus F, Quesneville H. Deep investigation of Arabidopsis thaliana junk DNA reveals a continuum between repetitive elements and genomic dark matter. PLoS One 2014, 9:e94101.

26. Bed'Home B, Coullin P, Guillier-Gencik S, Moulin S, et al. Characterization of the atypical karyotype of the black-winged kite Elanus caeruleus (Falconiformes: Accipitridae) by means of classical and molecular cytogenetic techniques. Chromosome Res. 2003;11:335-343.

27. Habermann FA, Cremer M, Walter J, Kreth G, von Hase J, et al. Arrangements of macro- and microchromosomes in chicken cells. Chromosome Res. 2001;9:569-584.

28. International Chicken Genome Sequencing Consortium. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature 2004;432:695-716.

29. Kelley DR, Salzberg SL. Detection and correction of false segmental duplications caused by genome mis-assembly. Genome Biol. 2010;11:R28.

30. Ye L, Hillier LW, Minx P, Thane N, Locke DP, Martin JC, et al. A vertebrate case study of the quality of assemblies derived from next-generation sequences. Genome Biol. 2011;12:R31.

31. Zhang Q, Backström N. Assembly errors cause false tandem duplicate regions in the chicken (Gallus gallus) genome sequence. Chromosoma. 2014;123:165-168.

32. M, Smith J, Burt DW, Aken BL, Antin PB, et al. Third Report on Chicken Genes and C Schmid hromosomes 2015; Cytogenet Genome Res. 2015;145:78-179.

33. Eden FC, Hendrick JP, Gottlieb SS. Homology of single copy and repeated sequences in chicken, duck, Japanese quail, and ostrich DNA. Biochemistry 1978;17:5113-5121.

34. Olofsson B, Bernardi G. Organization of nucleotide sequences in the chicken genome. Eur J Biochem. 1983;130:241-245.

35. Tiersch TR, Wachtel SS. On the evolution of genome size of birds. J Hered. 1991;82:363-368.

36. Organ CL, Shedlock AM, Meade A, Pagel M, Edwards SV. Origin of avian genome size and structure in non-avian dinosaurs. Nature. 2007;446:180-184.

37. Mendonça MA, Carvalho CR, Clarindo WR. DNA content differences between male and female chicken (Gallus gallus domesticus) nuclei and Z and W chromosomes resolved by image cytometry. J Histochem Cytochem. 2010;58:229-235.

38. Gregory TR. Animal Genome Size Database. (2015) http://www.genomesize.com. Accessed 10 Sep 2015.

39. Doležel J, Bartoš J,Voglmayr H, Greilhuber J. Letter to the editor: Nuclear DNA Content and Genome Size of Trout and Human. Cytometry 2003;51A:127–128.

40. Shang WH, Hori T, Toyoda A, Kato J, Popendorf K, et al. Chickens possess centromeres with both extended tandem repeats and short non-tandem-repetitive sequences. Genome Res. 2010;20:1219-1228.

41. Nanda I, Schmid M. Localization of the telomeric (TTAGGG)n sequence in chicken (Gallus domesticus) chromosomes. Cytogenet Cell Genet. 1994;65:190-193.

42. Delany ME, Krupkin AB, Miller MM. Organization of telomere sequences in birds: evidence for arrays of extreme length and for in vivo shortening. Cytogenet Cell Genet. 2000; 90:139-145.

43. Delany ME, Gessaro TM, Rodrigue KL, Daniels LM. Chromosomal mapping of chicken mega-telomere arrays to GGA9, 16, 28 and W using a cytogenomic approach. Cytogenet Genome Res. 2007;117:54-63.

44. Maslova A, Zlotina A, Kosyakova N, Sidorova M, Krasikova A. Three-dimensional architecture of tandem repeats in chicken interphase nucleus. Chromosome Res. 2015; 23:625-639.

45. Su MH, Delany ME: Ribosomal RNA gene copy number and nucleolar-size polymorphisms within and among chicken lines selected for enhanced growth. Poult Sci. 1998;77:1748-1754.

46. Aird D, Ross MG, Chen WS, Danielsson M, Fennell T, et al. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome Biol. 2011;12:R18.

47. Krueger F, Andrews SR, Osborne CS. Large scale loss of data in low-diversity illumina sequencing libraries can be recovered by deferred cluster calling. PLoS One. 2011;6:e16607.

48. Nakamura K, Oshima T, Morimoto T, Ikeda S, Yoshikawa H, et al. Sequence-specific error profile of Illumina sequencers. Nucleic Acids Res. 2011;39:e90.

49. Benjamini Y, Speed TP. Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic Acids Res. 2012;40:e72.

50. Dabney J, Meyer M. Length and GC-biases during sequencing library amplification: a comparison of various polymerase-buffer systems with ancient and modern DNA sequencing libraries. Biotechniques. 2012;52:87-94.

51. Oyola SO, Otto TD, Gu Y, Maslen G, Manske M, et al. Optimizing Illumina next-generation sequencing library preparation for extremely AT-biased genomes. BMC Genomics. 2012;13:1.

52. van Heesch S, Mokry M, Boskova V, Junker W, Mehon R, et al. Systematic biases in DNA copy number originate from isolation procedures. Genome Biol. 2013 Apr 24;14(4):R33.

53. Miller MM, Robinson CM, Abernathy J, Goto RM, Hamilton MK, et al. Mapping genes to chicken microchromosome 16 and discovery of olfactory and scavenger receptor genes near the major histocompatibility complex. J Hered. 2014;105:203-215.

54. Newcomer EH. Accessory chromosomes in the domestic fowl. Genetics. 1955;40:587-588.

55. Friedman-Einat M, Cogburn LA, Yosefi S, Hen G, Shinder D, et al. Discovery and characterization of the first genuine avian leptin gene in the rock dove (Columba livia). Endocrinology. 2014;155:3376-3384.

56. Lovell PV, Wirthlin M, Wilhelm L, Minx P, Lazar NH, et al. Conserved syntenic clusters of protein coding genes are missing in birds. Genome Biol. 2014;15:565.

57. Hron T, Pajer P, Pačes J, Bartůněk P, Elleder D. Hidden genes in birds. Genome Biol. 2015;16:164.

58. Arthur RR, Straus NA. DNA-sequence organization in the genome of the domestic chicken (Gallus domesticus). Can J Biochem 1978, 56:257–63.

59. Epplen JT, Leipoldt M, Engel W, Schmidtke J. DNA sequence organisation in avian genomes. Chromosoma 1978, 69:307–321.

60. Wicker T, Robertson JS, Schulze SR, Feltus FA, Magrini V, et al. The repetitive landscape of the chicken genome. Genome Res 2005, 15:126-136.

61. Institute for Systems Biology: Chicken genomic dataset. http://www.repeatmasker.org/species/galGal.html (2014). Accessed 10 Sep 2015.

62. Bigot Y, Hamelin MH, Periquet G. Molecular analysis of the genomic organization of Hymenoptera Diadromus pulchellus and Eupelmus vuilleti. J Evol Biol;4:541-556.

63. Gu W, Castoe T, Hedges DJ, Batzer MA, Pollock DD. Identification of repeat structure in large genomes using repeat probability clouds. Anal Biochem 2008, 380:77–83.

64. De Koning APJ, Gu W, Castoe TA, Batzer MA, Pollock DD. Repetitive elements may comprise over two-thirds of the human genome. PLoS Genet 2011, 7:e1002384.

65. Price AL, Jones NC, Pevzner PA. De novo identification of repeat families in large genomes. Bioinformatics 2005;21:351–358.

66. Smit AFA,Hubley R. RepeatModeler 1.0.8 website. http://www.repeatmasker.org/RepeatModeler.html (2008) Accessed 2015 Sep 14.

67. Buisine N, Quesneville H, Colot V. Improved detection and annotation of transposable elements in sequenced genomes using multiple reference sequence sets. Genomics 2008;91:467-475.

68. Benchmark_Proposal_URGI. http://cgl.cs.mcgill.ca/wp-content/uploads/2014/06/Benchmark_Proposal_URGI_version.docx (2014) Accessed 2015 Sep 14.

69. Yi G, Qu L, Liu J, Yan Y, Xu G, Yang N. Genome-wide patterns of copy number variation in the diversified chicken genomes using next-generation sequencing. BMC Genomics. 2014;15:962.

70. Hori T, Suzuki Y, Solovei I, Saitoh Y, Hutchison N,et al. Characterization of DNA sequences constituting the terminal heterochromatin of the chicken Z chromosome. Chromosom Res 1996;4:411-426.

71. Yamada K: Empirical Analysis of Transcriptional Activity in the Arabidopsis Genome. Science 2003, 302:842-846.

72. Trayhurn P: Of genes and genomes – and dark matter. Br J Nutr 2004, 91:1.

73. Ponting CP, Grant Belgard T: Transcribed dark matter: Meaning or myth? Hum Mol Genet 2010, 19:162-168.

74. Melhem N, Devlin B: Shedding new light on genetic dark matter. Genome Med 2010, 2:79.

75. Pennisi E: Shining a light on the genome's "dark matter". Science 2010, 330:1614.

76. Jenks S: Navigating the genome's "dark matter". J Natl Cancer Inst 2013, 105:673–674.

77. Jiang J: The "dark matter" in the plant genomes: non-coding and unannotated DNA sequences associated with open chromatin. Curr Opin Plant Biol 2015, 24:17–23.

78. Brosius J: Genomes were forged by massive bombardments with retroelements and retrosequences. Genetica 1999, 107:209–238.

79. Wheeler TJ, Eddy SR. nhmmer: DNA homology search with profile HMMs. Bioinformatics. 2013;29:2487-2489.

80. Wheeler TJ, Clements J, Eddy SR, Hubley R, Jones TA, et al. Dfam: a database of repetitive DNA based on profile hidden Markov models. Nucleic Acids Res. 2013;41:D70-82.

81. Novák P, Neumann P, Macas J. Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. BMC Bioinformatics. 2010;11:378.

82. Novák P, Neumann P, Pech J, Steinhaisl J, Macas J. RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. Bioinformatics. 2013;29:792-793.

83. Vitte C, Panaud O. Formation of solo-LTRs through unequal homologous recombination counterbalances amplifications of LTR retrotransposons in rice Oryza sativa L. Mol Biol Evol. 2003;20:528-540.

84. Schwartz DE, Tizard R, Gilbert W. Nucleotide sequence of Rous sarcoma virus. Cell 1983;32:853-869.

85. Joliot V, Boroughs K, Lasserre F, Crochet J, Dambrine G, et al. Pathogenic potential of myeloblastosis-associated virus: implication of env proteins for osteopetrosis induction. Virology 1993;195:812-819.

86. Lerat E, Birot AM, Samarut J, Mey A. Maintenance in the chicken genome of the retroviral-like cENS gene family specifically expressed in early embryos. J Mol Evol. 2007;65:215-227.

87. Carré-Eusèbe D, Coudouel N, Magre S. OVEX1, a novel chicken endogenous retrovirus with sex-specific and left-right asymmetrical expression in gonads. Retrovirology. 2009;6:59.

88. Krupovic M, Koonin EV. Polintons: a hotbed of eukaryotic virus, transposon and plasmid evolution. Nat Rev Microbiol. 2014;13:105-115.

89. Piriyapongsa J, Polavarapu N, Borodovsky M, McDonald J. Exonization of the LTR transposable elements in human genome.BMC Genomics. 2007;8:291.

90. Piskurek O, Jackson DJ. Transposable elements: from DNA parasites to architects of metazoan evolution. Genes. 2012;3:409-422.

91. Tajnik M, Vigilante A, Braun S, Hänel H, Luscombe NM, et al. Intergenic Alu exonisation facilitates the evolution of tissue-specific transcript ends. Nucleic Acids Res. 2015; Sep 22 [Epub ahead of print].

92. Jordan IK, Rogozin IB, Glazko GV, Koonin EV. Origin of a substantial fraction of human regulatory sequences from transposable elements. Trends Genet. 2003;19:68-72.

93. Kang MI, Rhyu MG, Kim YH, Jung YC, Hong SJ, et al.The length of CpG islands is associated with the distribution of Alu and L1 retroelements. Genomics. 2006;87:580-590.

94. Estécio MR, Gallegos J, Dekmezian M, Lu Y, Liang S, et al. SINE retrotransposons cause epigenetic reprogramming of adjacent gene promoters. Mol Cancer Res. 2012;10:1332-1342.

95. Elferink MG, van As P, Veenendaal T, Crooijmans RP, Groenen MA. Regional differences in recombination hotspots between two chicken populations. BMC Genet. 2010;11:11.

96. Groenen MA, Wahlberg P, Foglio M, Cheng HH, Megens HJ, et al. A high-density SNP-based linkage map of the chicken genome reveals sequence features correlated with recombination rate. Genome Res. 2009;19:510-519.

97. Duvick J, Fu A, Muppirala U, Sabharwal M, Wilkerson MD, Lawrence CJ, et al. PlantGDB: a resource for comparative plant genomics.Nucleic Acids Res. 2008;36:D959-965.

98. Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, et al. Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res. 2012;40:D1178-1186.

99. Albertin CB, Simakov O, Mitros T, Wang ZY, Pungor JR, et al. The octopus genome and the evolution of cephalopod neural and morphological novelties. Nature. 2015;524:220-224.

100. International Chicken Genome Consortium : Gallus-gallus-5.0. http://www.ncbi.nlm.nih.gov/genome/?term=Gallus+gallus (2016) Accessed 2016 Feb 15.

101. Zhang G, Cowled C, Shi Z, Huang Z, Bishop-Lilly KA, et al. Comparative analysis of bat genomes provides insight into the evolution of flight and immunity. Science. 2013;339:456-460.

102. Tanabe H, Habermann FA, Solovei I, Cremer M, Cremer T. Non-random radial arrangements of interphase chromosome territories: evolutionary considerations and 37-45.

103. Perrat PN, DasGupta S, Wang J, Theurkauf W, Weng Z, et al. Transposition-driven genomic heterogeneity in the Drosophila brain. Science. 2013;340:91-95.

104. Li W, Prazak L, Chatterjee N, Grüninger S, Krug L, et al. Activation of transposable elements during aging and neuronal decline in Drosophila. Nat Neurosci. 2013;16:529-531.

105. Erwin JA, Marchetto MC, Gage FH Mobile DNA elements in the generation of diversity and complexity in the brain. Nat Rev Neurosci. 2014;15:497-506.

106. Upton KR, Gerhardt DJ, Jesuadian JS, Richardson SR, Sánchez-Luque FJ, et al. Ubiquitous L1 mosaicism in hippocampal neurons. Cell. 2015;161:228-239.

107. Wang Z, Qu L, Yao J, Yang X, Li G, et al. An EAV-HP insertion in 5' Flanking region of SLCO1B3 causes blue eggshell in the chicken. PLoS Genet. 2013;9:e1003183.

107. Majumdar S, Singh A, Rio DC: The human THAP9 gene encodes an active P-element DNA transposase. Science. 2013;339:446-448.

109. Henssen AG, Henaff E, Jiang E, Eisenberg AR, Carson JR, et al. Genomic DNA transposition induced by human PGBD5. Elife. 2015 Sep 25;4. pii: e10565.

Figure Legends

Fig. 1 Sizes of chromosomes in galGal4. Backgound areas in pink indicate macrochromosomes and those in green indicate microchromosomes. Note, the chromosome numbering set up in caryology does not form a decreasing series in size with galGal4 for chromosomes 6, 11, 15, 16, 17, 18, 19, 22, 23, and 25. LGE22 (LGE22C19W28_E50C23) and LGE64 are two linkage groups whose scaffolds are assembled in two chromosomes but currently have no assigned microchromosomes. The 29 assembled chromosomes plus the two sex chromosomes W and Z and the two LGE contain a total of 1,004,801,586 Mbp. There is also a U chromosome not shown in the graph which contains all the unplaced scaffolds (14,093 sequences containing 42,130,513 bp).

Fig. 2 Strategy for detecting and annotating repeats in galGal4. Our strategy comprised five successive steps: 1, definition of the number of repeats; 2, number of SSRs; 3, number of TEs; 4, definition of dark matter; 5, definition of CNVs. The final products of each of these 5 steps were stored in the bed or gff annotation files (yellow boxes). Arrows show the chronology of events in the processes in each step. Black ellipses show the various states of the genome analysed; blue boxes indicate the programs used; green boxes indicate the intermediate library produced by a given process; red boxes indicate the end of a process before editing the bed or gff annotation files. The purple box in step 5 indicates the source of the annotation file used for CNVs.

Fig. 3 Features of SSRs in Galgal4. A, Coverage of former and new SRR annotations among chromosomes and galGal4. Red bars describe the RM annotations and green bars TRF annotations. The three samples on the right describe the average coverages in chromosomes and linkage groups (Assembled), in the sum of the unassembled scaffolds (Unassembled) and in the complete galGal4 model. **B**, Coverage of each type of SSR in the galGal4 chromosomes. The coverages of simple repeats corresponding to polyA (in red) and polyC (dark blue) stretches are at the bottom of each

bar, microsatellites (yellow) are just above them, minisatellites (light blue) are above them, and uppermost are the two types of tandem arrays, large tandem arrays (orange) and satellite DNAs (green).

Fig. 4 Coverage of TEs and DM annotations in galGal4 chromosomes. A, Percentage coverage of each chromosome by repeats resulting from the REPET annotation (STEP3, Figure 2). **B**, Percentage coverage of each chromosome by TE segments resulting from the DM annotation (STEP4, Figure 2). In **A** and **B**, blue bars indicate the coverage of non-LTR retrotransposons (CR1), red bars LTR-retrotransposons and solo LTRs, yellow bars DNA transposons, green bars repeats of undetermined origin, and kaki bars indicate z-reps (a repeat unique for chromosome Z). In **B**, chromosome 32 was removed because there was no annotation. **C**, Percentage coverage of each chromosome by the RM annotation (blue bars), the REPET annotation (STEP4, Figure 2; red bars) and the sum of TE and DM annotations (STEP3 and 4, Figure 2; green bars).

Fig. 5 Features of annotations calculated by Red, REPET, and RM. A, Venn diagram showing the overlaps between the annotation files calculated with Red (RED), TRF (SSR), and REPET (TE+DM), and CNVs [11]. Values correspond to coverage percentages in galGal4. **B**, Distributions of annotations sizes calculated with Red, TRF (SSR) and REPET (DM, TE and TE+DM), and those of the CNVs [11]. DM annotations were split into two batches corresponding to DM annotations that extend pre-existing annotations produced with the same TE model (DM extended) and those that are new (DM new). **C**, Size distributions of LINE, LTR, TIR and SSR annotations calculated with RM together with those obtained with REPET or TRF for the same categories. Vertical axes in A and B indicate $\log_{10}$(sizes) in bp. The red lines in the box plot indicate the median value, the ends of grey boxes the quartile 1 and 3 values, the ends of whisker the 10th and 91st percentiles of the size

distribution, and the black stars the highest and the values above or below the 1.5 interquartile range respectively within 1.5 interquartile range of the highest or the lowest quartile.

Fig. 6 Sequence organization of retroCalimero (A), retroSaturnin (B) and retroTux (C). Red boxes indicate 361-bp LTRs in retroCalimero (6837-bp), 334-bp LTRs in retroSaturnin (4624-bp) and 498-bp LTRs in retroTux (5800-bp). Cyan boxes indicated polypurine tracts (ppt) just upstream of the 3' LTR. Yellow boxes indicate regions of interrupted coding frames for Gag or RT detected on the sense strand and green boxes these regions on the anti-sense strand. We found interrupted frames encoding an RT on the sense strand and a Gag-like protein (so-called natural cytotoxicity triggering receptor 3 ligand 1 precursor among blastx hits obtained with the nucleic acid database at the NCBI website) on the anti-sense strand in the inner regions of retroTux. Nucleic acid sequences are shown in Additional file 9.

Fig. 7 Graph showing the expected and observed abundances of copies of TEs (A), genes (B), S/MAR (C) and CpG islands (D) in galGal 4 chromosomes. Each box was calculated from 100000 permutations and represents the 98% distributions obtained per chance. Red crosses above the boxes indicate over-representation of the element in the chromosome and blue crosses under-representation (p>99% in both cases). Pink backgound areas indicate macrochromosomes and green areas indicate microchromosomes. All galGal4 chromosomes were analysed except chromosome 32, which was too small (1028 bases).

Fig. 8 Coverages of TEs in galGal4 chromosomes with respect to the numbers of genes (A, B and C), CpG islands and S/MAR (D). Histograms in **A** and **B** show the coverages of TE copies annotated by REPET and those of the [TE+DM] annotation in each chromosome. The names of each of the 34 models are indicated in the left margin. The 3 bars near the abscissa describe data for

the 34 TE models (all), those of the ISB annotations (ISB), and the proportions of the exonic, genic and intergenic regions in galGal4 (GG4). A grey background indicates one of the four TE types in galGal4: LINE, LTR, TIR and U (undetermined). The name is shown in the right margin. Histogram in **B** shows the proportions of TEs (percent coverage or number of copies). Background areas in green indicate TE data from the ISB annotation, light purple indicates the REPET (TE) annotations, and blue indicates the [TE+DM] annotations. The bar (GG4) near the abscissa shows the proportions of exons, introns and intergenes in galGal4. In **A**, **B** and **C,** the exons (non-coding and coding) are shown in red, genes are in yellow and the intergene regions are in green. A purple vertical bar indicates the size of the intergene regions in galGal4. The histogram in **D** represents the coverage/ percentages of CpG islands, S/MAR elements and TEs inserted in CpG islands and S/MAR elements (blue), the 3-kbp distal and 3-kbp proximal ends of CpG islands and S/MAR elements (green) and in the rest of the chromosomes (purple).

Fig. 9 TE density in galGal4 chromosomes. Histograms of TE model densities calculated for all galGal4 chromosomes, except chromosome 32 (too small; 1028 bp). **A**, All TE models, **B**, CR1, **C** Kronos, **D** retroCalimero, **E**, putative_LTR_group 9 and **F**, Charlie. The number of copies for each dataset are indicated in parentheses. Results for all other TE models are shown in additional files 14 and 15.

Fig. 10 Density of TE hot spots in galGal4 chromosomes. Histograms of TE hot spot density calculated for all galGal4 chromosomes, except chromosome 32 (too small; 1028 bp). Hot spot are defined (p>99%) using permutation assays with **A**, All TE models, **B**, Kronos, **C**, putative_LTR_group4, **D** retroCalimero, **E**, Charli and **F**, mariner1_GG. The number of copies for each dataset are indicated in parenthesis. Results for all other TE models are shown in additional files 16 and 17.

Conclusion

L’annotation des séquences répétées est une étape négligée lors de la publication d’un génome. Elle est produite en utilisant des banques de séquences consensus généralistes, c’est-à-dire des familles d’ETs les plus nombreuses et mieux connues de chaque génome. Cette stratégie est simple et rapide, mais a pour effet de créer une couche d’annotations ne couvrant pas la totalité des répétitions dispersées et de créer de fausses annotations. Ceci est illustré dans le génome du poulet où l’annotation RM contient, entre autres, des SINEs ou des TIRs de l’espèce Chompy alors qu’ils sont inexistants.

Nous avons mis au point une méthode d’annotation globale de tous les types de répétitions d’un génome eucaryote. Celle-ci s’appuie sur des programmes faisant de la détection *de novo* des séquences répétées au lieu d’utiliser des librairies de consensus de référence pour l’essentiel extra-spécifiques.

L’annotation des séquences satellites est faite par le programme TRF. Généralement employé pour annoter les micro-satellites et les mini-satellites, nous avons fait le choix de rechercher les répétitions en tandem dont la taille de l’unité de répétition peut aller jusqu’à 2 kpb, permettant ainsi de détecter les séquences satellites.

L’annotation des ETs se fait en deux étapes, l’annotation des copies récentes et conservées puis l’annotation des copies anciennes, petites et dégradées, la matière noire. Ces deux annotations utilisent le pipeline REPET qui permet de créer une librairie de consensus d’ETs *de novo*, puis de l’utiliser comme base de référence. Pour la détection *de novo* des ETs, nous avons mis en place une validation des consensus générés pour sélectionner uniquement ceux de haute qualité et éliminer les consensus chimères regroupant plusieurs ETs et les faux positifs. Cela a permis de créer une banque de 499 séquences consensus décrivant les différents types d’ETs. La triple annotation et le post-traitement des résultats effectué par la partie TEannot de REPET permettent de produire une annotation de qualité avec un taux de faux positifs très faible grâce au post-traitement de l’annotation et dont le taux de couverture est supérieur à l’annotation produite par RM avec Repbase.

Pour la première fois, nous avons réalisé l’annotation de la matière noire dans un génome de vertébré de grande taille (c.a.d. > à 1 Gpb). Nous avons appliqué une méthode élaborée pour l’analyse du génome de *Arabidopsis thaliana*. Elle consiste à ne plus utiliser les consensus pour annoter le génome, mais les copies d’éléments transposables détectées lors de

la première étape. Cela permet de retrouver les copies les plus dégénérées en utilisant la diversité des copies précédemment annotées et ainsi de compléter l'annotation des éléments transposables.

Ces travaux ont permis de définir une nouvelle façon de décrire les ETs. Communément, les ETs sont décrits à l'aide d'une seule séquence consensus. Cependant, cela ne permet pas de décrire toute la variabilité des copies générées par leur dérive et l'accumulation des mutations. En s'inspirant des modèles proposés dans RepeatExplorer et REPET, nous avons proposé de décrire les ETs grâce à des modèles. Ils consistent en une série de consensus décrivant la séquence de l'ET complet ainsi que toutes les différentes versions présentes dans le génome. Nous avons appliqué ce concept à notre base de séquence consensus, ce qui nous a permis de former 34 modèles d'ETs. On retrouve les ETs précédemment décrits par Wicker et dans Repbase (CR1, Charlie, Galluhop, Charlie-Galluhop, Kronos, Hitchcock, Soprano, Birddawg, EAV, EAV-HP, ERV2, ERV7, ERV11, Mariner1_GG), mais aussi de nouveaux rétrotransposons à LTR (retroSaturnin, retroCalimero, retroTux), une série de modèles suspectés d'être des solo-LTR de rétrotransposons (putative_LTR_group4-9-12-22-28-30), une série de modèles dont le type ne peut pas être déterminé, et enfin un modèle regroupant les répétitions liées au méga-satellite du chromosome Z (Z-REP).

Cette nouvelle couche d'annotations de haute résolution des ETs nous a permis d'étudier la dynamique des ETs dans le génome. Nous avons analysé leur distribution dans les chromosomes par rapport aux gènes, aux îlots de CpG et sMAR. Nous avons démontré qu'il existe 3 types de chromosomes, les macro- (1, 2, 3, 4, Z), les médiums (5, 6, 7, 8 and 9) et les microchromosomes (10 à 32). Les macro-chromosomes sont enrichis en ETs et en sMARs, pauvres en gènes et en îlots de CpG. Les microchromosomes sont riches en gènes et îlots de CpG, pauvres en ETs et sMARs. Les médiums sont pauvres en ETs et îlots de CpG, riches en gènes et ont une quantité variable de sMARs.

Nous avons étudié la distribution des ETs entre les chromosomes et mis en évidence plusieurs profils d'insertion variant entre les différents modèles. Nous avons étudié la distribution des ETs au sein de chaque chromosome et démontré la présence de régions favorables à l'insertion des ETs (« hot spots »). Ces régions diffèrent en fonction des modèles d'ETs. Cependant, les résultats ne permettent pas de conclure que ces « hot spots » sont issus

de préférence d'insertions. En effet, les copies annotées ayant accumulé de nombreuses mutations qui les rendent inactives, il n'est pas possible de savoir si les profils obtenus sont issus de préférences d'insertion dans certaines régions des chromosomes ou si elles proviennent de l'incapacité du génome à éliminer les ETs dans ces régions.

Cette nouvelle méthode d'annotation des répétitions de tout type permet une amélioration aussi bien quantitative que qualitative. Cependant, l'évaluation de la quantité globale de répétitions s'est avérée insatisfaisante avec les outils utilisés. Les deux programmes, P-clouds et Red, sont capables de détecter tous les types de répétitions dans un génome et de les cartographier. Malheureusement, les annotations obtenues ne couvrent pas la totalité des annotations produites pour les ETs et les répétitions en tandem. Cela nous a amenés à remettre en cause les résultats obtenus avec ces programmes et à conclure que les algorithmes sous-jacents sont encore à améliorer.

Aboutir à une méthode d'annotation permettant l'obtention de consensus de qualité et une annotation de toutes les copies d'ETs, des plus jeunes aux plus anciennes, a exigé un long travail d'étalonnage. Il a fallu travailler sur de nombreux points de REPET afin de comprendre son fonctionnement et ainsi calibrer et configurer les différents composants. En effet, lors des premières analyses, les résultats de couverture des ETs étaient prometteurs et pouvaient dépasser 30 %. Après analyse de la couche d'annotations, nous avons remarqué que plus de 95 % des copies avaient été annotées par des consensus créés par la branche d'analyse structurale de REPET. Comme LTR_HARVERST recherche seulement deux longues répétitions terminales d'une taille maximale donnée, lorsqu'il est exécuté sur le modèle de la poule rouge de jungle, il génère des consensus à partir de fragments de CR1 qu'il considère comme des LTRs. Ce biais a pour effet de générer un très grand nombre de consensus de grande taille correspondant à une portion d'ADN unique du génome (c.à.d. de l'ADN non répété). Ces constatations nous ont amenés à exclure la partie de détection structurale de la détection *de novo*. Malgré cette optimisation, certains consensus produits annotaient une seule copie dans le génome. Ceci est problématique car dans le processus de détection par similarité, les consensus sont générés à partir de trois répétitions. Malgré les tentatives pour agir sur les paramètres d'identité de REPET, nous n'avons pu éliminer ces consensus. Avec l'aide de l'URGI, nous avons mis au point une sélection des consensus par l'annotation qui consiste à ne conserver que les consensus annotant au moins trois copies dans le génome.

REPET est doté d'un programme, PASTEC, capable de classer les consensus *de novo* dans la classification de Wicker. Cependant, lorsqu'il est utilisé pour analyser les consensus produits à partir du génome du poulet, cette classification est fausse ou peu précise pour un très grand nombre d'entre eux. De plus, la dernière étape de TEdenovo est censée regrouper les consensus proches et ainsi former des familles d'ETs. Lors des analyses faites sur le génome du poulet, cette étape n'a pas permis de former de familles. Ces problèmes nous ont obligés à analyser une à une les séquences consensus produites pour leur assigner un type et les associer à une famille en les comparant aux séquences connues et en recherchant manuellement des structures particulières permettant leur classification.

L'implémentation de REPET nous a également posé des problèmes. La prise en charge des erreurs par le pipeline est minimale, les messages lancés sont très vagues et n'indiquent pas clairement l'origine de l'erreur. Ceci est gênant, car lors des étapes de REPET réalisées en parallèle sur un cluster de calcul, si une des tâches est en erreur, il est obligatoire de relancer toutes les tâches de l'étape. Cet inconvénient génère une perte de temps de calcul considérable, surtout si l'erreur en jeu n'est pas corrigée au premier essai.

Notre nouvelle méthode d'annotation a permis de faire des progrès sur l'étude de la dynamique du génome de poulet, mais l'analyse d'un génome de grande taille reste un challenge difficilement applicable en routine. En effet, même avec sa petite taille pour un génome vertébré, les analyses REPET sont longues, très consommatrices de calculs et ont besoin d'être améliorées, que ce soit d'un point de vue stratégie ou algorithmique. Dans le cadre de la thèse, nous avons réalisé des analyses sur le génome entier du poulet. Cette opération a l'avantage de permettre de détecter les ETs avec peu de copies dispersées sur plusieurs chromosomes, mais la quantité de calculs nécessaire pour analyser la totalité des séquences est d'autant plus grande. Ce coût peut être réduit en procédant à une première analyse sur un échantillon des séquences du génome, ce qui n'est pas sans inconvénient. Elle permettrait de détecter les ETs les plus fréquents et de les annoter, puis de produire un génome sans ces copies. Le génome ainsi obtenu présenterait un taux de répétition plus faible et serait analysable plus rapidement ou sur un cluster de calcul de taille plus modeste. Il serait aussi possible de commencer par une annotation à l'aide d'une base d'ETs de qualité d'une espèce proche, de valider les consensus de cette base en conservant uniquement les consensus

ayant au moins 3 copies complètes, puis de supprimer les copies des consensus sélectionnés du génome avant de procéder à l'analyse *de novo*.

La réduction du temps de calcul peut passer par l'exécution d'une étape de l'annotation. Afin de supprimer les annotations qui pourraient être issues du hasard, TEannot procède à une annotation sur le génome dans lequel les bases ont été distribuées aléatoirement. Cela permet de déduire un score en dessous duquel il est impossible de savoir si la copie a été obtenue aléatoirement ou si elle est réelle. Les scores déterminés par REPET pourraient être appliqués aux autres TEannot sans avoir à recommencer l'annotation sur le génome aléatoire.

Les temps de calcul pourraient être également optimisés en améliorant certaines étapes du pipeline. Dans le pipeline TEdenovo, l'étape de clustering suivant la comparaison génome contre génome n'est pas parallélisée. Ces analyses peuvent être longues et consommatrices de calculs. Il serait possible d'accélérer leur exécution en réalisant un post-traitement des données brutes afin de former des pré-clusters, c'est-à-dire réaliser un clustering à faible contrainte pour former des groupes pouvant contenir plusieurs familles proches d'ETs, pour ensuite analyser chacune d'entre elles en parallèle à l'aide de GROUPER, RECON, PILER.

L'un des points à améliorer pour pouvoir appliquer notre méthode d'annotation en routine concerne la classification automatique des consensus. Analyser manuellement chaque séquence consensus pour leur attribuer un type est très chronophage. Le programme PASTEC, chargé de cette étape dans REPET, utilise des programmes (agents), dont l'objectif est de noter un consensus vis-à-vis d'un type d'ET. Étant initialement développés et testés sur des génomes de plantes, les agents actuellement en place ne pas sont efficaces pour classer les ETs de vertébrés. Cependant, dans ce système, il est possible d'ajouter de nouveaux agents spécialisés dans la reconnaissance de type d'ETs spécifiques des vertébrés.

Lors du développement de la méthode, nous avons eu accès à un grand nombre de machines et de nombreux processeurs en utilisant le cluster de GénoToul. Cependant, l'utilisation de ressources partagées nationalement présente des contraintes au niveau du nombre de processeurs utilisables en même temps, le nombre de calculs, le stockage des données, le nombre de tâches soumises simultanément, etc. La rapide baisse des coûts du matériel informatique, l'augmentation du nombre de cœurs par processeur et les améliorations

possibles sur la méthode d'annotation rendent envisageable l'acquisition du matériel nécessaire pour réaliser les annotations localement. Aujourd'hui, un serveur équipé de deux processeurs de 22 cœurs physiques (88 cœurs virtuels au total), 256 Go de RAM et 10 To de stockage coûte environ 26 000 €. L'acquisition d'une telle machine présente plusieurs avantages. Elle serait en mesure de procéder à l'annotation des ETs dans des temps raisonnables (une semaine à une semaine et demie) et à celle de la matière noire en 1 mois et demi. De plus, maximiser le nombre de processeurs par machine permet de se contenter d'un seul serveur et ainsi de diminuer la maintenance et la consommation électrique.

Ces dernières années, les progrès réalisés par les programmes d'annotations offrent la possibilité de créer des annotations de haute qualité s'ils sont utilisés correctement. Les méthodes *de novo* nous ont permis de redécouvrir la population en ETs du génome du poulet et de remettre en cause les connaissances jusqu'alors acceptées de tous. Notre annotation de la matière noire démontre que les répétions sont bien plus nombreuses que celles attendues. Ces informations non recherchées pourraient être pourtant d'une grande importance. Bien que les copies puissent provenir d'une élimination active, il est possible que certains fragments aient été cooptés grâce à un élément régulateur dont ils sont porteurs. Il a déjà été démontré que certains éléments transposables agissent sur l'expression des gènes, chaque fragment de matière noire annoté représente une voie de régulation potentielle. La recherche de la matière noire dans les génomes devrait à l'avenir faire partie des procédures standards d'annotation des génomes.

L'annotation *de novo* d'une espèce déjà étudiée, le poulet, a montré que la diversité en ETs était bien plus faible que ce que l'on attendait. RM utilise 317 consensus pour annoter le génome, alors que notre nouvelle base ne recense que 34 familles d'ETs. Nos modèles rassemblent pourtant 499 consensus. Cela est dû à la grande précision de ceux-ci qui décrivent un très grand nombre de copies dérivées des CR1. Pourtant, lors de nos tentatives pour définir les familles de CR1, nous n'avons pu former que 8 groupes de CR1s alors que Repbase en comporte 22. Cette différence de diversité met en cause plusieurs études tentant d'estimer les pics de multiplication des CR1 au cours du temps. Pour que les prochaines études se basant sur la diversité d'éléments transposables soient viables, il sera donc nécessaire de passer par une étape de détection *de novo* pour cerner la totalité de la diversité des copies d'une espèce d'ET. Cette faible diversité en ETs est une caractéristique inhabituelle

pour un génome de grande taille. En effet, la diversité observée dans le génome du poulet est extrêmement faible vis-à-vis de ce qui est connu dans des génomes comme celui de l'homme et de la souris qui contiennent 558 et 287 familles d'ETs, ceux de plantes comme A*rabidopsis thaliana* ou le riz qui comportent 461 et 992 familles d'ETs. Ce passage de 317 à 34 sortes d'ETs différents nous amène à nous questionner sur la diversité des ETs dans les génomes des autres espèces. Refaire l'analyse *de novo* de génomes déjà annotés permettra de déterminer la véritable diversité en ETs.

Les travaux de cette thèse pourront être utilisés pour annoter les espèces proches du poulet (galliformes) à l'aide de la banque de consensus que nous avons créée, ou réaliser des annotations de haute qualité sur les espèces éloignées des autres clades en procédant à l'analyse complète pour détecter les populations d'ETs.

Bibliographie

Agarwal P, States DJ: The Repeat Pattern Toolkit (RPT): analyzing the structure and evolution of the C. elegans genome. Proc Int Conf Intell Syst Mol Biol 1994, 2:1–9.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: Basic local alignment search tool. J Mol Biol 1990, 215:403–410.

Andraszek K, Smalec E: Structure and functions of lampbrush chromosomes. BioTechnologia 2011, 4:337–344.

Andreozzi L, Federico C, Motta S, Saccone S, Sazanova AL, Sazanov AA, Smirnov AF, Galkina SA, Lukina NA, Rodionov A V., Carels N, Bernardi G: Compositional mapping of chicken chromosomes and identification of the gene-richest regions. Chromosom Res 2001, 9:521–532.

Arthur and RR, Straus NA. DNA-sequence organization in the genome of the domestic chicken (Gallus domesticus). Can J Biochem 1978, 56:257–63.

Axelsson E, Webster MT, Smith NGC, Burt DW, Ellegren H: Comparison of the chicken and turkey genomes reveals a higher rate of nucleotide divergence on microchromosomes than macrochromosomes. Genome Res 2005, 15:120–5.

Balakrishnan CN, Mukai M, Gonser R a, Wingfield JC, London SE, Tuttle EM, Clayton DF: Brain transcriptome sequencing and assembly of three songbird model systems for the study of social behavior. PeerJ 2014, 2:e396.

Bao W, Kojima KK, Kohany O: Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob DNA 2015, 6:11.

Bao Z, Eddy SR: Automated De Novo Identification of Repeat Sequence Families in Sequenced Genomes. 2002:1269–1276.

Bed'Home B, Coullin P, Guillier-Gencik S, Moulin S, et al. Characterization of the atypical karyotype of the black-winged kite Elanus caeruleus (Falconiformes: Accipitridae) by means of classical and molecular cytogenetic techniques. Chromosome Res. 2003;11:335-343.

Belancio VP, Roy-Engel AM, Deininger PL: All y'all need to know 'bout retroelements in cancer. Semin Cancer Biol 2010, 20:200–210.

Benson G: Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 1999, 27:573–80.

Biémont C, Vieira C: What transposable elements tell us about genome organization and evolution: the case of Drosophila. Cytogenet Genome Res 2005, 110:25–34.

Bire S, Rouleux-Bonnin F. Transposable elements as tools for reshaping the genome: it is a huge world after all! Methods Mol Biol. 2012;859:1-28.

Bire S, Casteret S, Piégu B, Beauclair L, Moiré N, Arensbuger P, Bigot Y: Mariner Transposons Contain a Silencer: Possible Role of the Polycomb Repressive Complex 2. PLOS Genet 2016, 12:e1005902.2.

Böhne A, Brunet F, Galiana-Arnoux D, Schultheis C, Volff JN: Transposable elements as drivers of genomic and biological diversity in vertebrates. Chromosom Res 2008, 16:203–215.

Bourque G, Leong B, Vega VB, Chen X, Lee YL, Srinivasan KG, Chew J-L, Ruan Y, Wei C-L, Ng HH, Liu ET: Evolution of the mammalian transcription factor binding repertoire via transposable elements. Genome Res 2008, 18:1752–1762.

Brandström M, Ellegren H. Genome-wide analysis of microsatellite polymorphism in chicken circumventing the ascertainment bias. Genome Res. 2008;18:881-887.

Callinan a, Batzer M a., Callinan P a: Retrotransposable Elements and Human Disease. Genome Dyn 2006, 1:104–115.

Camiolo S, Porceddu A: Gff2Sequence, a New User Friendly Tool for the Generation of Genomic Sequences. BioData Min 2013, 6:15.

Capy P: Structure et évolution des éléments transposables. J Soc Biol 2004, 198:393–398.

Carré-Eusèbe D, Coudouel N, Magre S: OVEX1, a novel chicken endogenous retrovirus with sex-specific and left-right asymmetrical expression in gonads. Retrovirology 2009, 6:59.

Castoe T a, Hall KT, Guibotsy Mboulas ML, Gu W, de Koning a PJ, Fox SE, Poole AW, Vemulapalli V, Daza JM, Mockler T, Smith EN, Feschotte C, Pollock DD: Discovery of highly divergent repeat landscapes in snake genomes using high-throughput sequencing. Genome Biol Evol 2011, 3:641–53.

Chen Y, Zhou F, Li G, Xu Y: MUST: A system for identification of miniature inverted-repeat transposable elements and applications to Anabaena variabilis and Haloquadratum walsbyi. Gene 2009, 436:1–7.

Clement T, Kutish GF, Nezworski J, Scaria J, Nelson E, Christopher-Hennings J, Diel DG: Complete Genome Sequence of a Highly Pathogenic Avian Influenza Virus (H5N2) Associated with an Outbreak in Commercial Chickens, Iowa, USA, 2015. Genome Announc 2015, 3:e00613–15.

Coffin J, Hughes S, Varmus H: Retroviruses. Cold Spring Harbor Laboratory Press; 1997.

Curcio MJ, Derbyshire KM: The outs and ins of transposition: from mu to kangaroo. Nat Rev Mol Cell Biol 2003, 4:865–77.

de Koning APJ, Gu W, Castoe TA, Batzer MA, Pollock DD: Repetitive elements may comprise over two-thirds of the human genome. PLoS Genet 2011, 7:e1002384.

Delany ME, Gessaro TM, Rodrigue KL, Daniels LM. Chromosomal mapping of chicken mega-telomere arrays to GGA9, 16, 28 and W using a cytogenomic approach. Cytogenet Genome Res. 2007;117:54-63.

Delany ME, Krupkin AB, Miller MM. Organization of telomere sequences in birds: evidence for arrays of extreme length and for in vivo shortening. Cytogenet Cell Genet. 2000; 90:139-145.

Doležel J, Bartoš J,Voglmayr H, Greilhuber J. Letter to the editor: Nuclear DNA Content and Genome Size of Trout and Human. Cytometry 2003;51A:127–128.

Doran TJ, Cooper CA, Jenkins KA, Tizard ML V.: Advances in genetic engineering of the avian genome: “Realising the promise.” Transgenic Res 2016.

Du C, Caronna J, He L, Dooner HK: Computational prediction and molecular confirmation of Helitron transposons in the maize genome. BMC Genomics 2008, 9:51.

Eden FC, Hendrick JP, Gottlieb SS. Homology of single copy and repeated sequences in chicken, duck, Japanese quail, and ostrich DNA. Biochemistry 1978;17:5113-5121.

Edgar RC, Myers EW: PILER: identification and classification of genomic repeats. Bioinformatics 2005, 21 Suppl 1:i152–8.

Edgar RC: MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acid Res 2004, 32:1792–1797.

Ellegren H: The avian genome uncovered. Trends Ecol Evol 2005, 20:180–6.

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

Epplen JT, Leipoldt M, Engel W, Schmidtke J. DNA sequence organisation in avian genomes. Chromosoma 1978, 69:307–321.

Faulkner GJ: Retrotransposons: mobile and mutagenic from conception to death. FEBS Lett 2011, 585:1589–94.

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

Finnegan DJ: Eukaryotic transposable elements and genome evolution. Trends Genet 1989, 5(C):103–107.

Finnegan J: Transposable elements. Curr Biol 1992, 2:861–867.

Flutre T, Duprat E, Feuillet C, Quesneville H: Considering transposable element diversification in de novo annotation approaches. PLoS One 2011, 6:e16526.

Friedman-Einat M, Cogburn LA, Yosefi S, Hen G, Shinder D, et al. Discovery and characterization of the first genuine avian leptin gene in the rock dove (Columba livia). Endocrinology. 2014;155:3376-3384.

Girgis HZ: Red: an intelligent, rapid, accurate tool for detecting repeats de-novo on the genomic scale. BMC Bioinformatics 2015, 16:227.

Gobes SMH, Zandbergen M a, Bolhuis JJ: Memory in the making: localized brain activation related to song learning in young songbirds. Proc Biol Sci 2010, 277(June):3343–3351.

Gregory, T.R. (2005). Animal Genome Size Database. http://www.genomesize.com

Gu W, Castoe T a, Hedges DJ, Batzer M a, Pollock DD: Identification of repeat structure in large genomes using repeat probability clouds. Anal Biochem 2008, 380:77–83.

Ha T: Journal of General Virology. J Gen Virol 2008(C):2–3.

Habermann F a, Cremer M, Walter J, Kreth G, von Hase J, Bauer K, Wienberg J, Cremer C, Cremer T, Solovei I: Arrangements of macro- and microchromosomes in chicken cells. Chromosome Res 2001, 9:569–84.

Häsler J, Samuelsson T, Strub K: Useful “junk”: Alu RNAs in the human transcriptome. Cell Mol Life Sci 2007, 64:1793–800.

Hawkins JS, Kim H, Nason JD, Wing R a, Wendel JF: Differential lineage-specific amplification of transposable elements is responsible for genome size variation in Gossypium. Genome Res 2006, 16:1252–61.

Hillier LW: Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature 2004, 432:695–716.

Hori T, Suzuki Y, Solovei I, Saitoh Y, Hutchison N, Ikeda J, Macgregor H, Mizuno S: Characterization of DNA sequences constituting the terminal heterochromatin of the chicken Z chromosome. Chromosom Res 1996, 4:411–426.

Hron T, Pajer P, Pačes J, Bartůněk P, Elleder D. Hidden genes in birds. Genome Biol. 2015;16:164.

http://repeatmasker.org/genomicDatasets/RMGenomicDatasets.html

Huang Y, Zhang H, Li X, Hu S, Cai L, Sun Q, Li W, Deng Z, Xiang X, Zhang H, Li F, Gao L: Detection and genetic characteristics of H9N2 avian influenza viruses from live poultry markets in Hunan Province, China. PLoS One 2015, 10:1–12.

Hughes AL, Hughes MK: Small genomes for better flyers. Nature 1995:391.

Jarvis ED, Mirarab S, Aberer AJ, Li B, Houde P, Li C, Ho SYW, Faircloth BC, Nabholz B, Howard JT, Suh A, Weber CC, da Fonseca RR, Li J, Zhang F, Li H, Zhou L, Narula N, Liu L, Ganapathy G, Boussau B, Bayzid MS, Zavidovych V, Subramanian S, Gabaldon T, Capella-Gutierrez S, Huerta-Cepas J, Rekepalli B, Munch K, Schierup M, et al.: Whole-genome analyses resolve early branches in the tree of life of modern birds. Science (80-) 2014, 346:1320–1331.

Kalyanaraman A, Aluru S: Efficient algorithms and software for detection of full-length LTR retrotransposons. J Bioinforma … 2006, 4:197–216.

Kapitonov V V, Jurka J: A universal classification of eukaryotic transposable elements implemented in Repbase. Nat Rev Genet 2008, 9:411–2; author reply 414.

Kapitonov V V, Jurka J: Helitrons on a roll: eukaryotic rolling-circle transposons. Trends Genet 2007, 23:521–9.

Kapitonov V V., Jurka J: RAG1 core and V(D)J recombination signal sequences were derived from Transib transposons. PLoS Biol 2005, 3:0998–1011.

Katoh K, Misawa K, Kuma K, Miyata T: MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res 2002, 30:3059–3066.

Kazazian HH: Mobile Elements: Drivers of Genome Evolution. Science (80-) 2004, 303:1626–1632.

Kelley DR, Salzberg SL. Detection and correction of false segmental duplications caused by genome mis-assembly. Genome Biol. 2010;11:R28.

Kolpakov R: mreps: efficient and flexible detection of tandem repeats in DNA. Nucleic Acids Res 2003, 31:3672–3678.

Krasikova A, Fukagawa T, Zlotina A: High-resolution mapping and transcriptional activity analysis of chicken centromere sequences on giant lampbrush chromosomes. Chromosome Res 2012, 20:995–1008.

Krishnan J: Code in Non-coding. *Proc Indian Natl Sci Acad* 2015, 81:609–628.

Kurtz S, Narechania A, Stein JC, Ware D: A new method to compute K-mer frequencies and its application to annotate large repetitive plant genomes. BMC Genomics 2008, 9:517.

Lerat E: Identifying repeats and transposable elements in sequenced genomes: how to find your way through the dense forest of programs. Heredity (Edinb) 2010, 104:520–33.

Liu H, Wang X, Wang J, Zhao Y: Genome Sequences of an H5N1 Highly Pathogenic Avian Influenza Virus Isolated from Vaccinated Layers in China in 2012. Genome … 2013, 1:2012–2013.

Liu L, Li Y, Li S, Hu N, He Y, Pong R, Lin D, Lu L, Law M: Comparison of next-generation sequencing systems. *J Biomed Biotechnol* 2012, 2012:251364.

Lovell PV, Wirthlin M, Wilhelm L, Minx P, Lazar NH, et al. Conserved syntenic clusters of protein coding genes are missing in birds. Genome Biol. 2014;15:565.

Loytynoja A, Goldman N: webPRANK: a phylogeny-aware multiple sequence aligner with interactive alignment browser. BMC Bioinformatics 2010, 11(C):579.

Lucier J-F, Perreault J, Noël J-F, Boire G, Perreault J-P: RTAnalyzer: a web application for finding new retrotransposons and detecting L1 retrotransposition signatures. Nucleic Acids Res 2007, 35(Web Server issue):W269–74.

Marçais G, Kingsford C: A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics 2011, 27:764–770.

Maumus F, Quesneville H: Deep investigation of Arabidopsis thaliana junk DNA reveals a continuum between repetitive elements and genomic dark matter. PLoS One 2014, 9:e94101.

McCarthy EM, McDonald JF: LTR STRUC: A novel search and identification program for LTR retrotransposons. Bioinformatics 2003, 19:362–367.

McClintock B: The origin and behavior of mutable loci in maize. Proc Natl Acad Sci 1950, 36:344–355.

McQueen HA, Siriaco G, Bird AP: Chicken microchromosomes are hyperacetylated, early replicating, and gene rich. Genome Res 1998, 8:621–630.

McQueen, H.A. et al. (1996) CpG islands of chicken are concentrated on microchromosomes. Nat. Genet. 12, 321–324

Mello C V., Clayton DF: The opportunities and challenges of large-scale molecular approaches to songbird neurobiology. Neurosci Biobehav Rev 2015, 50:70–76.

Mello C V.: The Zebra finch, taeniopygia guttata: An avian model for investigating the neurobiological basis of vocal learning. Cold Spring Harb Protoc 2014, 2014:1237–1242.

Mendonça MA, Carvalho CR, Clarindo WR. DNA content differences between male and female chicken (Gallus gallus domesticus) nuclei and Z and W chromosomes resolved by image cytometry. J Histochem Cytochem. 2010;58:229-235.

Miele V, Penel S, Duret L: Ultra-fast sequence clustering from similarity networks with SiLiX. BMC Bioinformatics 2011, 12:116.

Morgulis A, Gertz EM, Schäffer AA, Agarwala R: WindowMasker: Window-based masker for sequenced genomes. Bioinformatics 2006, 22:134–141.

Nanda I, Benisch P, Fetting D, Haaf T, Schmid M: Synteny conservation of chicken macrochromosomes 1-10 in different avian lineages revealed by cross-species chromosome painting. Cytogenet Genome Res 2011, 132:165–181.

Nanda I, Schmid M. Localization of the telomeric (TTAGGG)n sequence in chicken (Gallus domesticus) chromosomes. Cytogenet Cell Genet. 1994;65:190-193.

Nelson MI, Pollett S, Ghersi B, Silva M, Simons MP, Icochea E, Gonzalez AE, Segovia K, Kasper MR, Montgomery JM, Bausch DG: The Genetic Diversity of Influenza A Viruses in Wild Birds in Peru. PLoS One 2016, 11:e0146059.

Novák P, Neumann P, Macas J: Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. BMC Bioinformatics 2010, 11:378.

Novak P, Neumann P, Pech J, Steinhaisl J, Macas J: RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. Bioinformatics 2013, 29:792–793.

O'Hare TH, Delany ME. Genetic variation exists for telomeric array organization within and among the genomes of normal, immortalized, and transformed chicken systems. Chromosome Res. 2009;17:947-964.

Ohno S: So much “junk” DNA in our genome. *Evol Genet Syst* 1972, 23:366–370.

Oliver KR, Greene WK: Mobile DNA and the TE-Thrust hypothesis: supporting evidence from the primates. Mob DNA 2011, 2:8.

Olofsson B, Bernardi G. Organization of nucleotide sequences in the chicken genome. Eur J Biochem. 1983;130:241-245.

Organ CL, Shedlock AM, Meade A, Pagel M, Edwards SV. Origin of avian genome size and structure in non-avian dinosaurs. Nature. 2007;446:180-184.

Orgel LE, Crick FHC: Selfish DNA: the ultimate parasite. *Nature* 1980, 284:604–607.

Pardue M-L, DeBaryshe PG: Retrotransposons Provide an Evolutionarily Robust Non-Telomerase Mechanism to Maintain Telomeres. Annu Rev Genet 2003, 37:485–511.

Peaston AE, Evsikov A V, Graber JH, de Vries WN, Holbrook AE, Solter D, Knowles BB: Retrotransposons regulate host genes in mouse oocytes and preimplantation embryos. Dev Cell 2004, 7:597–606.

Piégu B, Bire S, Arensburger P, Bigot Y: 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 2015, 86:90–109.

Piegu B, Guyot R, Picault N, Roulin A, Sanyal A, Saniyal A, Kim H, Collura K, Brar DS, Jackson S, Wing R a, Panaud O: Doubling genome size without polyploidization: dynamics of retrotransposition-driven genomic expansions in Oryza australiensis, a wild relative of rice. Genome Res 2006, 16:1262–9.

Ponicsan SL, Kugel JF, Goodrich JA: Genomic gems: SINE RNAs regulate mRNA production. Curr Opin Genet Dev 2010, 20:149–55.

Poynter G, Huss D, Lansford R: Japanese quail: An efficient animal model for the production of transgenic avians. Cold Spring Harb Protoc 2009, 4:1–7.

Primmer CR, Raudsepp T, Chowdhary BP, Møller AP, Ellegren H: Low frequency of microsatellites in the avian genome. Genome Res 1997, 7:471–82.

Quesneville H, Bergman CM, Andrieu O, Autard D, Nouaud D, Ashburner M, Anxolabehere D: Combined evidence annotation of transposable elements in genome sequences. PLoS Comput Biol 2005, 1:166–75.

Quesneville H, Bergman CM, Andrieu O, Autard D, Nouaud D, Ashburner M, Anxolabehere D: Combined evidence annotation of transposable elements in genome sequences. PLoS Comput Biol 2005, 1:166–75.

Quinlan AR, Hall IM: BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 2010, 26:841–842.

Rastogi A, Gupta D: GFF-Ex: a genome feature extraction package. BMC Res Notes 2014, 7:315.

Rebollo R, Horard B, Hubert B, Vieira C: Jumping genes and epigenetics: Towards new species. Gene 2010, 454:1–7.

Rebollo R, Romanish MT, Mager DL: Transposable Elements: An Abundant and Natural Source of Regulatory Sequences for Host Genes. Annu Rev Genet 2011, 46:120913153128008.

RepeatMasker Open-4.0 [http://www.repeatmasker.org]

Rho M, Choi J-H, Kim S, Lynch M, Tang H: De novo identification of LTR retrotransposons in eukaryotic genomes. BMC Genomics 2007, 8:90.

Romanov MN, Farré M, Lithgow PE, Fowler KE, Skinner BM, O'Connor R, Fonseka G, Backström N, Matsuda Y, Nishida C, Houde P, Jarvis ED, Ellegren H, Burt DW, Larkin DM, Griffin DK: Reconstruction of gross avian genome structure, organization and evolution suggests that the chicken lineage most closely resembles the dinosaur avian ancestor. BMC Genomics 2014, 15:1060.

Saha S, Bridges S, Magbanua Z V, Peterson DG: Empirical comparison of ab initio repeat finding programs. Nucleic Acids Res 2008, 36:2284–94.

Sanger, F., et al. DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences 74, 5463–5467 1977

Santiago N, Herráiz C, Goñi JR, Messeguer X, Casacuberta JM: Genome-wide analysis of the Emigrant family of MITEs of Arabidopsis thaliana. Mol Biol Evol 2002, 19:2285–93.

Schmid M, Smith J, Burt DW, Aken BL, Antin PB, et al. Third Report on Chicken Genes and Chromosomes 2015; Cytogenet Genome Res. 2015;145:78-179.

Schmutz J, Martin J, Terry a, Couronne O, Grimwood J, Lowry S, Gordon L a, Scott D, Xie G, Huang W, Hellsten U, Tran-Gyamfi M, She X, Prabhakar S, Aerts a, Altherr M, Bajorek E, Black S, Branscomb E, Caoile C, Challacombe JF, Chan YM, Denys M, Detter JC, Escobar J, Flowers D, Fotopulos D, Glavina T, Gomez M, Gonzales E, et al.: The DNA sequence and comparative analysis of human chromosome 5. Nature 2004, 431(September):268–274.

Schwartz A, Chan DC, Brown LG, Alagappan R, Pettay D, Disteche C, McGillivray B, De La Chapelle A, Page DC: Reconstructing hominid Y evolution: X-homologous block, created by X-Y transposition, was disrupted by Yp inversion through LINE-LINE recombination. Hum Mol Genet 1998, 7:1–11.

Seidl AH, Sanchez JT, Schecterson L, Tabor KM, Wang Y, Kashima DT, Poynter G, Huss D, Fraser SE, Lansford R, Rubel EW: Transgenic quail as a model for research in the avian nervous system: A comparative study of the auditory brainstem. J Comp Neurol 2013, 521:5–23.

Shang WH, Hori T, Toyoda A, Kato J, Popendorf K, et al. Chickens possess centromeres with both extended tandem repeats and short non-tandem-repetitive sequences. Genome Res. 2010;20:1219-1228.

Smith J, Bruley CK, Paton IR, Dunn I, Jones CT, Windsor D, Morrice DR, Law a S, Masabanda J, Sazanov A, Waddington D, Fries R, Burt DW: Differences in gene density on chicken macrochromosomes and microchromosomes. Anim Genet 2000, 31:96–103.

St John J, Quinn TW: Identification of novel CR1 subfamilies in an avian order with recently active elements. Mol Phylogenet Evol 2008, 49:1008–14.

Stijn van Dongen, *Graph Clustering by Flow Simulation*. PhD thesis, University of Utrecht, May 2000.

Studer A, Zhao Q, Ross-Ibarra J, Doebley J: Identification of a functional transposon insertion in the maize domestication gene tb1. Nat Genet 2011, 43:1160–1163.

Szak ST, Pickeral OK, Makalowski W, Boguski MS, Landsman D, Boeke JD: Molecular archeology of L1 insertions in the human genome. Genome Biol 2002, 3:research0052.1–research0052.18.

Tempel S: Using and Understanding RepeatMasker. 2012, 859. [Methods in Molecular Biology]

Thompson JD, Higgins DG, Gibson TJ: CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 1994, 22:4673–4680.

Tiersch TR, Wachtel SS. On the evolution of genome size of birds. J Hered. 1991;82:363-368.

Tu Z: Eight novel families of miniature inverted repeat transposable elements in the African malaria mosquito, Anopheles gambiae. Proc Natl Acad Sci U S A 2001, 98:1699–1704.

Völker M, Backström N, Skinner BM, Langley EJ, Bunzey SK, et al. Copy number variation, chromosome rearrangement, and their association with recombination during avian evolution. Genome Res. 2010;20:503-511.

Walker EL, Robbins TP, Bureau TE, Kermicle J, Dellaporta SL: Transposon-mediated chromosomal rearrangements and gene duplications in the formation of the maize R-r complex. EMBO J 1995, 14:2350–2363.

Walters RD, Kugel JF, Goodrich J a: InvAluable junk: the cellular impact and function of Alu and B2 RNAs. IUBMB Life 2009, 61:831–7.

Wicker T, Guyot R, Yahiaoui N, Keller B: CACTA transposons in Triticeae. A diverse family of high-copy repetitive elements. Plant Physiol 2003, 132:52–63.

Wicker T, Robertson JS, Schulze SR, Feltus FA, Magrini V, Morrison J a, Mardis ER, Wilson RK, Peterson DG, Paterson AH, Ivarie R: The repetitive landscape of the chicken genome. Genome Res 2005, 15:126–36.

Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, Paux E, SanMiguel P, Schulman AH: A unified classification system for eukaryotic transposable elements. Nat Rev Genet 2007, 8:973–82.

Wong LH, Choo KHA: Evolutionary dynamics of transposable elements at the centromere. Trends Genet 2004, 20:611–616.

Xu Z, Wang H: LTR-FINDER: An efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res 2007, 35:265–268.

Ye L, Hillier LW, Minx P, Thane N, Locke DP, Martin JC, et al. A vertebrate case study of the quality of assemblies derived from next-generation sequences. Genome Biol. 2011;12:R31.

Yi G, Qu L, Liu J, Yan Y, Xu G, Yang N. Genome-wide patterns of copy number variation in the diversified chicken genomes using next-generation sequencing. BMC Genomics. 2014;15:962.

Yuri T, Kimball RT, Harshman J, Bowie RCK, Braun MJ, Chojnowski JL, Han K-L, Hackett SJ, Huddleston CJ, Moore WS, Reddy S, Sheldon FH, Steadman DW, Witt CC, Braun EL: Parsimony and model-based analyses of indels in avian nuclear genes reveal congruent and incongruent phylogenetic signals. Biology (Basel) 2013, 2:419–444.

Zhang G, Li C, Li Q, Li B, Larkin DM, Lee C, Storz JF, Antunes A, Greenwold MJ, Meredith RW, Odeen A, Cui J, Zhou Q, Xu L, Pan H, Wang Z, Jin L, Zhang P, Hu H, Yang W, Hu J, Xiao J, Yang Z, Liu Y, Xie Q, Yu H, Lian J, Wen P, Zhang F, Li H, et al.: Comparative genomics reveals insights into avian genome evolution and adaptation. Science (80-) 2014, 346:1311–1320.

Zhang Q, Backström N. Assembly errors cause false tandem duplicate regions in the chicken (Gallus gallus) genome sequence. Chromosoma. 2014;123:165-168.

Zhang Q, Edwards S V.: The evolution of intron size in amniotes: A role for powered flight? Genome Biol Evol 2012, 4:1033–1043.

Annexes

Annexe 1 : Espèces aviaires séquencées par le phylogenomic project

Liste des espèces aviaires séquencées.

Espèce	Nom Commun	Nom Commun (Anglais)	Ordre	Niveau Assemblage	Date de Publication (aaaa/mm/jj)	ID ncbi	Phylogenomic Project (GigaScience accession)
Gallus gallus	Coq bankiva	Chicken	Galliformes	Chromosome	2004/02/29	111	
Taeniopygia guttata	Diamant mandarin	Zebra finch	Passeriformes	Chromosome	2008/09/03	367	
Meleagris gallopavo	Dindon sauvage	Turkey	Galliformes	Chromosome	2010/09/07	112	
Melopsittacus undulatus	Perruche ondulée	Budgerigar	Psittaciformes	Scaffold	2011/11/28	10765	doi:10.5524/100059
Geospiza fortis	Géospize à bec moyen	Medium Ground-finch	Passeriformes	Scaffold	2012/07/11	13302	doi:10.5524/100040
Ficedula albicollis	Gobe-mouche à collier	Collared flycatcher	Passeriformes	Chromosome	2012/11/30	11872	
Pseudopodoces humilis	Mésange de Hume	Ground tit	Passeriformes	Scaffold	2013/01/09	15114	
Amazona vittata	Amazone de Porto Rico	Puerto Rican parrot	Psittaciformes	Scaffold	2013/01/18	15170	
Columba livia	Pigeon biset	Pigeon	Columbiformes	Scaffold	2013/02/04	10719	doi:10.5524/100007
Falco cherrug	Faucon sacre	Saker falcon	Falconiformes	Scaffold	2013/02/05	14103	
Falco peregrinus	Faucon pèlerin	Peregrine Falcon	Falconiformes	Scaffold	2013/02/05	132	doi:10.5524/101006
Anas platyrhynchos	Canard colvert/mallard	Beijing Duck (Mallard)	Anseriformes	Scaffold	2013/04/02	2793	doi:10.5524/101001
Zonotrichia albicollis	Bruant à gorge blanche	White-throated sparrow	Passeriformes	Scaffold	2013/04/26	17510	
Ara macao	Ara rouge	Scarlet macaw	Psittaciformes	Scaffold	2013/05/20	17770	
Coturnix japonica	Caille du Japon	Japanese quail	Galliformes	Chromosome	2013/07/09	113	
Serinus canaria	Serin des Canaries	Atlantic canary	Passeriformes	Scaffold	2014/01/06	11251	
Lyrurus tetrix	Tétras lyre	Black grouse	Galliformes	Scaffold	2014/03/08	31278	
Colinus virginianus	Colin de Virginie	Northern bobwhite	Galliformes	Scaffold	2014/03/20	12817	
Egretta garzetta	Aigrette garzette	Little Egret	Pelecaniformes	Scaffold	2014/04/29	31706	doi:10.5524/101002
Phaethon lepturus	Phaéton à bec jaune	White-tailed Tropicbird	Phaethontiformes	Scaffold	2014/04/30	31929	doi:10.5524/101033
Phoenicopterus ruber ruber	Flamant des Caraïbes	American Flamingo	Phoenicopteriformes	Scaffold	2014/04/30	31928	doi:10.5524/101035

Tyto alba	Chouette effraie	Barn Owl	Strigiformes	Scaffold	2014/04/30	31927	doi:10.5524/101039
Merops nubicus	Guêpier écarlate	Carmine Bee-eater	Coraciiformes	Scaffold	2014/05/01	31978	doi:10.5524/101029
Mesitornis unicolor	Mésite unicolore	Brown Mesite	Mesitornithiformes	Scaffold	2014/05/01	32034	doi:10.5524/101030
Nestor notabilis	nestor kéa	Kea	Psittaciformes	Scaffold	2014/05/01	33272	doi:10.5524/101031
Pelecanus crispus	Pélican frisé	Dalmatian Pelican	Pelecaniformes	Scaffold	2014/05/01	31933	doi:10.5524/101032
Haliaeetus albicilla	Pygargue à queue blanche	White-tailed Eagle	Accipitriformes	Scaffold	2014/05/02	31975	doi:10.5524/101027
Leptosomus discolor	Courol vouroudriou	Cuckoo-roller	Leptosomiformes	Scaffold	2014/05/02	32000	doi:10.5524/101028
Fulmarus glacialis	Fulmar boréal	Northern Fulmar	Procellariiformes	Scaffold	2014/05/03	31971	doi:10.5524/101025
Gavia stellata	Plongeon catmarin	Red-throated Loon	Gaviiformes	Scaffold	2014/05/03	7634	doi:10.5524/101026
Cariama cristata	Cariama huppé	Red-legged Seriema	Cariamiformes	Scaffold	2014/05/05	31967	doi:10.5524/101020
Colius striatus	Coliou rayé	Speckled Mousebird	Coliiformes	Scaffold	2014/05/05	31969	doi:10.5524/101023
Eurypyga helias	Caurale soleil	Sunbittern	Eurypygiformes	Scaffold	2014/05/05	31970	doi:10.5524/101024
Acanthisitta chloris	Xénique grimpeur	Rifleman	Passeriformes	Scaffold	2014/05/06	32002	doi:10.5524/101015
Calypte anna	Colibri d'Anna	Anna's Hummingbird	Apodiformes	Scaffold	2014/05/06	32060	doi:10.5524/101004
Picoides pubescens	Pic mineur	Downy Woodpecker	Piciformes	Scaffold	2014/05/06	32059	doi:10.5524/101012
Struthio camelus australis	Autruche	Common Ostrich	Struthioniformes	Scaffold	2014/05/06	122	doi:10.5524/101013
Aptenodytes forsteri	Manchot empereur	Emperor Penguin	Sphenisciformes	Scaffold	2014/05/07	32061	doi:10.5524/100005
Podiceps cristatus	Grèbe huppé	Great-crested Grebe	Podicipediformes	Scaffold	2014/05/07	7975	doi:10.5524/101036
Pterocles gutturalis	Ganga à gorge jaune	Yellow-throated Sandgrouse	Pterocliformes	Scaffold	2014/05/07	32063	doi:10.5524/101037
Pygoscelis adeliae	Manchot Adélie	Adelie Penguin	Sphenisciformes	Scaffold	2014/05/07	17559	doi:10.5524/100006
Apaloderma vittatum	Trogon à queue barrée	Bar-tailed Trogon	Trogoniformes	Scaffold	2014/05/08	32169	doi:10.5524/101016
Caprimulgus carolinensis/ Antrostomus carolinensis	Engoulevent de Caroline	Chuck-will's-widow	Caprimulgiformes	Scaffold	2014/05/08	32067	doi:10.5524/101019
Cathartes aura	Urubu à tête rouge	Turkey Vulture	Accipitriformes	Scaffold	2014/05/08	7839	doi:10.5524/101021
Tinamus guttatus	Tinamou à gorge blanche	White-throated Tinamou	Tinamiformes	Scaffold	2014/05/08	32250	doi:10.5524/101014
Charadrius vociferus	Pluvier kildir	Killdeer	Charadriiformes	Scaffold	2014/05/09	32124	doi:10.5524/101007
Nipponia nippon	Ibis nippon	Crested Ibis	Pelecaniformes	Scaffold	2014/05/09	7967	doi:10.5524/101003
Chlamydotis macqueenii	Outarde de Macqueen	MacQueen's Bustard	Otidiformes	Scaffold	2014/05/12	34357	doi:10.5524/101022

Phalacrocorax carbo	Grand Cormoran	Great Cormorant	Suliformes	Scaffold	2014/05/12	32171	doi:10.5524/101034
Corvus brachyrhynchos	Corneille d'Amérique	American Crow	Passeriformes	Scaffold	2014/05/13	32035	doi:10.5524/101008
Manacus vitellinus	Manakin à col d'or	Golden-collared Manakin	Passeriformes	Scaffold	2014/05/13	31979	doi:10.5524/101010
Opisthocomus hoazin	Hoazin huppé	Hoatzin	Opisthocomiformes	Scaffold	2014/05/13	31992	doi:10.5524/101011
Aquila chrysaetos	Aigle royal	Golden eagle	Accipitriformes	Scaffold	2014/05/23	32031	
Balearica regulorum gibbericeps	Balearica regulorum	Grey-crowned Crane	Gruiformes	Scaffold	2014/06/02	17144	doi:10.5524/101017
Buceros rhinoceros silvestris	Calao rhinocéros	Rhinoceros Hornbill	Bucerotiformes	Scaffold	2014/06/02	32403	doi:10.5524/101018
Cuculus canorus	Coucou gris	Common Cuckoo	Cuculiformes	Scaffold	2014/06/12	32170	doi:10.5524/101009
Tauraco erythrolophus	Touraco de Pauline	Red-crested Turaco	Musophagiformes	Scaffold	2014/06/12	32247	doi:10.5524/101038
Corvus cornix	Corneille mantelée	Hooded crow	Passeriformes	Scaffold	2014/08/04	18230	
Haliaeetus leucocephalus	Pygargue à tête blanche	Bald Eagle	Accipitriformes	Scaffold	2014/08/04	32665	doi:10.5524/101040
Chaetura pelagica	Martinet ramoneur	Chimney Swift	Apodiformes	Scaffold	2014/09/04	33278	doi:10.5524/101005
Anser cygnoides	Oie cygnoïde	Swan goose	Anseriformes	Scaffold	2015/04/08	31397	
Zosterops lateralis	Zostérops à dos gris	Silvereye	Passeriformes	Scaffold	2015/09/14	40104	
Amazona aestiva	Amazone à front bleu	Blue-fronted parrot	Psittaciformes	Scaffold	2015/10/30	40915	
Calidris pugnax	Combattant varié	Ruff	Charadriiformes	Scaffold	2015/11/04	41082	
Sturnus vulgaris	Étourneau sansonnet	Common starling	Passeriformes	Scaffold	2015/11/20	41647	
Parus major	Mésange charbonnière	Great tit	Passeriformes	Chromosome	2016/01/20	12863	

Les lignes grises indiquent les espèces séquencées par le phylogenomic project.

Annexe 2 : DensityMap - Additional File 1 : Algorithmes

Main

1. Check mandatory options
2. Read optional options and set default values if needed
3. Image size computation
    1. Height computation
        1. For each GFF files
            1. Increment number of chromosome to plot
            2. Check the first line for gff version
            3. Check the second line for chromosome length
            4. Store the highest chromosome length
        2. if auto_scale_factor
            1. while picture height > max picture height
                1. compute picture height
                2. end loop if condition is satisfied
                3. multiply scale factor by 10
        3. Store picture height
    2. Width computation
        1. For each couple of type=value
            1. define number of strand to plot
            2. store total number of strand to plot
        2. Compute and store picture width
4. Ask user if picture dimensions are fine
5. Create pictures
6. Load colours and colours schemes from colours.txt
7. if background is set draw background
8. if title add title to the picture
9. if show_scale add scale to the picture (drawScale)
10. Prepare data structure to store intervals for each strands
11. Draw strands of all chromosomes: For each GFF
    1. Open GFF file and read chromosome length

2. Clean data structure of previous intervals
3. For each chromosome of the GFF file
    1. skip line if current line type is not in user's type list
    2. if current line type = centromere, Store centromere positions
    3. store intervals in data structure

12. if label_strand_rotation Apply rotation to labels
13. Remove stroke of each window
14. Save picture

ProcessData

1. For each type
    1. for each strand
        1. sort intervals
2. For each type
    1. for each strand
        1. Reduce intervals (removeIntervalRedundancy)
3. For each type
    1. for each strand
        1. set position of first window in picture
4. For each type
    1. for each strand
        1. draw all windows in picture (drawPixels)

removeIntervalRedundancy

1. For each intervals collection
    1. Load first interval (Once)
    2. if current interval start < previous interval end
        1. replace previous interval end by current interval end
    3. else if current interval end < previous interval end
        1. remove interval
    4. else (current interval start > previous interval end)
        1. set current intervals as previous interval
2. Return reduce intervals collection

drawScale

1. Search the max number of scale ticks, start with 10 bases by ticks
    1. while number of scale ticks > max number of scale ticks
        1. compute number of ticks
        2. end loop if condition is satisfied
        3. multiply number of base by tick by 10
2. Add scale unit to picture
3. Add number of base by window to picture
4. Draw scale to picture
5. Draw last tick to picture
6. Draw other ticks to picture

drawPixels

1. Get unique id for svg strand group
2. open svg strand group
3. For each window
    1. load number of bases covered by previous interval (interval spanning several window(s), interval size > window size)
    2. while end of interval collection is not reached and interval is in current window
        1. if interval is in current window
            1. increment number of covered bases by interval size
        2. if interval is across current and next window (and more)
            1. compute number of windows covered by the intervals
            2. increment number of covered bases for the current window
            3. For each windows fully covered by interval
                1. store covered bases for each window
            4. Store the last window number of bases covered by the interval
    3. Compute percentage of window covered bases (=density)
    4. Draw window to the picture with colour corresponding to the density
4. Draw centromere to picture
5. Draw sequence name label
6. Draw type name label
7. Draw strand name label
8. Close strand group

Annexe 3 : DensityMap - Additional File 2 : Manuel

DensityMap

1. Purpose

The visualization of genomic data is a considerable challenge. Many tools, such as Gbrowse, Jbrowse, and Abrowse, are available for displaying the data for small genomic loci, but few have attempted to visualize whole chromosomes or genomes. Phenogram and CviT offer two solutions to this problem but each has its own limitations. We needed a program suitable for treating repeated sequences, which can be very numerous in some genomes (45 % in the human genome, over 90 % in the wheat genome). We therefore designed DensityMap, which can represent the density (number of base pairs covered) of one or more features along chromosomes using a series of windows. The features correspond to the third column of the GFF annotation files (format: http://www.sequenceontology.org/gff3.shtml) used as inputs. The program generates high quality SVG pictures that can be edited.

This program makes it possible to represent the contents of GFF files simply. It allows the user to select the data to be plotted, provides automatic chromosome scaling, and an output picture that is easily configured using the many graphics options provided.

2. Installation

2.1 Requirements

DensityMap is in Perl script that needs only one requirement to work, GD::SVG (tested with version 0.33-1). GD::SVG libraries can be installed on a Debian-based system using the apt-get package manager:

sudo apt-get install libgd-svg-perl

It also use POSIX libraries that are included in Linux distributions for ceiling and floor functions. Term::ANSIColor libraries can be installed as options for more colourful output. Download the Perl module from cpan:

https://metacpan.org/pod/Term::ANSIColor.

Then install it following the standard Perl module installation procedure:

```
perlMakefile.PL
make
make test
sudo make install
```

Activate the colour terminal output by uncommenting lines 16 and 17 in DensityMap.pl script.

2.2 DensityMap Installation

Download the DensityMap archive or clone repository from github:

https://github.com/sguizard/DensityMap.

The program can be launched directly from the unzipped archive or you can execute the install script for an all-users installation:

```
sudo install.sh
```

It will create a directory DensityMap in /usr/local/share and copy all files in it. It then updates the path of colours.txt in DensityMap.pl and creates a symbolic link to /usr/local/bin.

3. Use

Mandatory options

- -i, -input
  - string
  - Name of Gff file(s)
- -o , -output_img_name
  - string
  - Name of the output image, this extension .svg will be automatically added
- -ty, -type_to_draw
  - string
  - List of type to draw, strand to plot and colour scale to use
  - Format: "Type1=strand;Type2=strand=8"
    - Type (third column of GFF): match, gene, CDS, etc.
    - Strand:
      - - → strand -
      - + → strand +
      - both → strand - and strand +
      - fused → Combination of strand - and strand +
      - all → strand - and strand + and fused

Generic options

- -v, -verbose: more text explanation
- -h, -help: This help
- -for, -force: Automatically answers yes to picture size validation

Density options

- -c, -colour_scale:
  - integer
  - colour scale to use
  - (Default = 7)
- -sc, -scale_factor:
  - integer
  - window length in bp
  - (Default = 1000)
- -a, -auto_scale_factor:
  - integer
  - Max picture height in pixel
- -ro, -rounding_method:
  - string
  - floor or ceil
  - (Default = floor)
- -gc:
  - integer
  - colour scale to use
  - Create a density map of the GC% along the chromosome, REQUIRE the presence of the fasta sequence in the ##FASTA section of the GFF file

Graphical options

- -ti, -title
  - string
  - Title to print on the picture
- -w, -win_size
  - integer
  - Height of window in pixel
  - Default: 1
- -sh, -show_scale
  - integer
  - Draw Scale, n = num max ticks
  - Default: 50
- -str_w, -str_width
  - integer
  - Strand width in pixel
  - Default: 50
- -str_s, -str_space
  - integer
  - Space between strands in pixels
  - Default: 50
- -sp, -space_chr
  - integer
  - space between chromosomes in pixels
  - default: 50
- -lm, -lmargin
  - integer
  - left margin in pixels

    - default: 50
- -rm, -rmargin
    - integer
    - right margin in pixels
    - default: 50
- -tm, -tmargin
    - integer
    - top margin in pixels
    - default: 50
- -bm, -bmargin
    - integer
    - bottom margin in pixels
    - default: 50
- -ba, background
    - color
    - fill background
    - default: no background
- -la, label_strand_rotation
    - integer
    - rotation degree of strand label
    - default: 0
- ft_f, ft_family:
    - string
    - font to use for text
    - default: “Helvetica”
- ft_s, ft_size:
    - integer

- font size
- default: 16

Running the program:

The GFF set as the input will only be valid if it respects the format defined in: http://www.sequenceontology.org/gff3.shtml. It also must contain the header sequence-region for each chromosome.

Example of valid file:

```
##sequence-region 2L 1 23513712
2L   RefSeq    gene 7529 9484 .    +    .    ID=gene2671
2L   RefSeq    gene 9839 21376     .    -    .    ID=gene2672
…
2L   RM   LTR  23512506  23512653  809  +    .
     Target=DM297_I-int 3285 3432
##sequence-region 2R 1 25286936
2R   RefSeq    gene 432219    705848    .    +    .
     ID=gene6156
…
##FASTA
>seq1
ALONGSEQ...
```

The mandatory options must be set to create a basic image of your data:

```
DensityMap.pl -i 2R.gff3 -ty "LTR=fused" -o 2R
```

This program is also executed with default graphical options (see above list) and will compute the size of the output picture and ask you if you want to continue or stop the execution. As you can see, the picture produced will be over 25,000 pixels high.

Two options are available for producing a reasonable picture height. You can define the window size to use to compute the density of feature using the -sc option. Or you can let the program choose for you and define a maximum picture height with the option -a.

You can try using the same command with the option -a 3000 to obtain a picture whose height does not exceed 3000 pixels and add the -ba white option to obtain a white background.

```
DensityMap.pl -i 2R.gff3 -ty "LTR=fused" -o 2R -a 3000 -ba white
```

The program will ask for the window size (scale factor) and ask you to validate the picture height. The resulting picture is 150 px by 2653 px, which is easily viewed (Figure 1). The default colour scale is blue to red: nucleotide windows with low-density features (LTR) to nucleotide windows with high-density features.

Several chromosomes can be visualized by pooling the GFF files on the -g option. Information for different types can also be displayed by defining them in the -ty option. The graphical options you can be used to add a scale and a title to the picture (Figure 2).

```
DensityMap.pl -i dmel.gff3 -o dmel -ty "LTR=fused;LINE=fused" -ba white -sc 20000 -sh 100 -title "LTR and LINE retrotransposon in Dmel genome"
```

This gives you a density map of the whole genome of *Drosophilia Melanogaster* describing the density of LINE and LTR retrotransposons. As you can see, the density of LTR retrotransposons is higher than that of LINE, but both TEs display a similar distribution with high concentrations at the ends of each chromosome and large areas devoid of transposable elements, except for chromosomes 4 and X which are almost devoid of transposable elements.

DensityMap can be also used to study low density features like rolling circle transposons. You can use another colour scale, like number 9. It is designed to set a blank pixel for windows with a density of 0 - 1 % and a red pixel for windows with a density over 1 %. You also modify the way the density is rounded. Densities are rounded down using floor method by

default, so a density of < 1 % is rounded to 0 %. Setting the rounding method (-ro) to ceiling rounds the value up to 1 %. You should use the ceil method if you need to maximize the visibility of low density features (Figure 3).

```
DensityMap.pl -i dmel.gff3 -o RC -ty "RC=fused" -ba white -sc 20000 -sh 100 -title "Rolling Circle transposons in Dmel genome" -c 9 -ro ceil
```

If you know the positions of the centromeres on the chromosomes you can add their coordinates to the GFF file with the type centromere. They will then be displayed on chromosomes as shown in Figure 3 for chromosome 3R.

You can modify all the picture margins (bottom, top, left, right), the space between strands and the width of the strands.

Customizing the colour scheme:

DensityMap is delivered with 10 colour scales for representing densities. If you find no suitable colour scale for your representation, you can create your own by modifying the colours.txt installation directory. A colour scale is composed of 101 colours, and each contains four pieces of information:

- An id formatted: <ID>_heatmap<COLOURNUMBER>
- A Red value, 0 to 255
- A Green value, 0 to 255
- A Blue value, 0 to 255

These specifications are separated by semi-colons.

Example of colour scale:

1_heatmap10;0;200;0

1_heatmap11;0;195;0

…

1_heatmap99;245;0;0

1_heatmap100;250;0;0

The DensityMap archive contains a script that reads the colours.txt file and creates a miniature of all the colour scales available (Figure 4). You only need to execute the scaleColorDrawer.pl script in the colours.txt directory without any arguments.

Figure 1

Figure 2

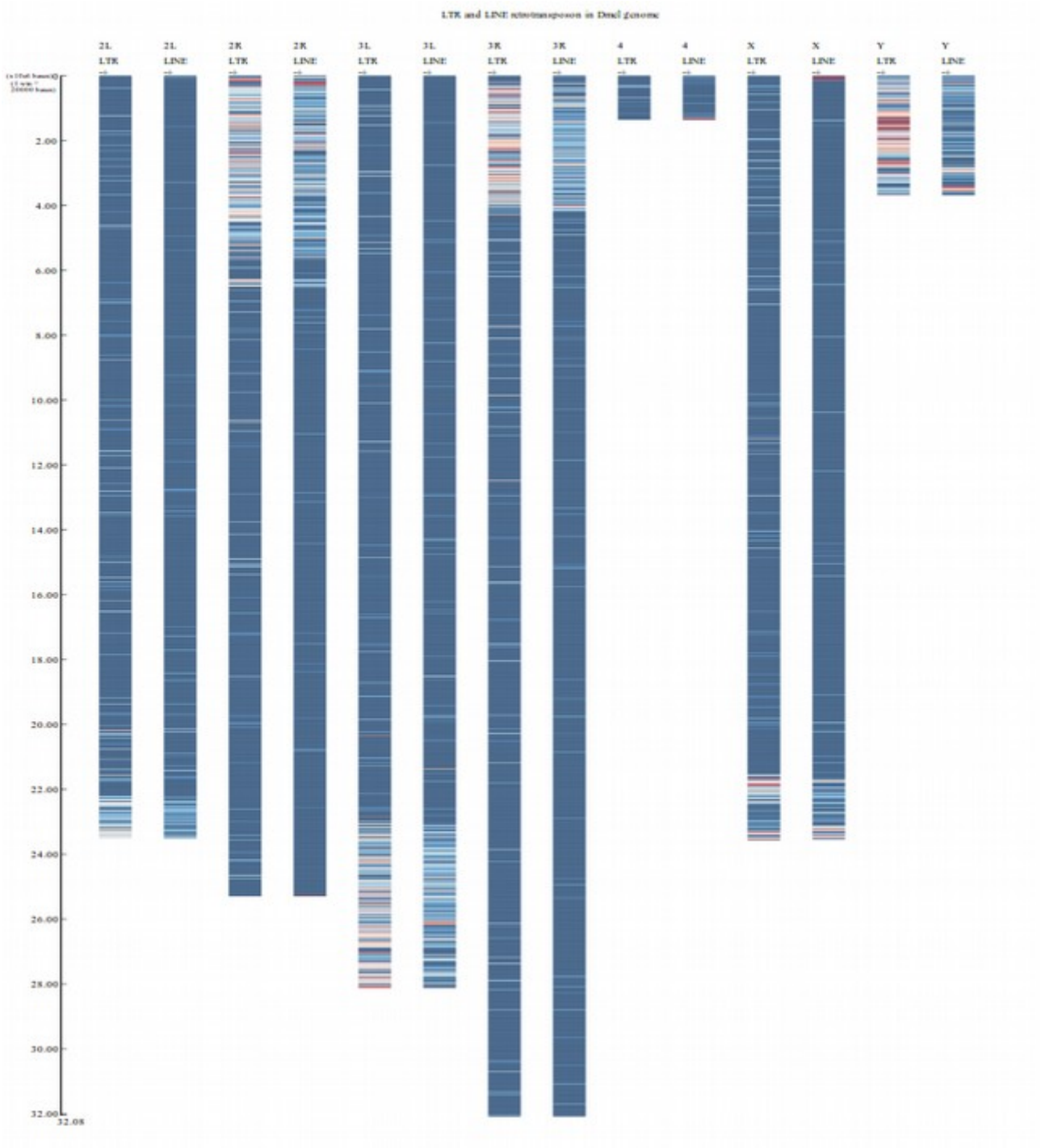

LTR and LINE retrotransposon in Dmel genome
2L LTR
2L LINE
2R LTR
2R LINE
3L LTR
3L LINE
3R LTR
3R LINE
4 LTR
4 LINE
X LTR
X LINE
Y LTR
Y LINE
2.00
4.00
6.00
8.00
10.00
12.00
14.00
16.00
18.00
20.00
22.00
24.00
26.00
28.00
30.00
32.00
32.08

Figure 3

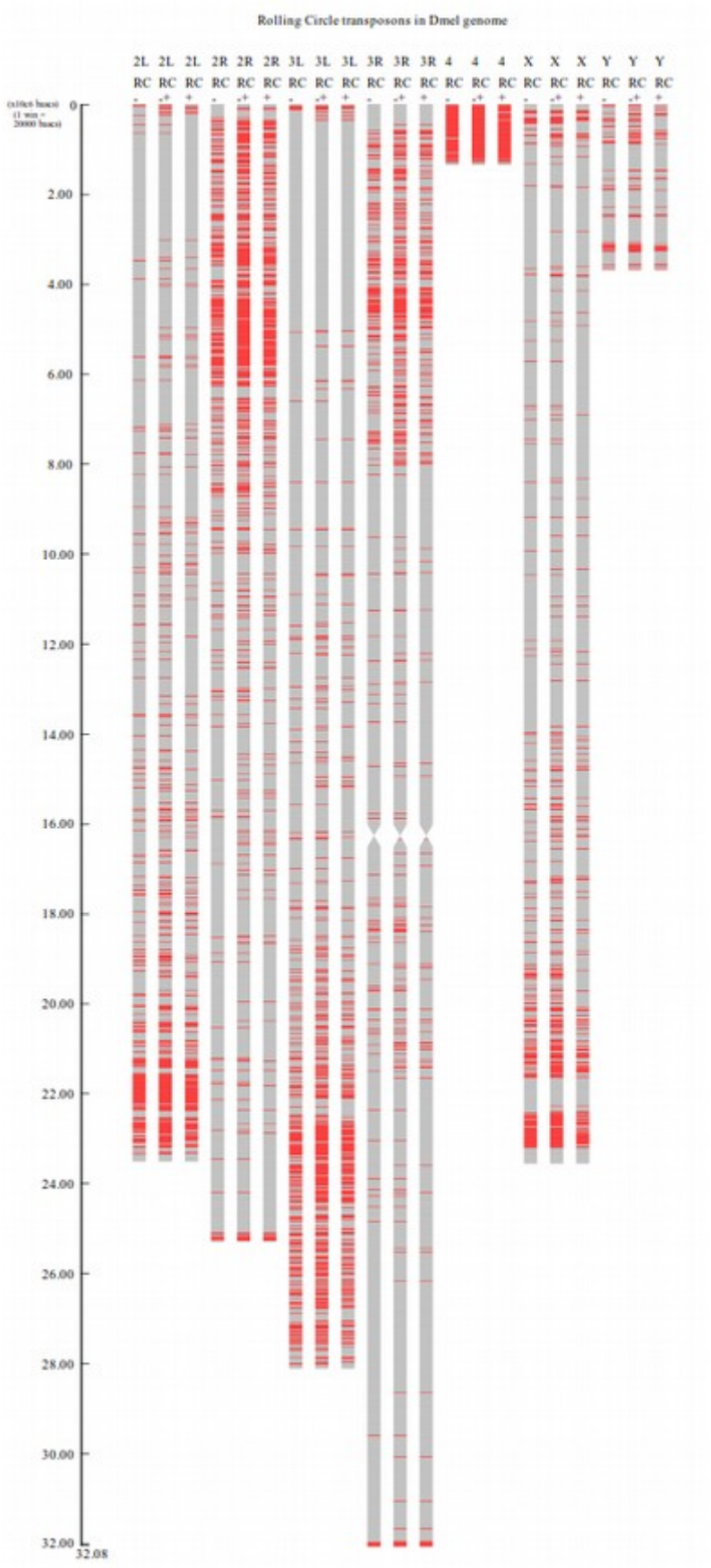
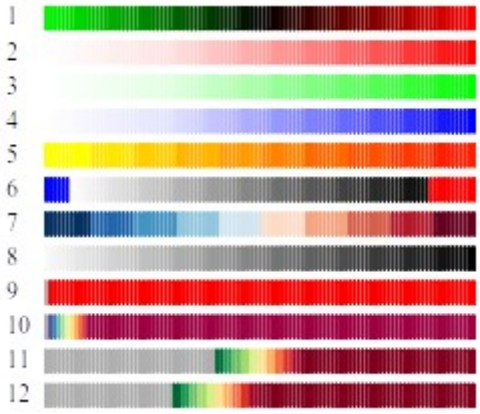

Rolling Circle transposons in Dmel genome
2L RC -
2L RC -,+
2L RC +
2R RC -
2R RC -,+
2R RC +
3L RC -
3L RC -,+
3L RC +
3R RC -
3R RC -,+
3R RC +
4 RC -
4 RC -,+
4 RC +
X RC -
X RC -,+
X RC +
Y RC -
Y RC -,+
Y RC +
0
2.00
4.00
6.00
8.00
10.00
12.00
14.00
16.00
18.00
20.00
22.00
24.00
26.00
28.00
30.00
32.00
32.08

Figure 4



Annexe 4 : Ré-annotation et re-découverte du modèle Galgal4 - Tableaux

Table 1. Amounts of repeated sequences found in chicken genome through done investigations

Method of investigation	% of moderately repeated and interspersed sequences	% of highly repeated sequences	Year of publication	References
Reassociation kinetic	20	10	1978	58
Reassociation kinetic	20	10	1978	59
ICGGC*	9.4	0.1	2004	28
Reassociation kinetic and sequencing	4.3	3 to 4	2005	60
ISB	9.74	1.73	2011	61

* International Chicken Genome Sequencing Consortium

Table 2. Percentages of SSRs found using ISB annotation or TRF in the Galgal4 model

Sequence type	RM	TRF	Increase factor
Assembled in chromosomes	1.54	3.73	2.42
Unassembled	5.67	12.24	2.16
Total in Galgal4	1.73	4.08	2.36

Table 3. Number and diversity of simple sequences repeats (SSRs) in Galgal4

SSRs type*		Number of arrays	Number of different repeated units[a]	% coverage in galGal4
Simple Repeat (stretches of A or T, and C or G)[b]		204434	2	PolyA : 0.355 PolyC : 0.022
Microsatellite [2-10] bp[b,c]		770202	2101	2.189
Minisatellite [11-60]bp[b,c]		12310	123	1.273
Tandem arrays [>60] bp[d]	Large tandem repeats	6	6	0.003
	Satellite DNAs	10136	9987	0.238

a, the threshold used to gather two repeated unit is a sequence similarity of 100 %; b, the minimal size for an array is 50 repeated units; c, between brackets are indicated the size of the repeated unit of each SSR type; d, the minimal size for an array is 2 to 50 repeated units of large tandem repeats et 51 to ∞ for a satellite DNA.

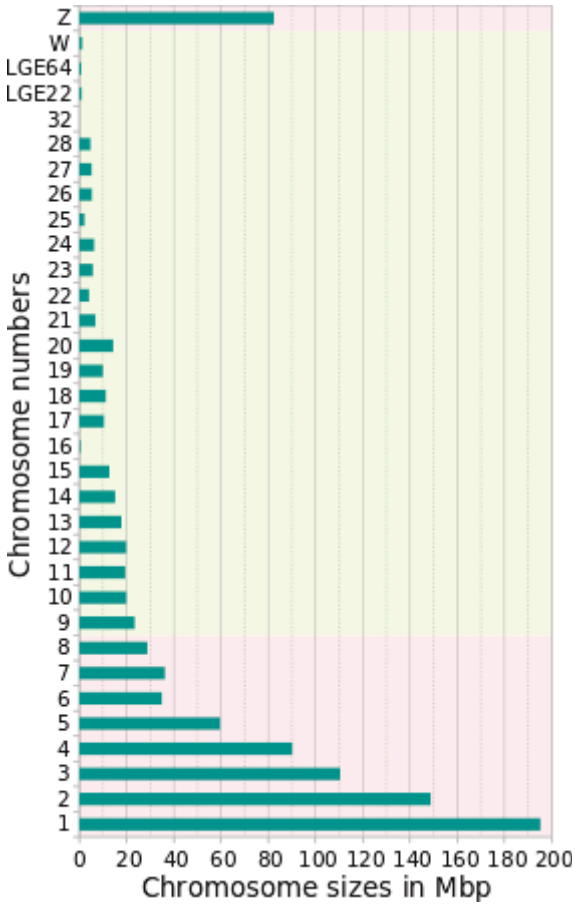
Table 4. Features and diversity of TE models found in the Galgal4 model based on the REPET and DM annotations (STEP3 + STEP4, Figure 2) after stack resolving and merging stacked and juxtaposed annotations.

Names of TE models	a	b	c*	d	e
CR1	308	LINE	413857	66.4707	11.8457
Ancestral_LTR_group_1	3	LTR	86	0.0138	0.0034
Ancestral_LTR_group_2	1	LTR	22	0.0035	0.0013
Ancestral_LTR_group_3	1	LTR	40	0.0064	0.0012
Ancestral_LTR_group_4	1	LTR	308	0.0495	0.0119
BIRDDAWG	10	LTR	6238	1.0019	0.2525
EAV	1	LTR	191	0.0307	0.0212
EAV-HP	7	LTR	765	0.1229	0.0496
ERV2	2	LTR	426	0.0684	0.0209
ERV7	10	LTR	2885	0.4634	0.1061
ERV11	1	LTR	512	0.0822	0.0168
Kronos	46	LTR	30732	4.9359	0.7377
putative_LTR_group4	2	LTR	835	0.1341	0.0137
putative_LTR_group9	1	LTR	170	0.0273	0.0017
putative_LTR_group12	17	LTR	1797	0.2886	0.05
putative_LTR_group22	3	LTR	1219	0.1958	0.0257
putative_LTR_group28	2	LTR	367	0.0589	0.0116
putative_LTR_group30	13	LTR	3847	0.6179	0.0996
retroCalimero	1	LTR	826	0.1327	0.0540
retroSaturnin	1	LTR	161	0.0259	0.0118
retroTux	2	LTR	2490	0.3999	0.1243
Soprano	19	LTR	3014	0.4841	0.1171
Charlie	3	TIR	37319	5.9939	0.5868
Charlie-Galluhop	5	TIR	67691	10.872	1.0296
Galluhop	2	TIR	4588	0.7369	0.1198
Mariner1_GG	10	TIR	5686	0.9132	0.1491
Hitchcock	4	undefined	27033	4.3418	0.4182
undetermined_group_1	3	undefined	2219	0.3564	0.0773
undetermined_group_2	2	undefined	1030	0.1654	0.0165

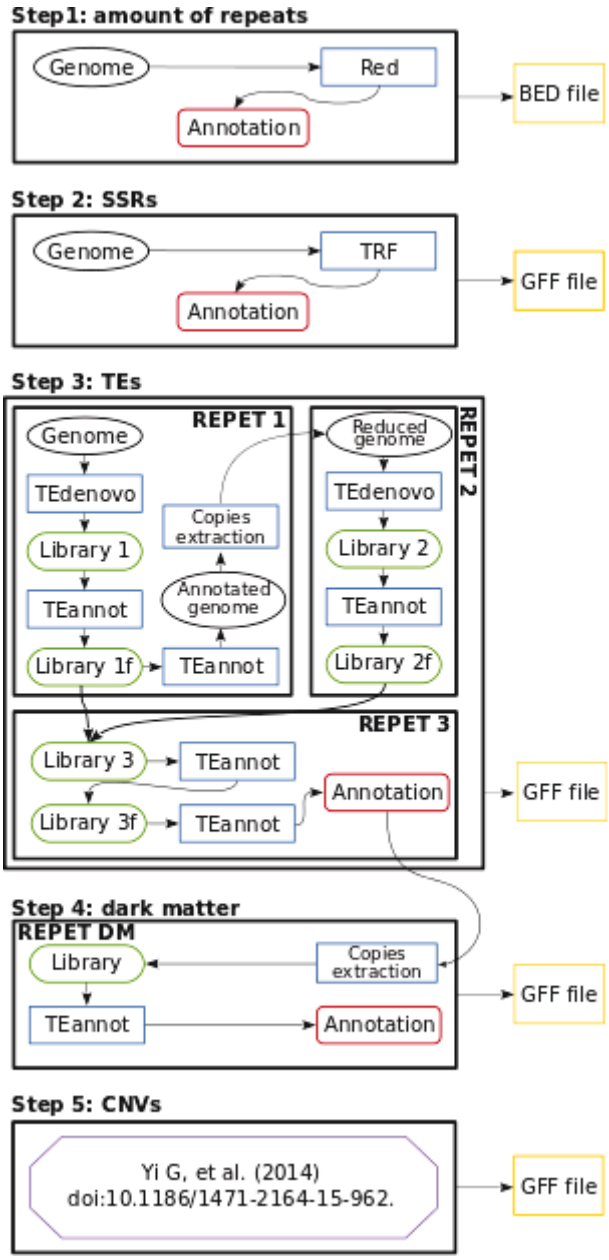
undetermined_group_3	2	undefined	174	0.0279	0.0045
undetermined_group_4	4	undefined	2550	0.4096	0.0423
undetermined_group_5	2	undefined	134	0.0215	0.0036
undetermined_group_6	1	undefined	372	0.0597	0.0100
Z_rep	9	undefined	3032	0.487	0.1476
Total	499		622616	100	16.1832**

a, Number of consensus ; b, TE types ; c, Total number of TE copies ; d, Percentage of the total number of TE copies ; e, Percentage of chromosome coverage; *, Post stack resolving and annotation merging are called copies all complete elements, internally deleted elements; 5' or 3' truncated elements and elements truncated at both ends (i.e. internal regions of a TE devoid of ends). **, this coverage value was more elevated than the 15.7% indicated in the main text because the coverage corresponding to the small TE copies nested in larger TEs were not removed for these calculations.

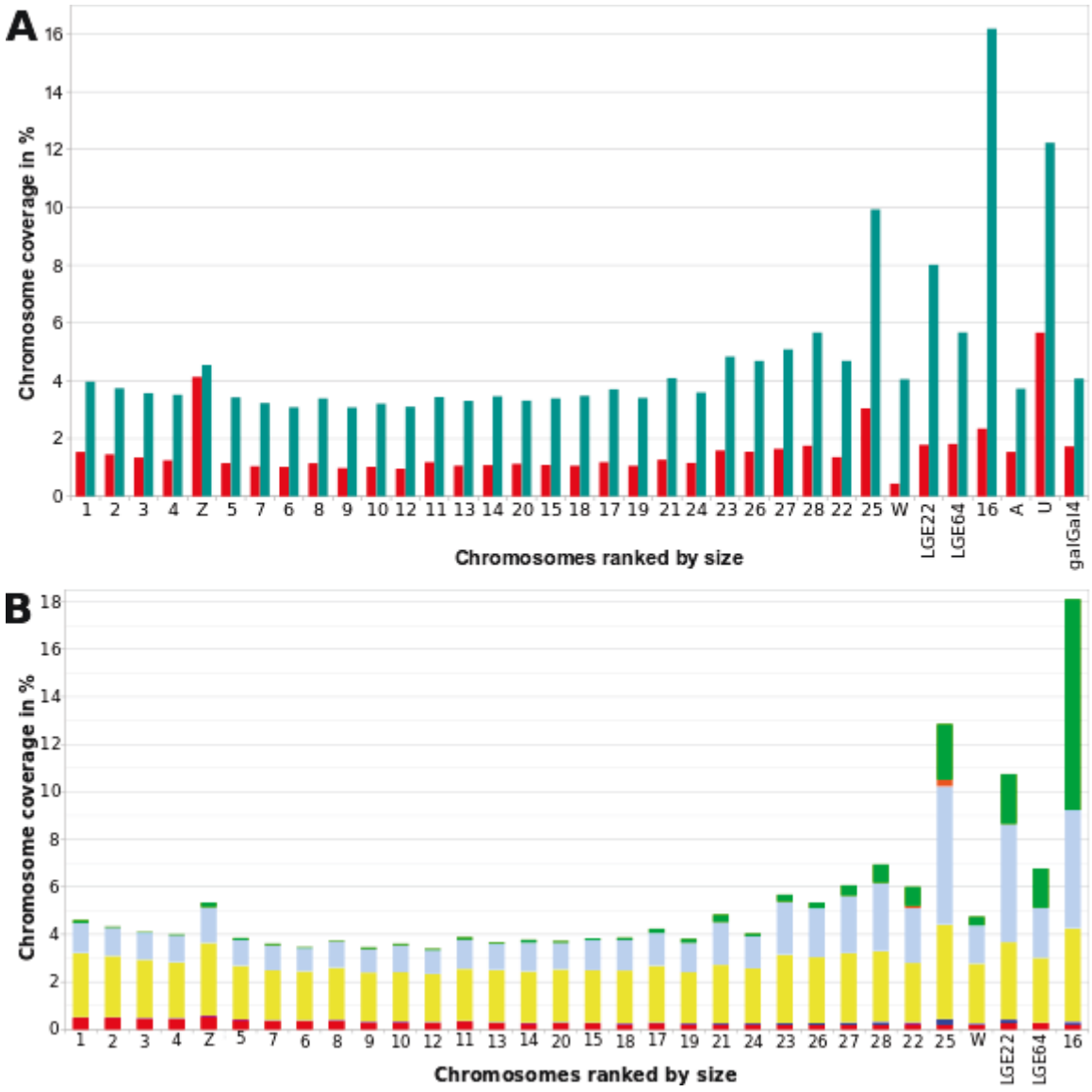
Annexe 5 : Ré-annotation et re-découverte du modèle Galgal4 - Figure 1



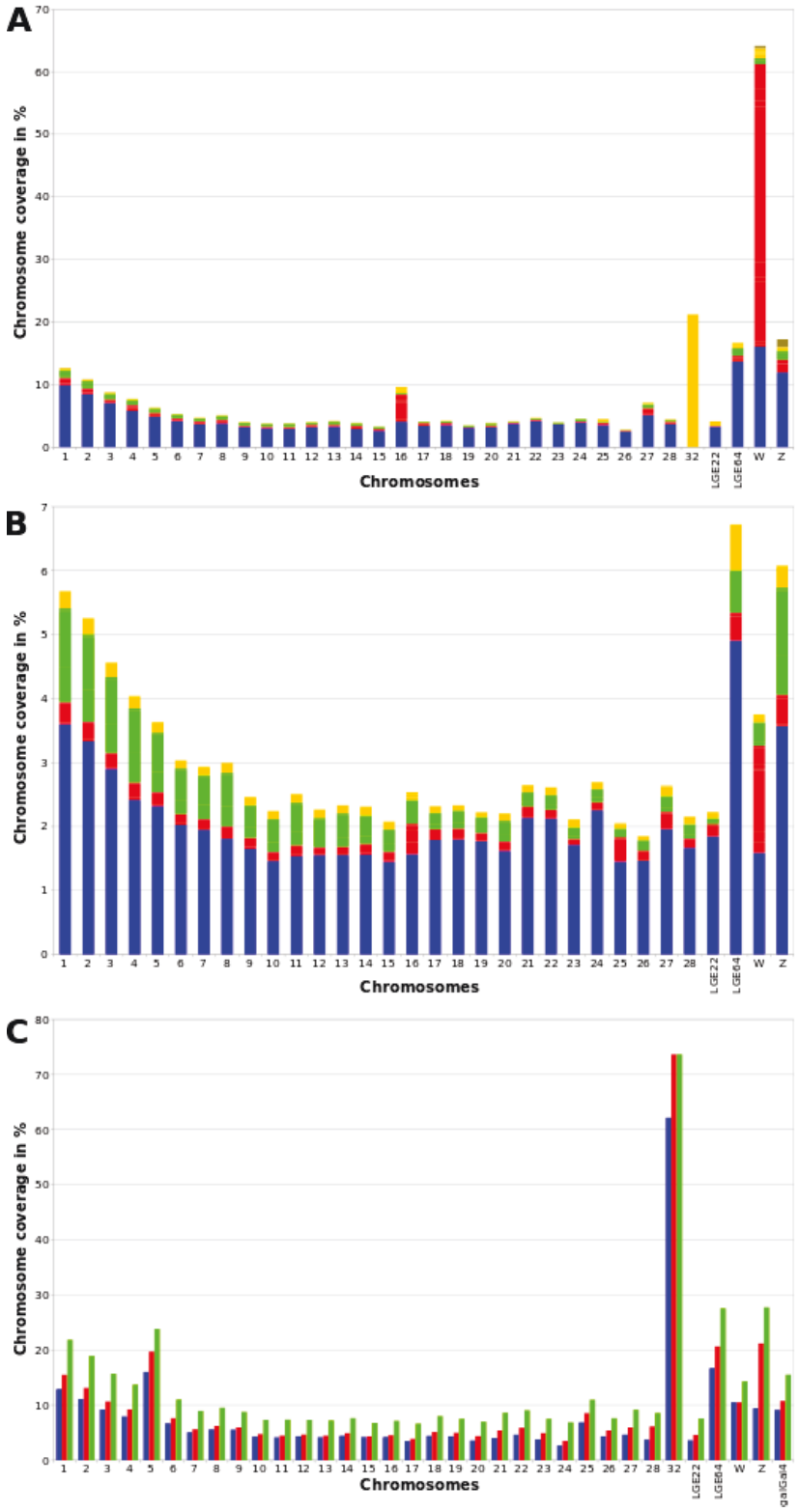
Annexe 6 : Ré-annotation et re-découverte du modèle Galgal4 - Figure 2



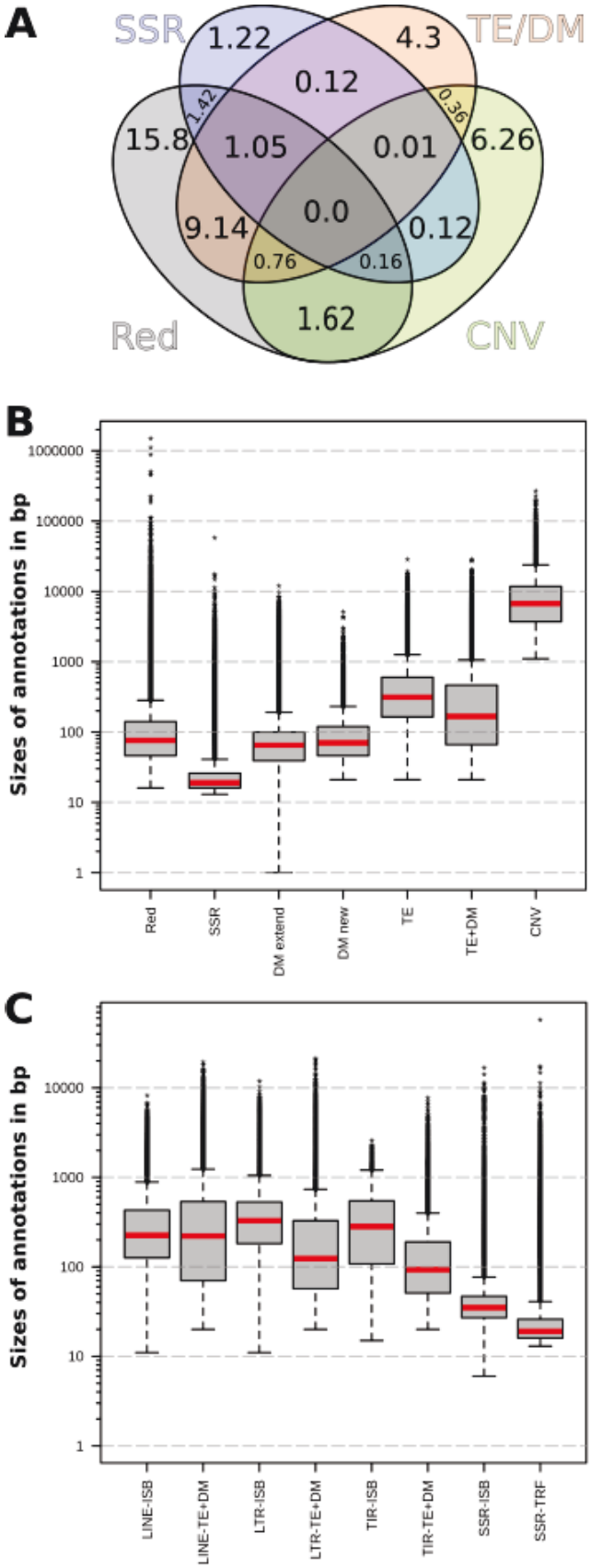
Annexe 7 : Ré-annotation et re-découverte du modèle Galgal4 - Figure 3



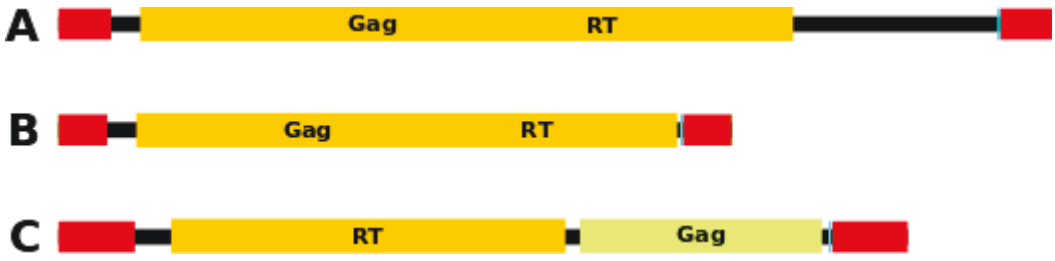
Annexe 8 : Ré-annotation et re-découverte du modèle Galgal4 - Figure 4



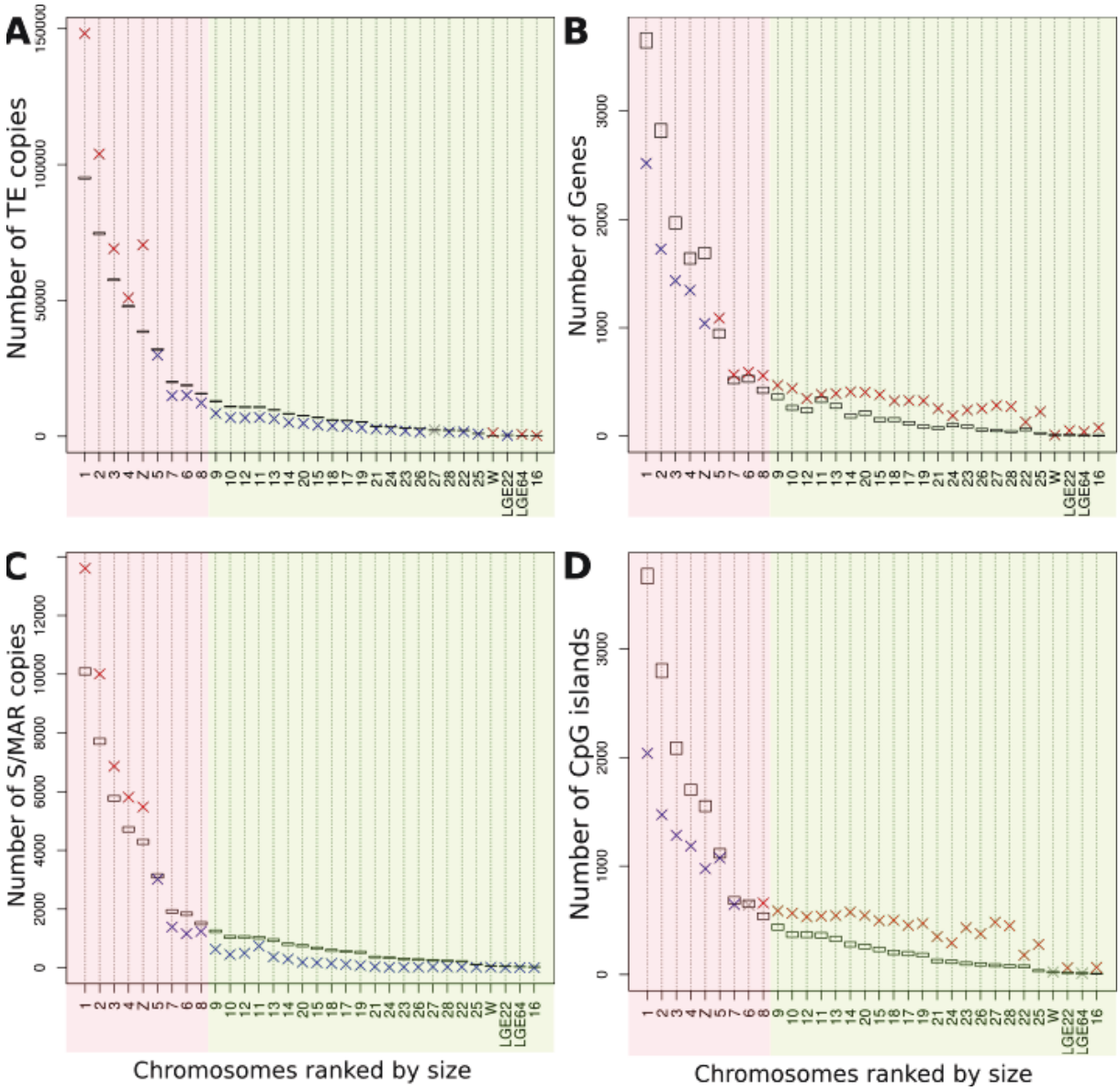
Annexe 9 : Ré-annotation et re-découverte du modèle Galgal4 - Figure 5



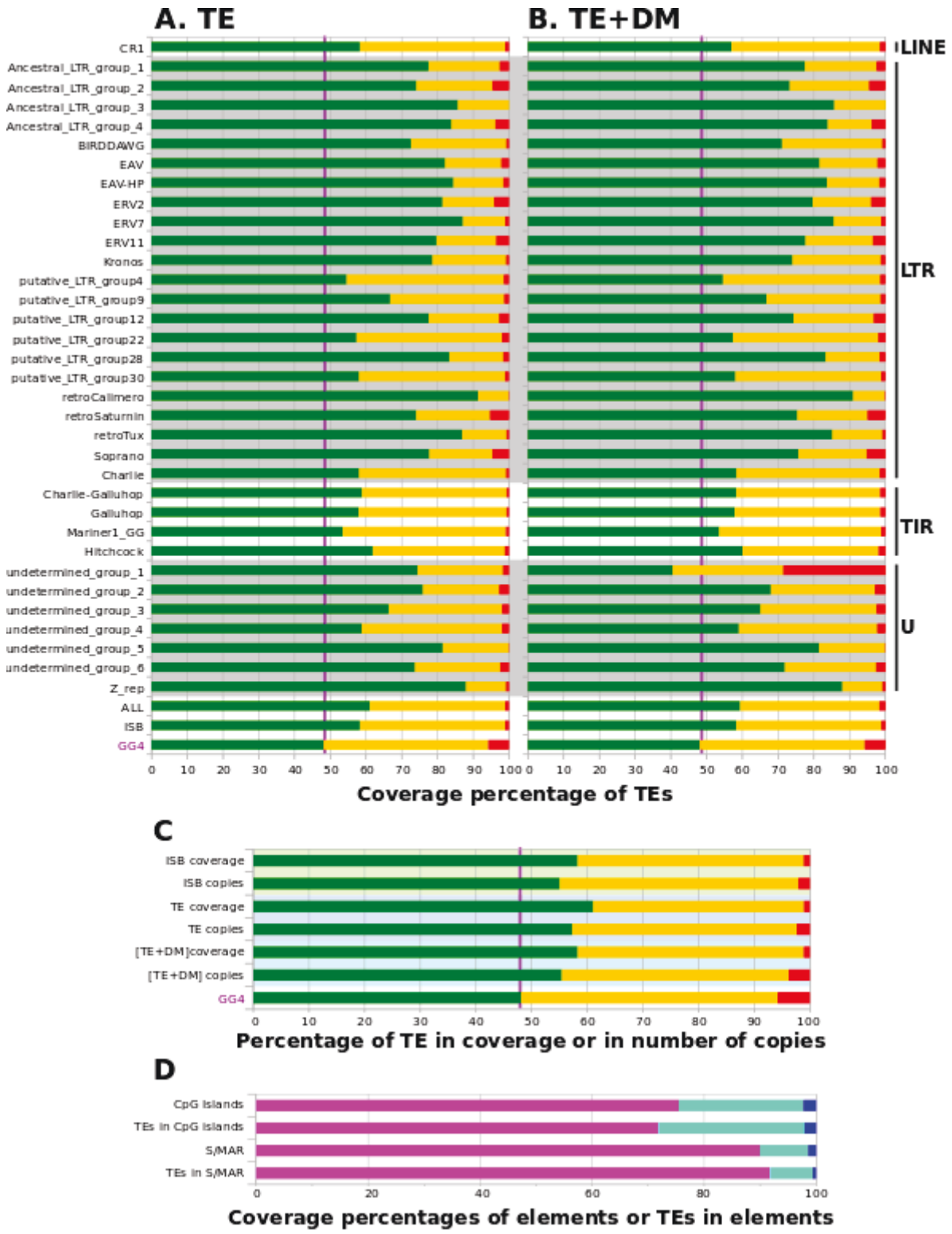
Annexe 10 : Ré-annotation et re-découverte du modèle Galgal4 - Figure 6



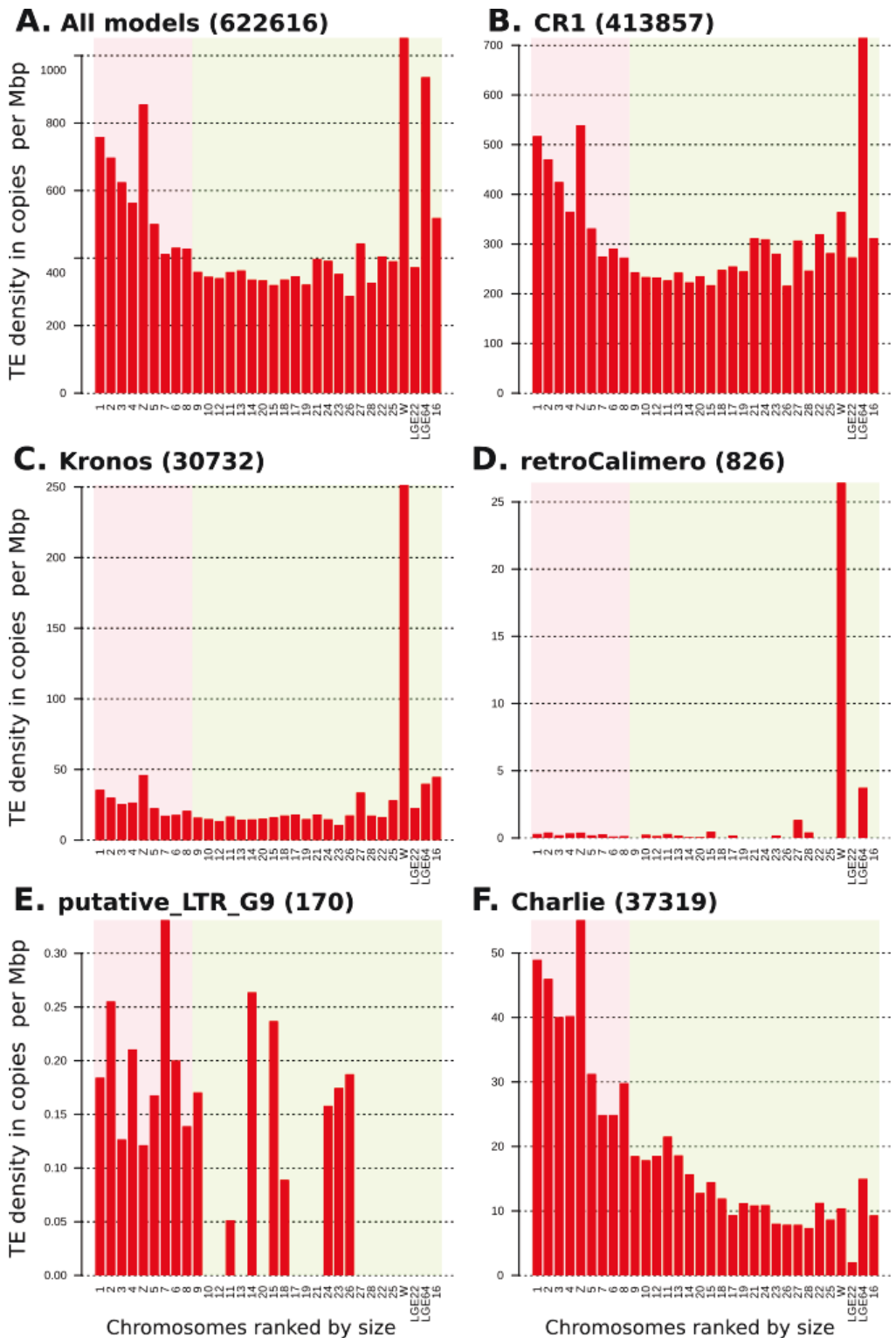
Annexe 11 : Ré-annotation et re-découverte du modèle Galgal4 - Figure 7



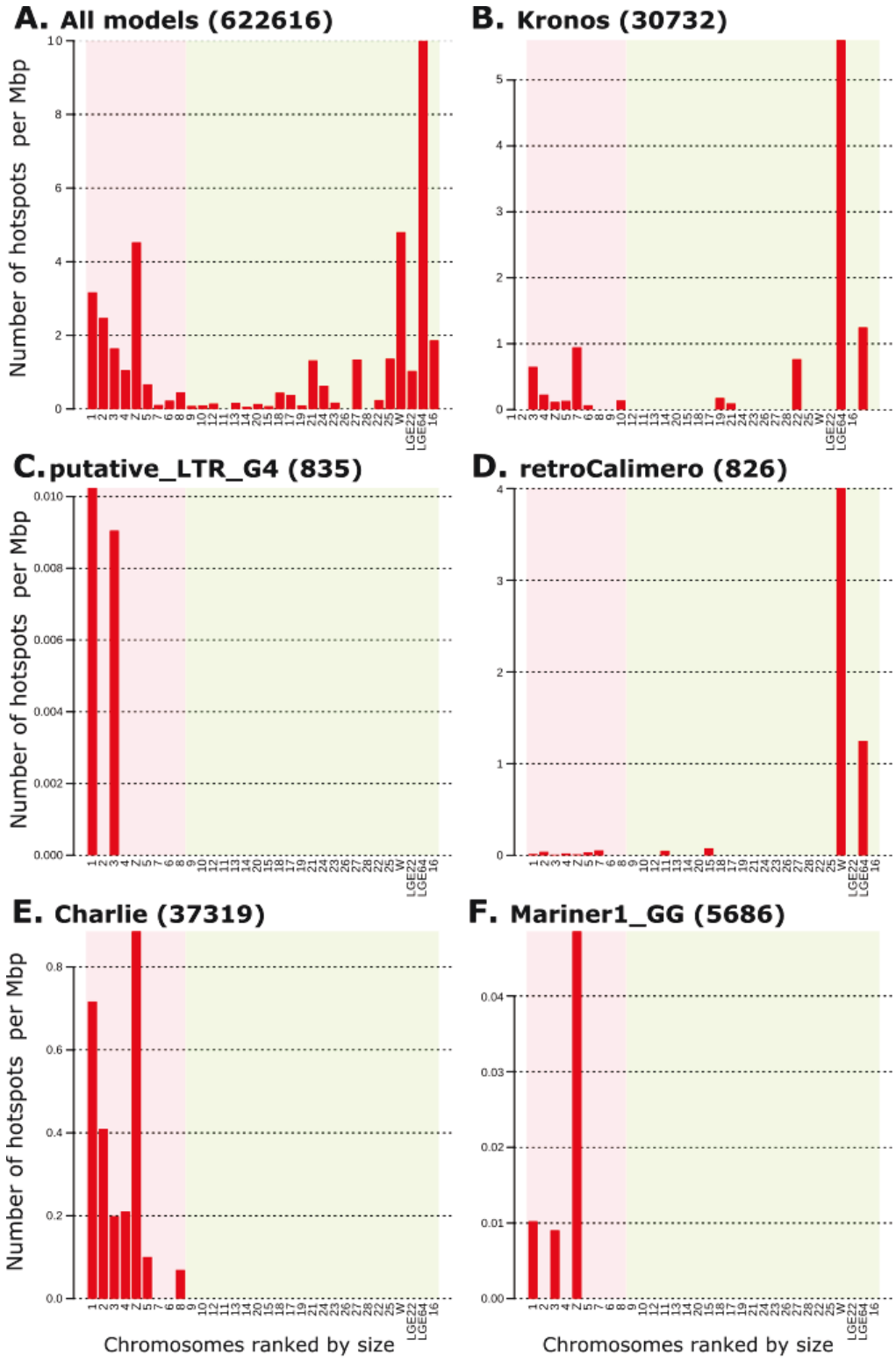
Annexe 12 : Ré-annotation et re-découverte du modèle Galgal4 - Figure 8



Annexe 13 : Ré-annotation et re-découverte du modèle Galgal4 - Figure 9



Annexe 14 : Ré-annotation et re-découverte du modèle Galgal4 - Figure 10



Annexe 15 : Ré-annotation et re-découverte du modèle Galgal4 - Additional file 1

Additional file 1: Conditions of use for programs P-clouds and Red.

P-clouds has previously been used to calculate the rate of repeats in the human genome (63-69 %; [64]) and in the Arabidopsis thaliana genome (11 %; [25]). This tool evaluates the amount of repeats by using a k-mer method . Because a batch of optimal parameters must be calculated for each genome analys, a benchmarking with 6 batches of 5 parameters (c4, c5, c8, c10, c100 and c200, see Figure and Material and methods, section “P-clouds” for the definition of each parameter) was assess on galGal4 using a oligo size of 16 nucleotides [63]. The most relaxed parameter set was c4 and the most stringent c200. The quality of each parameter set was verified by calculating for each the amount of exons within the P-clouds annotation. The highest repeat coverage was obtained with the least stringent parameter, c4. The amount of exons in the P-clouds annotation varying very little (between 5.42 % and 5.91 % from c200 to c4). We concluded that the c4 set was suitable and that the global amount of detectable repeats in galGal4 was 33 %.

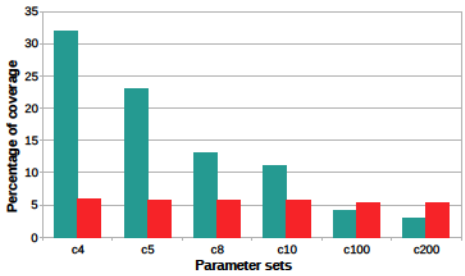


Fig. S1 P-clouds coverages using 6 parameters sets. Bars in blue represent the coverage in the P-clouds annotation corresponding to the repeats in the genome and those in red, that obtained for the exons.

Red is recently published tool [15] and therefore does not have the sam track record as P-coulds. This tool evaluates the amount of repeats using a k-mer method that locates repeats based on a machine learning approach. This program only requires the sequence of the genome model and an oligo size that is calculated using the same rule as P-clouds (i.e. 16 nucleotides). The global amount of detectable repeats in galGal4 obtained with Red was 29.9 %, a value close to that obtained with P-clouds.

Annexe 16 : Ré-annotation et re-découverte du modèle Galgal4 - Additional file 2

Additional file 2: Evaluating the efficiency of P-clouds, REPET, and Red.

The amount of repeats in the genome model of *Anopheles gambiae* (AgamP3)[4] was determined using three programs P-Clouds, Red and REPET. The coverages obtained were then compared to those obtained with RepeatMasker (RM) in previous publications[5,6]. Results are summarized in Table 1 and 2.

Table 1 Amounts of repeats calculated by four programs : RM, REPET, P-clouds and Red in AgamP3.

	Coverage of AgamP3 in bp	Coverage of AgamP3 in %
AgamP3	277645527	100
Repeat amount identified by RM	41794934	15.1
Repeat amount identified by REPET	73501425	26.5
Repeat amount identified by P-clouds	74964293	27.4
Repeat amount identified by Red	78322870	28.2

4 VectorBase : Bioinformatics Ressource for Invertebrate Vectors of Human Pathogens. https://www.vectorbase.org/ (2007) Accessed 2015 Sep 14.

5 Arensburger P, Megy K, Waterhouse RM, Abrudan J, Amedeo P, et al. Sequencing of *Culex quinquefasciatus* establishes a platform for mosquito comparative genomics. Science. 2010;330: 86-88.

6 Holt RAG, Subramanian M, Halpern A, Sutton GG, Charlab R, et al.The Genome Sequence of the Malaria Mosquito Anopheles Gambiae. Science. 2002;298:129.

Table 2 Overlaps between repeat annotation obtained with RM, REPET, P-clouds and Red in AgamP3.

Pair-comparison between files	Percentage coverage of the first pair member by the second pair member
RM / P-clouds	61 %
RM / REPET	100 %
RM /Red	84 %
REPET / P-clouds	61 %
REPET / Red	99 %
P-clouds / Red	62 %

Taken together these results indicated that REPET and Red led to evaluations that comprised for both of them 100 % of the RM annotations. By contrast, even though P-Clouds reported percentages of repeats similar to those calculated by REPET and Red these repeats do not overlap 100% with the RM annotations, including a large number of repeat annotations that were specific to this method. Similar results were obtained with the dmel6_07 model[7] of the *Drosophila melanogaster* genome. These results led us to abandon the use of the P-clouds method for that of Red which has the advantage of having 100 to 1000 folds faster calculation time than that of REPET, RepeatScout and RepeatModeler packages[8]. In agreement with these observation, it should be noted that P-clouds was only annotated 57.3 % of repeats contained in the official TE annotation of *Arabidopsis thaliana*[9].

7 Flybase. ftp://ftp.flybase.net/genomes/Drosophila_melanogaster/dmel_r5.55_FB2014_01/ (1994) Accessed 2015 Sep 14.

8 Girgis HZ. Red: an intelligent, rapid, accurate tool for detecting repeats de-novo on the genomic scale. BMC Bioinformatics. 2015;16:227.

9 Arabidopsis thaliana TAIR V10 Tes.

https://urgi.versailles.inra.fr/gb2/gbrowse/tairv10_pub_TEs/ (2014) Accessed 2015 Sep 14.

Annexe 17 : Ré-annotation et re-découverte du modèle Galgal4 - Additional file 3

Additional file 3: Features of SSRs in galGal4

Simple sequences

For there were 189543 and 14891 loci respectively containing simple sequences, a polyA stretch (4372699 bp; 0.42% of galGal4) or a polyC stretch (283691 bp; 0.03% of galGal4) of at least 50 nucleotides (Table 3). With respect to the number of polyC stretches in galGal4, they were approximately 13-folds higher than those of polyA stretches. This observation cannot be explained by the activity of the CR1 non-LTR retrotransposons. Indeed, in contrast to non-LTR retrotransposons such as the human L1[10,11,12], avian CR1s did not contain a polyA at their 3' ends, but instead contained between 2 and 4 copies of an octamer repeat (ATTCTRTG) [60][13,14]. Another source of AT-rich elements in eukaryotic genomes are the scaffold/matrix attachment regions (S/MARs, that covered 2.279% in galGal4). These are DNA elements that are functionally repeated elements able to be attached by a nuclear matrix and where RNA polymerase II complex would locate when they are involved in transcription[15,16,17]. The intersection between the chromosomal positions of both kinds of elements revealed that polyA stretches only covered 1.4 % of S/MARs. In conclusion, the

10 Ichiyanagi K, Okada N. Mobility pathways for vertebrate L1, L2, CR1, and RTE clade retrotransposons. Mol Biol Evol. 2008;25:1148-1157.

11 Kaessmann H, Vinckenbosch N, Long M. RNA-based gene duplication: mechanistic and evolutionary insights. Nat Rev Genet. 2009;10:19-31.

12 Monot C, Kuciak M, Viollet S, Mir AA, Gabus C, et al. The specificity and flexibility of l1 reverse transcription priming at imperfect T-tracts. PLoS Genet. 2013;9:e1003499.

13 Coullin P, Bed'Hom B, Candelier JJ, Vettese D, Maucolin S, et al. Cytogenetic repartition of chicken CR1 sequences evidenced by PRINS in Galliformes and some other birds. Chromosome Res. 2005;13:665-673.

14 Liu GE, Jiang L, Tian F, Zhu B, Song J. Calibration of mutation rates reveals diverse subfamily structure of galliform CR1 repeats. Genome Biol Evol. 2009;1:119-130.

15 Girod PA, Nguyen DQ, Calabrese D, Puttini S, Grandjean M, et al. Genome-wide prediction of matrix attachment regions that increase gene expression in mammalian cells. Nat Methods. 2007;4:747-753.

16 Arope S, Harraghy N, Pjanic M, Mermod N. Molecular characterization of a human matrix attachment region epigenetic regulator. PLoS One. 2013;):e79262.

17 Harraghy N, Calabrese D, Fisch I, Girod PA, LeFourn V, et al. Epigenetic regulatory elements: Recent advances in understanding their mode of action and use for recombinant protein production in mammalian cells. Biotechnol J. 2015;10:967-978.

numerous stretches of polyA in galGal4 were therefore neither products of aborted events of CR1 retrotransposition nor parts of S/MAR elements.

Microsatellites

There were 770,202 loci containing one of the 2101 microsatellites found in galGal4 (Table 3). Previously Brandström et al [9] indicated that there were 1,615,000 microsatellites that were 6 bp or longer in galGal4. Using a similar size threshold, we found using TRF, 1,997,228 microsatellites. This suggests that this software is more efficient overall than sputnik. Among microsatellites, we expected to find (TTAGGG)n repeats that are located at telomer ends. Our investigations revealed that only 8 kbp of such repeats were detected in galGal4 whereas approximately 2-4% (i.e. ~21-42 Mbp) are present in the real RJF genome [42] This confirms that telomeres are currently not present in galGal4. To our knowledge, there has not been an analysis of minisatellites at the genome scale in the RJF genome since 1994[18,19]. We identified 12,310 loci containing one of the 123 different minisatellites found in galGal4 (Table 3). With respect to their numbers and the length of their repeated units, the sequence diversities of microsatellites and minisatellites were rather low (2101 and 123 different repeated units, respectively). We verified whether this reduced variability was due to their involvement in S/MAR elements and found that they only covered 5.8 % of S/MARs.

Large tandem arrays

Analysis of large tandem arrays (Table 3) revealed that only 6 arrays contained less than 50 tandemly repeated units. Above this threshold, 10,136 loci containing repeats ranging in size from 60 bp to 2 kbp were identified and considered as being satellite DNA arrays. Overall they represented 0.238% of galGal4 in spite of the absence of most large satellite DNA families from galGal4. Some of them were clustered using SiLiX[20] in 4797 to 8678 families, depending on the similarity and overlapping rates used to align sequences (Table S1 and S2). Such a number of families indicated that there were numerous satellite DNA families in the

18 Bruford MW, Burke T. Minisatellite DNA markers in the chicken genome. I. Distribution and abundance of minisatellites in multilocus DNA fingerprints. Anim Genet. 1994;25:381-389.

19 Bruford MW, Hanotte O, Burke T. Minisatellite DNA markers in the chicken genome. II. Isolation and characterization of minisatellite loci. Anim Genet. 1994;25:391-399.

20 Miele V, Penel S, Duret L. Ultra-fast sequence clustering from similarity networks with SiLiX. BMC Bioinformatics. 2011;12:116.

chicken genome that could be split in two types. Those that are abundant in telomeric and centromeric regions of some chromosomes (review in [5,44]) and that are widely absent in galGal4, and those described here, numerous, finally few repeated in galGal4 and diverse in DNA sequence. The distribution of these numerous few repeated satellite DNA arrays was therefore investigated in order to verify whether they were mainly concentrated near telomeres and centromeres, or whether they were interspersed along chromosome arms. Our results (Figure S2) confirmed that the density of these satellite DNA arrays was more elevated in microchromosomes and showed that they would not be concentrated in centromeric and telomeric regions, but distributed in the inner regions of the chromosome arms. Because the positions of every centromere in galGal4 chromosomes are not currently available, this last observation was not be tested. Nevertheless, these observations suggested that the profile of density, sequence diversity and distribution of minisatellites and the rare repeated satellite DNAs in numerous microchromosomes might be a signature reflecting a specific sequence organization and maybe specificities in their functioning or localization in avian nucleus, as previously proposed [44,101].

Table S1. SiLix clustering of satellite DNA repeated units with three parameter sets

Clustering 70 / 70	
a	b
2	616
3	184
4	81
5	44
6	22
7	22
8	13
9	10
10	7
11	3
12	4
13	5
16	3
17	1
18	1
20	1
21	1
22	2
24	2
27	1
28	2
33	1
37	2
43	1
54	1
65	1
77	1
83	1
93	1
96	1
110	1
113	1
114	1
266	1
317	1
332	1
440	1
600	1

Clustering 80 / 80	
a	b
2	522
3	175
4	38
5	40
6	15
7	9
8	8
9	5
10	3
11	5
12	3
13	4
14	1
16	2
17	1
19	1
20	1
21	1
22	1
23	1
28	1
29	1
30	1
43	1
49	1
51	1
53	1
57	1
63	1
65	1
67	1
98	1
103	1
108	1
125	1
206	1
301	1

Clustering 90 / 90	
a	b
2	362
3	116
4	25
5	24
6	7
7	11
8	3
9	3
10	3
12	1
13	2
14	1
15	1
17	1
21	1
24	1
27	1
30	2
34	1
65	1
70	1

a, Number of repeated units in each cluster

Table S2. Diversity of satellite DNA in Galgal4.

Parameters used for the clustering*	70 / 70	80 / 80	90 / 90
Numbers of clusters	1043	853	568
Number of loci in clusters	6233	4000	1877
Range of loci per cluster	2 to 600	2 to 301	2 to 70
Numbers of single loci	3754	5987	8110
Numbers of satellite DNA families	4797	6840	8678

* Similarity rate / overlapping rate used to compare two sequences

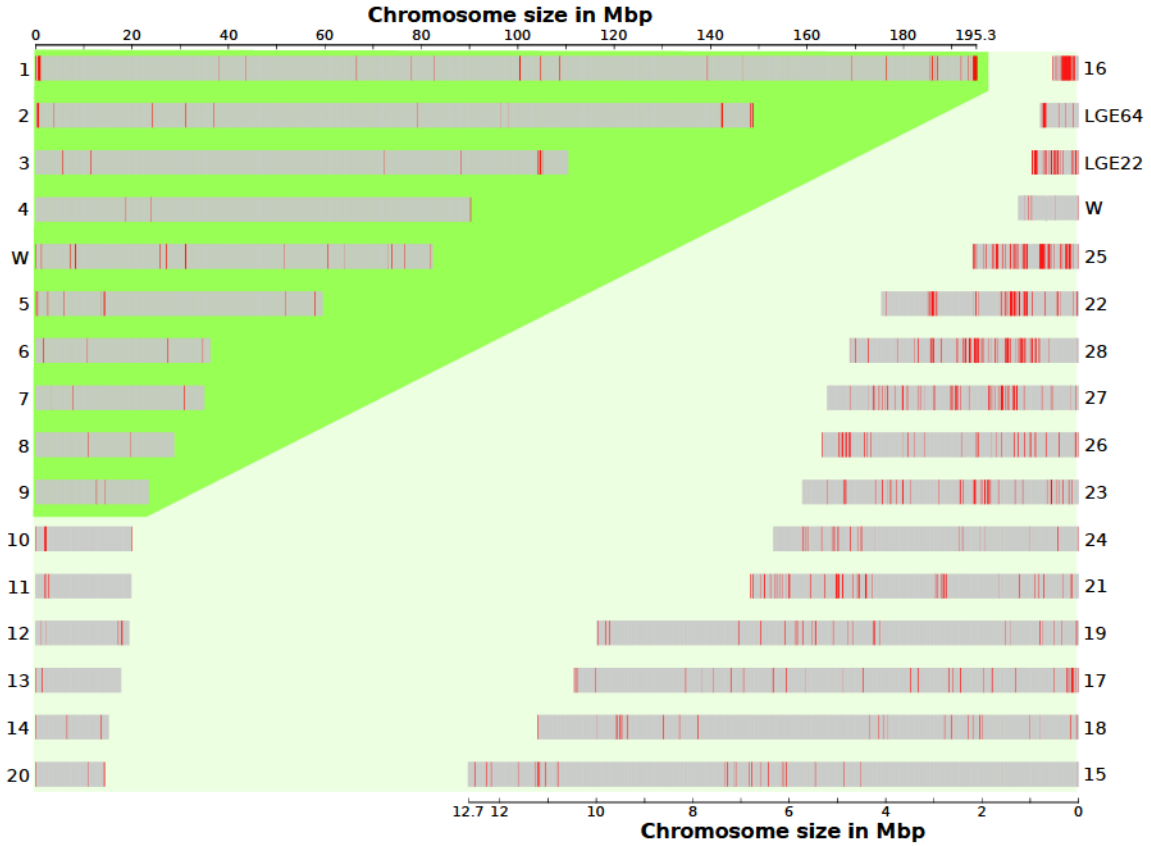
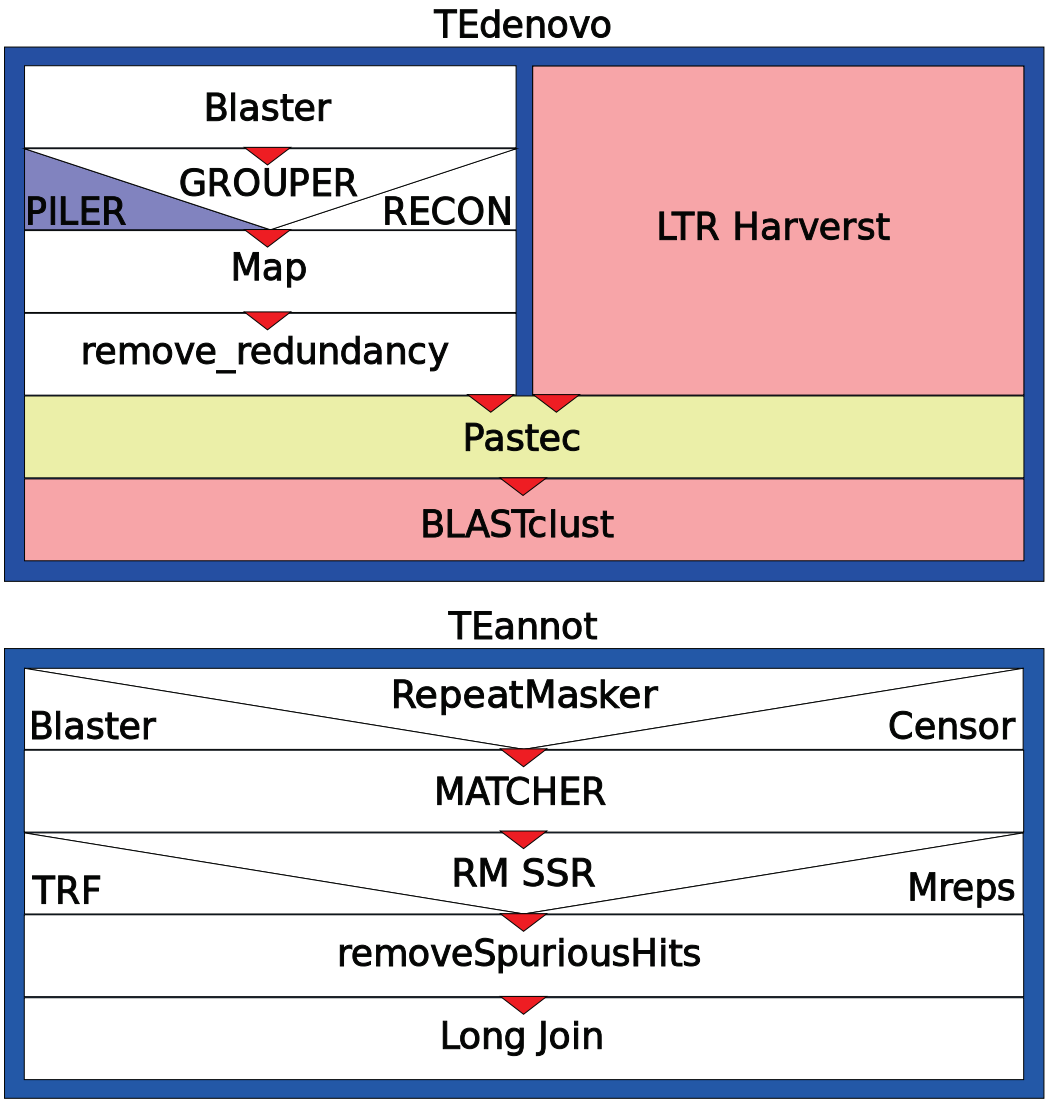


Fig. S2 Distribution of satellite DNAs in galGal4 chromosomes. The size scale of chromosomes 1 to 16 is on the top, that for others chromosomes was on bottom of the graphic. Red bars represented regions containing satellite DNA arrays. Background areas in pink contained the 9 macrochromosomes plus the W, that in green all galGal4 microchromosomes plus the Z. Due to its size, the data concerning the chromosome 32 were not drawn in the three graphics. The chromosomes maps were drawn with a homemade tool DensityMap.pl (available at http://chicken-repeats.inra.fr/) using windows to calculate densities of satellite DNA arrays in chromosomes 1 to 16 and other chromosomes that were respectively of 100 kbp and 10 kbp.

Annexe 17 : Ré-annotation et re-découverte du modèle Galgal4 - Additional file 4

Additional File 4: Schematic summary of programs intervening in both components of the REPET pipeline TEdenovo and TEannot.

In both components, programs intervened successively or in parallel from the top to the bottom as indicated by red triangles. In TEannot the three programs with a background in blue or pink were removed for our analyses, and the outputs of that in yellow needed to be verified by hand. It must be noticed that RepeatMasker is one of the program of the TEannot component.



Annexe 18 : Ré-annotation et re-découverte du modèle Galgal4 - Additional file 5

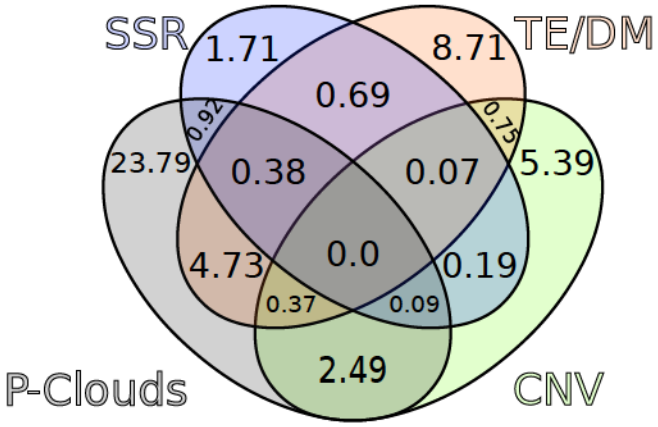
Additional file 5: TE coverages in each galGal4 chromosomes in the ISB, REPET TE, TEannot DM and TE+DM annotations

Chromosome	Coverage ISB (%)	Coverage REPET TE (%)	Coverage TEannot DM (%)	Coverage TE+DM (%)	Increase factor between [TE+DM] & ISB annotations
1	12,91	15,24	6,14	21,38	1,66
2	11,11	12,95	5,66	18,61	1,67
3	9,20	10,57	4,90	15,46	1,68
4	7,95	9,10	4,41	13,52	1,7
5	6,68	7,58	3,99	11,57	1,73
6	5,59	6,17	3,29	9,46	1,69
7	5,07	5,60	3,24	8,83	1,74
8	5,54	5,89	3,23	9,12	1,65
9	4,27	4,74	2,74	7,48	1,75
10	4,16	4,43	2,50	6,94	1,67
11	4,17	4,45	2,82	7,27	1,74
12	4,35	4,63	2,57	7,20	1,66
13	4,44	4,87	2,69	7,57	1,7
14	4,20	4,31	2,68	6,99	1,67
15	3,48	3,87	2,41	6,27	1,8
16	10,51	10,45	2,57	13,02	1,24
17	4,34	4,97	2,67	7,63	1,76
18	4,42	5,09	2,75	7,84	1,77
19	3,59	4,37	2,43	6,80	1,9
20	4,21	4,52	2,50	7,01	1,67
21	4,05	5,39	3,11	8,50	2,1
22	4,65	5,95	3,07	9,03	1,94
23	3,77	4,92	2,51	7,42	1,97
24	4,65	5,87	3,21	9,09	1,96
25	3,72	6,05	2,42	8,48	2,28
26	2,67	3,50	2,15	5,65	2,11
27	6,87	8,34	3,13	11,47	1,67
28	4,34	5,36	2,45	7,81	1,8
32	9,44	21,21	0	21,21	2,25
LGE22	3,62	4,57	2,86	7,44	2,05
LGE64	16,71	20,64	6,67	27,31	1,63
W	62,17	66,44	3,76	70,20	1,13
Z	16,03	19,30	6,40	25,70	1,6
Total	8,75	10,17	4,44	14,62	1,67

Annexe 19 : Ré-annotation et re-découverte du modèle Galgal4 - Additional file 6

Additional File 6: Intersections between annotation files calculated with P-clouds, TRF, and REPET.

Overlaps between the annotation files calculated with P-clouds, TRF (SSR), and REPET (TE+DM), and CNVs [11]. Values are percentage coverages in galGal4.



Annexe 20 : Ré-annotation et re-découverte du modèle Galgal4 - Additional file 7

Additional File 7: Correspondence between the names of consensus describing TEs in Repbase and ISB, and the TE models calculated with REPET

Repbase and ISB Names	Name in our annotation	TE types
CR1AVI, CR1_B, CR1-B2, CR1-C2, CR1-C3, CR1-C4, CR1-C, CR1-D2, CR1-D, CR1_F, CR1-F2, CR1-G, CR1_GG, CR1-H2, CR1-H, CR1L, CR1-X1_3end, CR1-X2_3end, CR1-Y1_Aves, CR1-Y2_Aves, CR1-Y2, CR1-Y3, CR1-Y4, CR1-Y	CR1	LINE
Absent	Ancestral_LTR_group_1	LTR
Absent	Ancestral_LTR_group_2	LTR
Absent	Ancestral_LTR_group_3	LTR
Absent	Ancestral_LTR_group_4	LTR
Birddawg_I, Birddawg_LTR, GGERV10_LTR, GGERV10_RT, GGERV11_RT, GGERV20_I, GGERV23_LTR	BIRDDAWG	LTR
GGERVK1, GGLTR1	EAV	LTR
GGLTR10A_LTR, GGLTR10B_LTR, GGLTR10C1_LTR, GGLTR10C2_LTR, GGLTR10C_I, GGLTR10D_I, GGLTR10D_LTR	EAV-HP	LTR
GGERVK10	ERV2	LTR
GGLTR7A_LTR, GGLTR7B_LTR, GGLTR7_I	ERV7	LTR
GGLTR11_LTR, GGLTR11_I	ERV11	LTR
GGERVL-A, GGERVL-C, GGLTR3B2, GGLTR3B3, GGLTR3B4, GGLTR3C1, GGLTR3C2, GGLTR3D, GGLTR3E1_LTR, GGLTR3E3_LTR, GGLTR3F1_LTR, GGLTR3G2, Kronos_I, Kronos_LTR	Kronos	LTR
GGLTR4A, GGLTR4B	putative_LTR_group4	LTR
GGLTR9_LTR	putative_LTR_group9	LTR
GGLTR12A, GGLTR12B, GGLTR12C	putative_LTR_group12	LTR
GGERV22_LTR	putative_LTR_group22	LTR
GGERV28_LTR	putative_LTR_group28	LTR
GGERV30_LTR, GGERV21_LTR	putative_LTR_group30	LTR
Absent	Retrocalimero	LTR
Absent	Retrosaturnin	LTR
Absent	RetroTux	LTR
Soprano_I, Soprano_LTR, ENS1B-LTR	Soprano	LTR
Charlie12_GG, Charlie12_Gga	Charlie	TIR
Charlie-Galluhop, Mariner1b_GG	Charlie-Galluhop	TIR
Galluhop	Galluhop	TIR
Mariner1_GG,	Mariner1_GG	TIR
Hitchcock_LTR	Hitchcock	
Absent	undetermined_group_1	undefined
Absent	undetermined_group_2	undefined
Absent	undetermined_group_3	undefined
Absent	undetermined_group_4	undefined
Absent	undetermined_group_5	undefined
Absent	undetermined_group_6	undefined
Z_Rep	Z_rep	undefined

Annexe 21 : Ré-annotation et re-découverte du modèle Galgal4 - Additional file 8

Additional File 8: Diversity of CR1 within galGal4

It was reported in Repbase[21] that CR1 are divided into 22 subfamilies in the RJF genome, and studies based on the 3' ends of CR1 copies classified them in 57 sub-families[2]. We reinvestigated this by performing two successive single linkage clusterings using SiLiX with thresholds to gather copies in a cluster fixed at a minimum of 80% overlap and 80% sequence identity, as recommended in Wicker et al [60]. The first clustering was performed using 308 CR1 consensuses, 82 of them having a small size that excluded them from the process. It gathered 226 consensuses within 8 clusters, 7 of them matching with the Repbase sub-families CR1-C, CR1-D, CR1_F, CR1-G, CR1_GG, CR1-H, and CR1-Y (Table S2). The eighth cluster was named CR1_like. In order to avoid creating false CR1 copies in the final [TE+DM] annotation of galGal4 by fusing juxtaposed annotations that were defined from consensuses related to different CR1 subfamilies, this division into 8 CR1 subfamilies was conserved for the step involving resolving and merging stacked and juxtaposed annotations. For the second clustering we used genomic copies annotated by 226 consensuses and containing the last 1000 bp located at the 3' end CR1 sequence. Results confirmed that copies belonging to one of the 6 clusters matching with CR1-C, CR1-D, CR1_GG, CR1-H, CR1-Y and CR1_like sub-families were gathered in the cluster linked to their annotation consensus. They also showed that copies belonging to one of the 2 clusters matching with CR1_F and CR1-G sub-families were gathered into 2 clusters that were different from those of their annotation consensus. Taken together these results suggested that there are be numerous CR1_F/CR1-G chimeric elements in galGal4. The emergence of chimeric elements between

21 Repbase at giri. http://www.girinst.org/repbase/ (2001). Accessed 10 Sep 2015.

non-LTR retrotransposons[22,23,24,25,26,27] followed by their amplification via retrotransposition is not a rare phenomenon in eukaryotic genomes. At the scale of the evolution of these genomes, it might therefore be proposed that the 22 to 57 sub-families previously described may have originated from frequent emergences of chimeric CR1 elements originating from a limited number of CR1 sub-families.

22 Losada A, Abad JP, Agudo M, Villasante A. The analysis of Circe, an LTR retrotransposon of Drosophila melanogaster, suggests that an insertion of non-LTR retrotransposons into LTR elements can create chimeric retroelements. Mol Biol Evol. 1999;16:1341-1346.

23 Buzdin A, Gogvadze E, Kovalskaya E, Volchkov P, Ustyugova S, et al. The human genome contains many types of chimeric retrogenes generated through in vivo RNA recombination. Nucleic Acids Res. 2003;31:4385-4390.

24 Gogvadze E, Barbisan C, Lebrun MH, Buzdin A. Tripartite chimeric pseudogene from the genome of rice blast fungus Magnaporthe grisea suggests double template jumps during long interspersed nuclear element (LINE) reverse transcription. BMC Genomics. 2007;8:360.

25 Buzdin A, Gogvadze E, Lebrun MH. Chimeric retrogenes suggest a role for the nucleolus in LINE amplification. FEBS Lett. 2007;581:2877-2882

26 Bao W, Jurka J. Origin and evolution of LINE-1 derived "half-L1" retrotransposons (HAL1). Gene. 2010;465:9-16.

27 Luchetti A, Mingazzini V, Mantovani B. 28S junctions and chimeric elements of the rDNA targeting non-LTR retrotransposon R2 in crustacean living fossils (Branchiopoda, Notostraca). Genomics. 2012;100:51-56.

Table S2 Features and diversity of CR1 models found in the Galgal4 model based on the REPET and DM annotations (STEP3 + STEP4, Figure 2) after stack resolving and merging stacked and juxtaposed annotations.

Names of TE models	a	b	c*	d	e
CR1-C	1	LINE	27113	4.3547	0.6970
CR1-D	1	LINE	56909	9.1403	0.9380
CR1_F_NV2	2	LINE	36222	5.8177	0.8303
CR1-G	2	LINE	9635	1.5475	0.2885
CR1_GG	2	LINE	10791	1.7332	0.3133
CR1-H	3	LINE	12544	2.0147	0.4253
CR1-Y	1	LINE	211266	33.9320	7.4710
CR1_like	296	LINE	49377	7.9306	0.8823
Total	308		413857	100	11.8457

a, Number of consensus ; b, TE types ; c, Total number of TE copies ; d, Percentage of the total number of TE copies ; e, Percentage of chromosome coverage; *, Post stack resolving and annotation merging are called copies all complete elements, internally deleted elements; 5' or 3' truncated elements and elements truncated at both ends (i.e. internal regions of a TE devoid of ends).

Annexe 22 : Ré-annotation et re-découverte du modèle Galgal4 - Additional file 9

Additional File 9: Nucleic acid sequences of retroCalimero, retroSaturnin and retroTux.

A. retroCalimero

TGTGGAGAGATTTATTGATGACTGGCTGACTGAGAAAAGCCATGCAGACACTGTGCAGCTGG
TTAAGCAACCTTGTTACGCGCAAGGTCAGGTATGAGAACTAGGAACAAAACAGGATGCCGAT
GGAGAAAGCCGGCAACAGTGTTTTGAGAAGTCTGTAAGAGAGTGTTTGCTGAGAGATAACCA
CAGAGCACAAAGATAGGCGTAGTTGACACTGACTAGCGAGCCAATCCTGAGCTCCACTTTAG
CAATATGTATGAGCCTATTATCCACACTATAAATGTACACGCAGCTGAAACAATAAACGAGA
TTTGCTGATCATCTACTATGGTCGTCGCATTTCTCCCGCTGATTTCCTACAAATGGTGACCC
CGACTGACCCAGTAACGGCGCACCACGGTGGGACGGCGTAGGAGTTCGGGTCCTAAAGCAGT
GAGGACGGCTAAAGCACTCAGGAACCAGAGATAGGAGTGCCGAGCCCTAGGTGACCTCATCG
GAGAGCCTGGCCGATACGGGCCGCCGAGACCTTGTAAGGCTGCAATAGCGCGCGCTGACAGA
TAAGGGATGGAAAGGCAAGCGGCTTATGATTTATTTACTTGTTTTCTGAGAAAACGACAAAT
TAAGGGAGTAGACTTACAGAAGGAACTCCCCGGGTTATTAGCTTATGGCAATGCTAAAGGGT
TTTTTCAAAACCCACATTTAGTACATGACCTTACAGAATGGCGTAAATTTGGGGATGCAATT
TGGCAAGCAGTTTTAGATGATGATAAGACAGCAAAAAAATGCGGGAAACTGTGGCGAGCAGT
TCACAATGAACTGTTGCAGCACCAAGCAGAAAAGAAGGCAGCGGAACAGGCCAGTGCTGCAC
AGGGGAGAAATAAGGATTACAATGGGTGGCCGGATTGTCCGTTGCCCCCCGCAGTGACTACC
TTAGTGTTGCCGCCGGCCGCGGAGCACGCCACATCAGCATGTTTGGAGCCAAGCGCCCCACC
CCCGCCCCTAGTTACGGCCGCGCCGGGTAGTTCACCTGCGCCTATTACAATGGGCCCAATAA
ACCCTCCGAGCAGTGAGCCGATCCCTGGGGCAGAGAGTGACCTAGCGGGGGCCATTGCACGG
GAGAGGCGGGAAGCTTGGGCGGCGCTGGCACGGGCCTGCATGGAGGCGGGGGATAACGACGC
AGTGGAGGCCGTGCGGCACATGGCGTTTCCGGTGATATATGCCTCTAATCCTGCGGGAAGGA
TGCAGGCCACCATCACAGCTTTAGATTGGAAATTATTATCACAGCTACGATCTACAGTTAGT
CAGTTTGGGGTAAAAAGCGAGCCAGCTAAGCAGATGTTGGATTATATTTGGAGTACGCAGAT
ATTGCTGCCATCCGATTGTTGGGCAATAGCAAAATTGATCTTTTCACAACATCAACAGCTGT
TGTTTAATGCATATTGGCAAGAACTGTGCCATCAGAGTGTCTCAAAGGCTAGGCAGCCGGGA
GACCCACTACATGGTGTAACTATCAAAGAACTCTTAGGGCTAGGGCCTTTTTTCAGAACACA
AGCCCAAGCATTATTAGGGCCAGATAAATGTCGAGAAACTATGTATTTAGCTAGACAGGCCA
TGGACAAGATTAAGGTGCCTGATGGATTGCCATTTTATATGGGGATCCGACAAGGTAGAGAT

GAGGACTTTGGGGCATTCATAGACAAGGTAGCCGGGGCTATTGAAAAGGCAGGGGTGCCAGA
GTATATGAGAGGAGTGATGTTAAAACAATGCGCGCTCCAAAATTGTAATTCAACAGCACGTA
GTATTCTGAATACTTTGAGGAGTAATTGGACTATTGAGGAGGCACTGGAAAAGCTATCAAGC
GTGCCGGTCGGGCCCCAAGCATTTCTGGTTGAGGCTATTAAGGAGTTAGGGGCAGGTTTAAA
GGCGCAGGCTGAGGCCTCTCACAATCAAGTGCTAGCAGCTCTTGCACCCTTACAAGCGTCTG
CGACAACAAATTTAGGCCCGAGGTCGCCTGTTGCTGGTCGCATCAAATGCTATCGCTGTGGT
GGCATGGGACATGTGCGTCGTCAGTGTCAAGCCACCGGTTCATGGTGTCAGACATGTCGTAT
GGACAACCACAATACCAATGCCTGTCGACGCCGGTCGGGAAACCCCCAGCCCAGCGCGAGAA
AGAAACGATGGCCGCGCGCAGACACAAGTAGCCGTTTCCAGCCAGCAGCAGCCCTGCAACCA
GCCACACCAGGAAGCCTTGGCTTGGACTTGGCAGCCGCAGTGACCACGACCTTAATGACCAC
AAAACCTGAGCGGGTGTCTACGGGGATCAGGGGACCAGTAATGATAAATGGAACCGCCGTTG
GGGCCCTTCTATTGGGGCGTTCTTCAGCATCGATGCTCGGACTTTTTGTCCTCCCTGGGGTA
ATAGATGCGGACTTTCAGGGCGAGATCCAGATTATGGTATACACCCCGTTTCCTCCAATAAA
AATTGAGAAAGGGCAATGGATAGCGCAGTTAGTACCCCTAGAGCAATTGACCAAAGCCTTAA
CCCTGTGTCAATTGTCTCCCTGAGGGGAGCAAGGGTTTGGCTCCTCAGGGGGCTTAGCATTG
CTATCACTAAATTTGCATGATCGACCTAAAAAACCAGTGACACTTAAATACAGAGAGGAAGA
AATCAAACTTCAAGGTCTATTAGATACAGGGGCCGATAGCAGCATAGTGAGCCCAGAAATTT
GGCCGCCTCATTGGCCACTGCAAGCGGCCATAGCCACAGTGACGGGTATAGGCGGACTATCC
TTAGCGAAAAAGTCGCCTCCCCTGCAAATCCACCTTGATGAACAAGTAGTCCATACCTCTGT
CTCGGTGGCACCCCTACCCCCCACCGTTCAATGCTTAATTGGCCGCAATGTGCTCTCTCAAC
TGGGGGCACTTCAACCCGGCCTGCCCAATCCTGCCATGCTACCTTACAATTGGCCATTACTA
ATTATTGACTTAAAGGACTGTTTCTTTACTATTACCCTGCACCCTCAGGACACTAAACGATT
TGCCTTCACATTGCCTGCATTGAATCGAGAACACCCTGATCAACGATTTGAATGGACAGTAT
TACCTCAGGGGATGAAAAATAGCCCCACGCTATGTCAACTATATGTCGATCATGCTCTACAA
CCACTCCGGCGAGAATGGAAACAAATGGTCATTTATCATTACATGGATGACATCCTTTTTGC
CCAGCCAGAGGCCTTCACACAGGAACAAATTTGGCAAATAGAAAAGACCCTAAATAGGGAAG
GACTTATGATTGCCCCTGAAAAGGTACAACTCTCTGCGCCCTGGAAGTACTTAGGATGGACA
CTGACTAACACGATAGTAACCCCACAGAAACTGCAACTAAATACTAAAATAGAGACTCTACA
CGATGCCCAAAGGTTACTGGGGGACTTACAGTGGCTGTGCCCTGTGGTGGGCATCCCAAACG
AACTCTTAGAGTCGTTGCGACCTTTGTTACGGGGCACTGACCCGGCCCAGCCTGTAACGGTG
ACAACGCAGCACAAACGTCTACTACAACAGATTATGGACTGCATTATACACGGCAGTGTTCG
GAGACGTGACCCTGACCTCCCCATACAGGTTATGGTATGGTATGGACCAAAGTATCTTTTAG
GAGCGTTGGCACAATCTAAAAAGAAAACGGGGGAGGTATGGGTACTAGAGTGGATCTGTCCC

TCACTGCAGCAATCAAAAACACTTCTTCAAAAAATTGAGCTCCTGGCAGAAGTGATTAAGAA
AGGGCGAGAACGTACCCTGCAAATCACAGGTATGGAGCCTGTGTGTGTACAGCTGCCAATGC
AGAAGGACACTCTGACATGGTATGTGCAGCATAGTCCAGAGTTACAGGATGCTCTCTTACGA
GCTGGAAGTACGGTTTCAATGGAAAAGATTCCGAACGCGCCGCTACATTGGATTGGTCAATG
GAGTTGGCTCCGGATACCAAAGCGGCCCGAGACGCCCTTGCAGAACGCAATCACGGCTTACA
CGGACGCGGGATGGAAGTCTAGAACAGCAGCAGTGACCTGGCAGCAGGGCGGCTCCTGGAGA
CATCACCTCATTGCAGCCGATGATAAGGACTCATTGCAAACATTTGAGCTGGTGGCCGTTGT
ATGGGCCATGATGAACTTAATCGACCCTCTTAATGTGGTCACCGACTCCCTTTATGTAGCTG
GAGTATGCCACCGAATAGAGGAGGCCTACATAAAGGAAGTGCAGAATCGGCGGCTGTACGAG
CTGTTCGTGCAGTTGCAGAGAGCAATCAGAATTAGGGAGCACTCATATGCAGTAATACATGT
TCGAGGTCATAAATGGGAGATAGACTTGGGGGAGGGAAATGCGAGAGCCGATCGCTTGGTGT
CGCTGGCACAGAGACCTTTAGTCTCCCAGCATGTCCTGGCCCGAGAGGCGCACTGTATGTTT
CACCAGAATGCCAAGGGGCTAAGAAGGGAATATCAGATAACATATGAGGACGCTAAGGCAAT
TGTTAGATCGTGCCCAGTGTGCAGCCACCATAATGGCGGTATGGGTCTCGGGCTAGGAGTTA
ACCCCAGGGGACTTAAAGCTAACGAAAATAGGCAGATGGATGTGACGCATGTGGGTGAGTTC
GGGCAGCTGAAATATGTACACGTGTCTATAGATACATATAGTCATTTTATGTGGGCCACCGC
TCAACCTGGAGAAAAGGCTATGTATGTAGAAAGGCATCTCAGTTGTTGCTTTGCAGTTATGG
GTATATCACTACAAATAAAAACAGACAGTGGCCCAGCCTACACTAGTCGTAGGCTTGGGGAG
TTTCTGCAAACATGGGGAGTAAAACACAGTACGGGTATCCCGAACTCGCCCACAGGCCAGGC
AATAGTAGAACAGGGAAATTGTGGATGGACGGCTGACCCAGCGAGGTCATGTTTTTCTCCCC
CTGGGTCAGTGCAAGTAAAACAGCTGCTATGGATCACTGTCTCACACAGAACAATGATGGGA
GTAGGCGAAAGGAATGCAGAGCTGCGATAGAGACCGCTGTGCCAGGTGCTGTGAGACAAGCT
GCAAGCAACTGCAAATTATGAGAACTGCGCCAGCGCTGCCCTCTGTTCTTCTGTCTCATTAT
ACTGCATGCTAATGGACTTTCAGGAACGTAATGAAAATGTATCTGAATTCCTGTCAACTATT
GATTATGTATTAGTTTGACCTATATCTATAGCACAATTTCTCCTGACGGTGTGCAAGTTAGG
TGGAGTGATCCCCCTTGCACTATCAATAGACATCAATAGACACATGCCTGCTCTATATCCTT
ACCGGATATAGGGTCTGATTTCCGCAACTCAATTTGCAACGAGAATAGATCCGCTCTGTTCG
GCTGCGGGATCGGTTGAGAGAGGGGACACCTTTGGCGCGCCCCAAGATTTCTTGGAAGGTCT
CCCTTCCTCACCCGATCACTGCAGGGACAGACAAGGACCATCCATTGGTGTATCAGGTATGT
GTATAGAAAAGGGGGTGCCCGTAAGGCGTGAGGAGACGTCCTACGCAGGGTATGAAGGAGCG
TGCCTGCTCCCCCCCCAGAGGTACGTGAGCTCACAGGTTAGGTTGCCGGCTGGGGTAGGACT
TGCAACTAATTGACTAATGTACTGGCGTTATTAAGTTTCAAGCTAAAAGACGTTGTCAGCAT
TGTGGATTATGTATAAACCATGTTTTTTGTTCTAAATATCAGAGTGAGTGTGGTGTGAGCGT

GTAAGGGAACGTGGACAAGCTTCTGTGAGATCTGTTCCCAAGTGTTTATAACTGAACAATAT
CACTATACGGTGTTAAGGGTTTGTGGAAACTTTGAGGGAATTAGGAATCAATAAGAATTTTG
TAAACATCATGGGAGGCTCCCAGGAGAAAGTTGTGTTTCCACAGGATCTGGCCCATTATGGG
TTCCCAGTAAATGGACTAAGCCAGCCCCAAACATTAGCAACCCATCTCAACCTGATGACAAC
AACAGCGACTCCGGCGAGCAATAGCCGCATGGTGTGCAAGACATCATATACAGTGCCATGAC
TCCCGAGCTGCAACTGGTATTCCCTTAGCAGATTGCAGAAAACATCTCATCAGAAAGGTCTC
TTGGACACTGGGAACTGCCAGGAGCAATTGGCGGTTACGAGACTCAGAAGCGCACTAATTGT
GGGGGTCGGGTACAGTGATTGATCCAGAGTGGGGACTGTGGTATTTTATTCATCGGGATATT
GAAACTGTAGTTAATACTGGTGTGCTGTGTAGTCTTTACTGTTAGTTTACTGACAAGGAATG
CTTTGTAGAAGTGTGTAACATAAATCTAGTGTATTCTTATAGATAGTGATTGGTATGCTGAT
AATCATGCTTTGTTTACTGTAATGCTAGCAAAAGATGGTTCAAGATGCAATTCATACAACCT
GGCTCGTACAAAAAGAAAACGGGGGAAATGTGGAGAGATTTATTGATGACTGGCTGACTGAG
AAAAGCCATGCAGACACTGTGCAGCTGGTTAAGCAACCTTGTTACGCGCAAGGTCAGGTATG
AGAACAAGGAACAAAACAGGATGCCGATGGAGAAAGCCGGCAACAGTGTTTTGAGAAGTCTG
TAAGAGAGTGTTTGCTGTGAGATAACCACAGAGCACAAAGATAGGCGTAGTTGACACTGACT
AGCGAGCCAATCCTGAGCTCCACTTTAGCAATATGTATGAGCCTATTATCCACACTATAAAT
GTACACGCAGCTGAAAGAATAAACGAGATTTGCTGATCATCTACTATGGTCGTTGCATTTCT
CCCGCTGATTTCCTACA

B. retroSaturnin

TGTGGAGAGCCTGTACACAGAGTGGCTACGTGACAAGGGCCACAATGGAATGCCACTCGTGA
AACAATAAAAGAAAGTATGTAACTAGAGGCAAGGACGGCTACTCGGGTAGATGGAAAACAGG
ATGCATATCTACAGCAATGAAGAAACAGTAACAGCAAGGCTAACCACGAAGGTCAAGATGAA
GTAATGCATAATCCGCCAATCCTGAGCTATGCTTTTGCAATATGTATAAGCTCATTACCTAC
TGTATAAATACAATACCAATGTGCCTAATAAATGAGACCTGTTGATCATGATAGCTTGAGAC
TCGTCTCCCGCGGTCTTTTCCGACAAATGGTGACCCCGACGTGATACAGACCGGGGAACTTT
GGCTGCCACGAACGGAGACTGCGTTGGGTTCGGGTCCCGCAGCTAGACGGACGGCGAAAGAG
TGAAATATTGGACTCCGCGCCTTGGGCGACCACAGCGGTGGGCTCAACCGATACGTGCTGCC
GAAGCTTGGGGGGCTGCAACAGCGCTGACGTAGGTAAGGATGGACAGGCAAGCAGCATATGA
TTTATTCACTGCCTTTCTGAGAAAGAGGCAATTCCGGGGGGTAGACTTAGAGAAGGAGCTCC
CTGCCCTGCTGAGTCATGCAGTATCACACGGGTTTTTCGCTGACCCTCACTTAGTACACACG

CTAGATGAGTGGCGACGGTATGGGGATAAGCTTTGGGAGTTGGTGTTGGAAGATGACAAGGT
AGCAAAAAGGATGAGTAAGTTGTGGAAGGCAGTGCATAACGAGTTGTTGATGGGTCAGGCGG
AGAAGAGGGCTGCGCAGGGAGCGCAGGCAGCTCAAGAAAAGAACAAAGATTATGGGTTGATC
TGGGATAACACTCCAAACCCCCCCTCCTTTTGCATGGTTATGCTACCCGCTGCTCCCCCCGC
ATCCGCAAATGGAGTGCAGGCTCAGTCAACCGCAGAGCCTGTGGAGGCACCCCCGCGTCGGG
TTTTAGGGCCATCAGCATTAGATGGTTCCGGGTCGGTTCCCGGTGCCGAGTCTGACCTTGCG
GAGGCTATAGCCAGGGAGAGACGGGAGTTGTGGACCACGTTAGCAAAACATGGGATGGAGGA
CGGAGATAATGAGCTTGTGCAGGCAGTGACTGAAAATCTTGCGTTCCCAGTTGTCTATACTC
CCACGCAGCAGGGCGGGTTACAAGCAGAAATTCGAACTTTAGATTGGAAAATGCTGTCCCAA
CTACGGGCTACAGTGGGGACATATGGGGTAACCAGTGAGCCGGCTAGACAGATGATGGAATA
TTTATTCAGTGCGCATATACTACTCCCGGCCGATATTAGAGGCATAGCTAGATTAATTTACA
CCCCCCATCAACTAATATTGTTTAACGCACACTGGCAGCAGGAAGCTGCAATATCGGCAGCA
GTGCAGAGGGGTCCGGGAGACCCTCTGGCAGGCATAACAATAGAACAGTTGATGGGTCTCGG
GGCATGGACTCGGATAGAGGCACAGGCAATTTCGGGACCTGATGTTTCCCGAGAAATAATGG
CTGTAGCCAGACGGGCTATGGACAAGATAAAGGCTCCTGGAGGTACACCTATTTACATGGGG
ATACGACAGGGGCGGGAAGAGCCGTTGGGTGATTTTGTGGATAAGGTTATGGATGCGATTAA
AAAGGCTAGCGTACCAGAGCACATGCAGGGGGCATTGCTTAAGCAGTGTGTGCTTCAGAATG
GGAACTCGAATACTCGGTCCCTTGTCAATACCCTGCCTGGGGATTGGTCGATCCCGGAGCTA
CTGGAAAAGGCTGCTACGGTCCCCTCGGGGACCCAGGCTTTCCTAGTGAATGCCCTGCAGAA
GATTGGGGATGGGTTACGGGAGCAGGCCAGAGCATTGCAGGAACAATCAAGGTCAGCTCAGA
CCCAGGTGTTAGCAGCCCTTGCGCCCCTCCAAGCTACCGTCTCCACTGGCCCCCACCCGCGC
GGCAACCCCCCGCATGAAGTGCTTTCGCTGTGGGAATGTCGGACACATCCGCCGTGAATGCCA
AGCAGATGGGGTCTGGTGCGGCAAGTGTCAATCGAACACCCACAACGCCAACGCCTGCCGCC
GAAAGTCGGGAAACGCGAGGAGCAGCGCGAACAGCAGCCGCGCCCGGACACAAGTAGCAGCT
GCAACATCGAACGCCCCCGGCAGCAACAACCCCAGCCTGCCACAAGCGGGAGCCTCGGCCTG
GACGTGGCAGCCTCAATAGACATTACCATCTTATCTAATCAACCCCACCGAATTGCTACCGG
ACTGTTTGGCCCTGTTATGATAAATGGACAAGCGGTGGGAGGGCTGCTGCTCGGAAGATCAT
CAGCCAGCATGTTAGGACTATTTGTCTTGCCTGGGGTCATAGATGCGGACTACCAAGGAGAA
ATTATGGTTATGGTGCATACCCCTTACCCGCCAATAAGGATTAACAAGGGGCAACGGATAGC
TCAGCTAGTTCCGCTGCCGCAACTAACGAAAGGGATGATACCCCTAAAGCAGGAACCGAGAG
AGCAAAAAGGATTCGGTTCGACTGGGGGACTAACTCTGCTCACTATTGACCTGTCAGATCGC
CCAAAGAAACCATGCACATTGTACTACCGAAATCAGAGCATTATTCTGACAGGATTATTGGA
CACGGGCGCAGATACATCGATCATTACCCCAGGAGAATGGCCGTCGGAATGGCCACTTCAAT

CGTCCACTACCACAGTGACCGGGGTTGGAGGAGTGACGTTAGCAAGCCGGACTCCCATGCTC
GCGGTAGAAATAGACGGCCGAAGAGCCACAGCTGTGTTTTCACTAGCCCTGCTTCCGCCCAC
AGTATCCTGCCTTATCGGTAGAGACGTGTTGGCTCAATTAGGCCTTGTCCTTACCAATGAAC
ACCCTTTAGGATAACGGCCATTGCTTGGACTTTCCCGCTGCCACTTACATGGACTACGAACA
ACCCGGTTGTGGTTAAGCAATGGCCACTAAAACGGGAAAGTCTTCTACAAGCACATCAGTTA
GTGCAAGAACAATATGCTCAGGGTCACCTGAAGTTGTCCACCAGCCCATGGAACACCCCAAT
ATTTGTGATTCAAAAGAAATCCGGGAAGTACCGTCTGCTGCACGACTTAAGAGCTGTTAATG
ATCAGATGGAGCCCATGGGGGCTCTTCAACCCGGCCTGCCTAATCCTGCCATGTTACCTGAA
GACTGGCCTATCCTTATCATCGACCTCAAGGACTGTTTTTTCACAATTGCTTTACACCCCCA
AGATACGAAACGTTTTGCCTTTACCTTACCGGCAATAAATAGGGGGGAGCCTGATAAGCGCT
TCGAATGGACAGTCCTACCACAGGGGATGCGTAACTCCCCAACTATTTGCCAGTTGTATGTT
GATGCTGCTTTACAACCGCTACGTAAAGAAATGCCCAACACTATCATCTACCATTACATGGA
TGACATTCTCTTCGCCCAGCAGGATCCGTTTACGGAGCAGCAGATTGAACGCGTCAAAACAG
TTCTAGCGGAATTCTCTTTGGTGGTTGCTCCTGAAAAGGTGCAAAGGTCTGCACCTTGGAAA
TATCTCGGATGGCAAATTACAGGGAAGCAAGTAACACCACAAAAATTGACTCTAGCTACTGA
CATCGCCACACTGAATGATGCACAAAAGCTCCTCGGAGACCTACAATGGTTGAAACCGGTCG
TAGGTCTCCCCAATACATTATTAGAAAAACTAATGCCCCTGCTAAAGGGTACAGATCCTTGC
ACCCCCATCTCCCTCACATTAGAGCAACAGGAAACCTTACAAGACATTACTGACTGCGTCTG
CAAGGGATTTGTCTCCCGCCTGGATCCTAACCTACCATTGGACCTTATGATTTGGAATAACA
GGACACACCTTCTTGGTGCAATCACTCAATTCCGAAAGAAAACGGGGGAGAGGGTGCTGGAA
TGGCTCTCGCCACCTCTTCAAAAACGAAAAACAATAACCACAAAAATTGAGAACCTTGCAGC
TGTTATTAGGAAAGGCCGCGTGCGAATATTGGAAGTCAGTGGTCGGGAGCCAGAGACTATTT
ATCTCCCAATGGAAAAGGCGACCTTGGATTGGTATTTGCTAAACTCGCTGGACTTAGCCGAA
GCACTCCTAGCCTCAGGGGCTAGGGTAAAGTCAGGTCCACTGGTCCCGCGAGCGCTGCAGTG
GCTTGGGAATGATTGCCAGACTGATCTCCAGTAGCTTTGCCGGTTTTGTTGCCCAGAAAGAA
AACGGGGGAAGTGTGGAGAGCCTGTACACAGAGTGGCTACGTGACAAGGGCCACAATGGAAT
GCCACTCGTGAAACAATAAAAGAAAGTATGTAACTAGAGGCAAGGACGGCTACTCGGGTAGA
TGGAAAACAGGATGCATATCTACAGCAATGAAGAAACAGTAACAGCAAGGCTAACCACGAAG
GTCAAGATGAAGTAATGCATAATCCGCCAATCCTGAGCTATGCTTTTGCAATATGTATAAGC
TCATTACCTACTGTATAAATACAATACCAATGTGCCTAATAAATGAGACCTGTTGATCATGA
TAGCTTGAGACTCGTCTCCCGCGGTCTTTTCCGACA

C. retroTux

TGCGGAATCAAACTCTATAACCCAGTTAAGTTTTAAAGCAGGTATGTTTACTCAGCCGGCCG
GGTGCAAGGAGGATCACTCCACCAAACTTGCACACTAGCAAGCAAAAACACTACAGATTTGT
AAGCCAAGTCCATACATATTCAGCAGAATTCTTGGAACCCATCTACATATTCATTATCCTTC
TGGAAATTCTGGAAATTTACATACTGCCCCTCCCCCCGCATGTGCGTTGTCCATAATTGGGG
TCTCCCTCAGGTGGTCATGGGGATGAAGTCAGAAGTCTTCCTCACAAAGGCTCAACATCTTC
CTTGGTCTGACCTTTTCACGGTGCATTCAAACTGGTCAAAGCCATTCATTTTGTATTGTTCC
TTTCTTTGTGTTTGCGTAGCTGACCATGAAATGTTGAGATGTCGTTCTCCTTCCCTAGCTTT
CAAAGCTAGCTTATTTAACCCTTATATGTTCAAACCCTAGACCCACCTGAGCTCCTGATCTC
AATCCCCCATTTTCTGAAGTCTAGCAAAACCTTGTGCTAGATGATCCCATTTTGAGAAAGAG
AATTCCTAAAATAACTCCCATTTTCCATCTTCTTTCTTAGCCTGTGTTTCTTCCAGTCCTCG
TAGCTATTCTCTGACCTTTACACAGCTATAACATGCATCGCAGAGAAATACATACTAAGATA
AGCAAAGCAATTACTATAATGCAAGCCATAAAGAGTTGCTTTAACCATGCAAAATTAGGTAT
CCAATTAGTGAGCTTTTTCCGTAATTCTCTAAAGCCAAAAGAAGTATCATCCAATGAGATTT
GGTGAAGTAACTTGGTACGTTCTCATGAGGCTTGCAGGTCAGTCTCTATTCTTGGAGTCTCG
TCTATATATAAACAACACCTCTGATTGATAATCATTCAATACTTCATCCCCCACCTGCATAG
AGATTTCCTTTTGTACAGTATAAGATTGCCCACAAAGAATCTCATGCGGGCTTAGTCCTTCT
TTGGTTCGGGTTTTAGTTCAAACTCTCAGAAGTGCCAAAGGCAGTGCCTGTGGCCAGGGAAT
CCCAGCCTCTTGGCCAAGTTTTACTATCTGCAATCTGATCAGGTAAATCATTTTTCCACCTG
TCCACTTAACTGCAGTCTATAAGGTGTATGCAATTGCCAAACAATTCCCAATAATGTGCTTA
TTTGTTGGACCACCTGGGCAAGAAAACGTGGGCCTCTATCTGATGGGACTGCTAAAGGAACC
CTGAACCTTGTAATTACTTCATATAACAGCACTTTCTTGGGCCTTATCAGTCCTACAAGGAA
ATACCTCTGGCCATCCTGAAAAGGTACCAGTTAGTACCAACATAAAGCGATACCCTCCTTTT
CTTGGGAGTTCCGAGAAATCAATTTGCCATTGTTGCCCTGGAAGGTTCCCTTTCCCAATAGG
CCCACTTGTACCCTGTTCCTGGTACTAGGATCGTACTTTGTAACGTCTGTTTGCCCAATACA
TTTTCTTATCCCCCCTTAGAACCAATTTCCATAGAATACTAAATGGCGCAACAAAGTGTCCG
TCTGGGTGCTTATCCATCTGTCAGCTTGCTCTCATCCCCTCTCTTTAGAATATTTTACAGAT
TCTGACTCAGGTTCAGAAATGTTAAGTTTACCGTCTGGAATTAAAGCCCTCTTCAACTACCT
GTTCTGCCACCTGATCCGCCATCTTATTTCCAGTTTCCTGTACAGTGTTACCTTTCTAATGT
CCTTTACAATGTATAATAGCCACACTCTTTGGTAGATTTACATAATGCTGTTTTTCCCTGTG
TGTGAGCAATCCTTGCTCTTTTCAGATGGTACTATAAGTGTGTAACATCCCATGTACATAAT
TAGCATCTGTTCATATATTTCACACTTTTGTCAAGGCAGCTATTTCAGCCTTCTGAGGGAAA

GTCCCCACAGGTAATGATCGTGCTTCGATTACCTTGTCAGTAGCAGCTAATGCATATCCTGC
CTTACGGTTTCCTTGTCTCACAAAACTGCTTCTGTCGGTAACCCAGGTGTCTTCTGCATCTA
AGGATTCATCTTTCAGGTCCGACTGGCAACAGTCCATAACTTCAGTTGTTTCAAAGCAGTCA
TGAAATTCTGGTTCTCCGGGAGTTCCGCTAAGGAAAGAGGCTGGGTTGACAATATTAGTTAC
AATTATTTCTGCATCATCCTGCTTTACCAGGATTGCCTGGTATTTCAGGAACCTTTGGGGGT
GAGAGCCAATGCCCCCCTTTTACCTCCAGTACAGTAGACACTGTATGAGATGTTAAAACTGC
TATTCATAAATTTACTGGCTTCTTGAATATTGAGCACTAGTGCAGCCACTGCCCTCAAACAC
TCTGGCCATCTCTTACCCACTTCATAAAGGGGCTTTACCAACAGTCCATAGCTGTGTATCCA
CAGTTGGCACCATCCTGTCATTCCTGAGAATGTATGGAGATCTTTCCTGGCAGTCCTTAAGA
TTCAAAGTCCAGCTGTAACCTCCTATCCCAGACACGTCACTTGCCGTTATTTGTGCCTTCTG
TTGAGAAACTTGATAACAAAGAAATCCAGCAAACTCACAGTCCATTACACACAGTTCTCTTT
TGTTTCGTGGCATCCATGCTCTGCAGTAAAGTTCCACCAGGGGATGGAGGGATCCAAATCTC
AGGCTCAAGAGTTGACTAGTTCTCAAAAATCATTGGGGTGTTCTCACAACCCTGGGTAACAC
TGCCCAGGTTAACTGAGTCTTTTTTTCAATATATTGGGGTTTTCTATTCAAAAGCAAATAAC
CTTAAGCTTCCTTTGGCCAAACCAGGCAGAAAAGGCATCCTTCAAATCTGGAAGGTAAACTC
CACCTACTCATTCCCAAACTTGGTTAATAAAGTGTATGGATTTACCACCATGGAGCGTACGC
ATTCAACGATTCTATTACTTGCTCTCAGATCTTGTACCAAACTATAGCACTTGCTATCCAAC
TTTTCATCAGCAAGACTGGAGTATTGTATTTGGATTCACATTCAATTAATAACCTGAAATCC
AGAAGTTATCTATTATTTTTTCCATCCTTCTACAGTTGCATATCCTCAAAGGGCACTGCTTT
ACCCTGACTGGGCTAGCTTTAGCTCTTGTTTAGAGCTGCATTTTGGCTCTACCAGGTATCCC
CGATACCCACACTCCTAAACCTACCTTATTTAAAATTTCTGTAACTTCTGGAGGTACAACAT
TGTTTTTCCCAGTTTGTGTTAAAGCCAGACTTAGAATTTTCAATCAGTTAATATTGTTTGAC
CCTTAGCTCCAGTTTTATTGCATTTAATCTCTGCATCTATTTGTTCTAATAAGTCTTACCAG
AATAACAGTTTCAGAGAACCTGTCATGTACAAGAATTTATGCATTCCTACTTATTTTCTCAG
CCTATATTTAATTGATCCCAAAAGGTAAGCTTTTTCTTGCTGGGCAGTAGCTCCTACTACTG
TTGCAAAATCATCTAAGAGTGGTAAGAGGTACCCGTATCCACCAGAAAATAAACCTCTTTAT
CCTGCTCTCCTAGCTGCAATTTAACCAGTGGATCTGATTTCTGTCCTCTTTTGTTCTGACAT
TCCTGTTACACCTTGCCTGCCTCAGCCTGCATCCTTCCCAAACTTTTCACTTTCCAAACCCT
CATAAACTACCAAAATTATATTTCAGTTTTACCTCCTCCTCTCCAAAGGCTTCCAGGGCTAA
CACTGGAATTTTGGGTTTCCTAGATCTCTGCAGATGAAGGCTATTCTCACTTCCAGTGCCCA
GTCCTCTCCACCTCCCCCCCCTTCACAGGTAACGCTGTTTACAACAGCAGTCTCTAACTCTG
GCTCCCTCTCACTCTTAGAGTTCTGCTCATGGGGGAACCACGGGGGAGGGTGGGGGGGTGTT
GCTCCCATGGCAGCGCAGATAAACTCACTTTCTCTTTTTCTCTTTGTCCACTATTCATACCT

TCAGTTACACTTCTGAATACCATTATACCACCAAAAACGTTTGTCCCAGTCCCTTCAGAACC
CATAAATCAAACAATATAGTTCTTCAGTTTCCATAAACACCCAGTACTTAGTCTCGTGATTA
GAGCACTCAATTATAAAAACTGTTCATGTTACTTCATTTACCTTTCCTCCTTAAGAATAACA
GTAACTGCAAGAAGGTATTATAGTTTAGGGTCCCATTCTGGGGCCACTTCTCATCATCTTCT
AATTTAAATATTAACCACCATTGATTGTAATACTTGATTAATGTTTTCTTACTCAGGGTGCC
TCCAGGTTCCCCTGCAGTATCCCACTGATATAGCAATACGCATCCCAAAGGGGTTTCTTTGG
GAATTTCCTTCCCCTGGGAGCCTCCCATGATGTTTACAAAATTCTTATTGGTTCTTTACTGC
CTTAACATTTCCACAGGTAACTTTCAATACTCTAAAATTTCCCCACACTAATCCCAAAAATA
CTCCATCCACGAGTCACAATGGTACTCTAGACATAAATGACACCACCAGCCCTGGTACACCA
CCTCACACACTAAACAAAGTATAACATAAATGGGCCAACATTCATTTGTTGGGTGGAATGTA
CACCTTTCCTCTGGTTCCTCTTCCCAATTTACAGAGTACAAACTCCCATTCTTATTCCAGTC
CATAAAATACATACAAACAGTGGACAGACATCACAAAGAGTCATCTTTCACTTTGTCTGTCT
CTGGCCGTTTCCCTCACGGGAGAACAGAACTGCGGATTGCGACTCCGCACACGCTCCGTGGC
TCTGCCACATCCCACTCACTCACACATACACACAAATGCTGACTTAAGTACAATCTGAGTTC
AAACACTTGAAAGAATGTTTTGAATGCATCAGTAGTTCAAATGCATGGTACAGAACATTTAC
AACAGATGGTAATACCACCTATGACTCCTACTATTATTGCCACAAAAAACCACTTAGCCATT
TAACAAAACGCACTGTTACCAACCACCTAGGAGAGATTTATTACTCCTTTTTTTCCACACGT
ACAACTTACAGGTCCCCCTAGGCGCAACCTAACCCATGAGCTCATTACCTCAGGGGCAGCTC
TATGGTTATACCCAGCATAGAACATCTCCTCACAACTTATGGGTTCCCCTTATTCCCATTCA
CATACCTGGTCCGCCGATAGATGGCCTTTGTCTATCCCTGCAGTGATCAAGAGAGGAAGGGG
AGTCCTTCCAAGAAACCTTGGGGTGCGCCTGAGGTATCCCCTCCTCAGCTGGTCCTGCAGCC
AGCCAGACAGTGTCCCATCTGGGGTGCCAAAATTGACATGCGGAATCAAACTCTATAACCCA
GTTAGGTTTTAAAGCAGGTATGTTTACTCAGCTGGCCGGGTGCAAGGAGGATCACTCCACCA
AACTTGCACACCAGCAAGCAAAAACACTACAGATTTGTAAGCCAAGTCCATACATATTCAGC
AGAATTCTTGGAACCCATCTACATATTCATTATCCTTCTGGAAATTTGTTTACATACTGCCC
CTCCCATGGTGCATGCATTGTCCGTAGTTGGGGTCTCCCTCAGGTGGTCATGGGGATGAAGT
CAGAAGTCTTCCTCACAAAGGCTCAACATCTTCCTTGGTCTGACCTTTTCACGGTGCATTCA
AACTGGTCAAAGCCATTCATTTTGTATTGTTCCTTTCTTTGTGTTTGCGTAGCTGACCATGA
AATGTTGAGATGTCGTTCTCCTTCCCTAGCTTTCAAAGCTAGCTTATTTAACCCTTATATGT
TCAAACCCTAGACCCACCTGAGCTCCTGATCTCA

Annexe 23 : Ré-annotation et re-découverte du modèle Galgal4 - Additional file 10

Additional file 10: Features of 8 repeat models that cannot be assigned to a known eukaryotic TE

Twenty seven consensuses gathered into 8 TE models (Table 5) had features that did not match with one of the three types described above or with any other known eukaryotic TE Seven of them (18 consensuses), Hitchcock and undetermined_group_1 to 6 (Additional file 7), were previously described as being solo LTRs [60]. In general, LTR elements had short inverted repeats at their ends, usually 5'-TG-3' and 5'-CA-3', but with extremely rare exceptions[28]. Solo LTR were flanked by a short 3-6 bp motif at both ends that varied in sequence from one locus to another. Here, copies of undetermined_group_1 to 6 had no 5'-TG-3' and 5'-CA-3' at ends. However, some of them had 3-6 bp repeats or relics at their ends. Together, this suggests that undetermined_group_1 to 6 were mobile sequences or were derived from them. However, the above evidence was not sufficient to assert that they were solo LTRs. With regard to undetermined_group_1, we observed that copies of this element displayed two hairpin loops on their ends. Such terminal structures have previously been described in some bacterial insertion sequences (IS), such as the IS with a HUH transposase[29]. Few loci with with potential HUH motifs were found in Galliforme species genomes with e-values less than 1 e-20. We considered that these data were not reliable because we failed to extend such similarities to other conserved motifs presents in HUH transposases.

The last nine consensuses were gathers into one model we named Z-rep, that describes repeats contained in a macro-satellite DNA mainly present at one end of the Z chromosomes and dispersed in the W chromosome. These nine consensuses described the ~22 kbp unit found in tandem in this satellite DNA. These were composed of a mix of old TE fragments originating from CR1 and LTR retroviruses. Interestingly, Z-rep sequences were only located in a single terminal block of 9 Mbp juxtaposed near a telomere involved in the lampbrush structure

28 Yin H1, Liu J, Xu Y, Liu X, Zhang S, et al. TARE1, a mutated Copia-like LTR retrotransposon followed by recent massive amplification in tomato. PLoS One. 2013;8:e68587.

29 Chandler M, de la Cruz F, Dyda F, Hickman AB, Moncalian G, et al. 2013. Breaking and joining single-stranded DNA: the HUH endonuclease superfamily. Nat Rev Microbiol. 11:525-538.

during the prophase I stage of meiosis[30], but they also spanned by spots over all the length of the chromosome Z (Figure S3), suggesting that this macro-satellite DNA might be a label of any parts of this chromosome.

Fig. S3 Dot plot representation of a pair sequence comparison calculated with the complete sequence of chromosome Z using Mummer[31]. All repeated segments along the chromosome Z were visualized by blue spots, excepted for those corresponding to Z_rep sequences that were highlighted by yellow spots. Horizontal and vertical axes were scaled in Mbp.

30 Andraszek K, Smalec E: Structure and functions of lampbrush chromosomes. BioTechnologia 2011, 4:337–344.

31 Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, et al. Versatile and open software for comparing large genomes. Genome Biol. 2004;5:R12.

Annexe 24 : Ré-annotation et re-découverte du modèle Galgal4 - Additional file 11

Additional File 11: Proteins encoded by the remnant polinton located in the RFJ and Turkey genomes.

>GG-Polinton-Integrase [chr2:79632317-79633476]
MYCNVRYLVEKDYKGRIGEISLVLESEKPSHQQLLFSPGQYNFSVHHLVDVSQLLLEDFHHS
PGRMESYHLFEVATKQIPSLNRSQVAIWLANQDAXSLGKATRKQCKRKELIVAEVKGRRQPQ
LVDVQQFSKQNGGGKXSFVVVGVLQKHAWAAGLKQKMGSEIVGGLXVIFIRGTXPQEMTAQG
KDFLKQPLKKLLQLYNNATTNKEHFSRTLKTKMXRYFIAHNAFCYSDTFSPFIKSYNYSFHR
ITTATEPGKVKPLKSMGEXIKIYSXSFKFNLNXRICTFPQKERPQSARKISEKLXAFIDXIF
IVAKTEKGDTCLLNRPQSQKNQRYLKKXIXIRTDYKIERIITTKGTGRKKELLVLLCGSLQI
FYSCTETSLYQDIQ

>GG-Polinton-MCP [chr2:79634258-79635332]
MNYAVQVSNNLLGFQAIPLTYRTAHTFCARKRNITYTLTAYRSYHCQNRSWKLLSQVLKTSL
KTLPKKKKKVFGYVSISVLVTPENDGEIKKRVRLTCTQKWPVPVFQKQKRKNQTAIWLRKKE
GKQCRQCSGLTARSGRTECLGPLHCLLFQDKLFLNSTDVTIKLMHSKDSICLMGSLATVACK
PVTACILLHVKKLWVALWLHLGQNEPLHVTVLNTWLTRRSWKHSTXLASSHVSNQENLFLGQ
LPKLHVIVFVDNDAFRGNYAXKPFNFKHKINILRIYTWLMMEPXQNHCSLIFKIGCCVKGYL
QLVQRISSRAKDHICXQPKGRIHPWLYPLCLWPISKSAMHQSLLIQLK

>Turkey-Polinton-DNApolymerase [chrZ30210529-30207921]|
[galGal4-chr2:79638800-79637500]
MDFIKLFMNMWMMAFQFLVHSAEGNDGYFVDSSKXGDNVQITQGGKLNCIEVLSMSTYFTDC
LFVSHDIKQLFQNLNIXLGRLFPLFFQHCXKIKVYVTTLSPLEYYEVDNVMLKEKTPLYDRK
TNQEKIFDLPKERCYYCQQDVQLLRGSCXXFMKXLMSIACYTFYEKKENENMTIYYCRDPFQ
HTTLFYTCLSRYRFMLLXNILSWKELTRIMKSNTWSIQNLNYFYELSPTASXTTAALGHDHF
PGSLFTLAEPFPNCHLTLLXDNSMPFPXVLSPSPESRVQHCPSTPLVRSFRPPXGFFSVSSA
LTXTQQGSSATPHILPSRPLPSTVLSNAFVFFILWHPKIPSAXSEATQLRAERDNHTPHPVA
VLGLAYPRGXLTLLAARTHSGLMFILSSTKTPRFLSAGLLSSILWLMYVAKKGKSXICLVLX
GTEHQIESXFLDSYMVIDMVLTAFKLNGCFFHGCIKCYDKNTWQGPHLIGAPKVLAVKTMRE

HVWTTFXWGDXRVFSXSLKRCLSAAGQVLCLYYKADIEEKYSYXKQPGLWHXQQRRSACVYL
VFLRTEQGSREELQTLXDLXESCFEXRVRNLFSDHIKLYLYQKHEASGXPASHRDDEENKXI
NKXMYXRLXAAGRCHFAAWARRKNSAIYKIAKLLLSCLWEKIQTEIQLSNTIVREQDELYCY
FFSPAFKVSSRDFTDKNIAILSXTLKEMXTRDFYDVLHLGIFVPXALCMPCCVYHISVIIMS
HLGARKDPSGNDSGLLMGELLXNEHITELASSSPKLYGYVRSGGRCCLKVKGISLTXXKARV
VTKIFKSLPYIIILGMHWKHISFVALEGSSNGR

Annexe 25 : Ré-annotation et re-découverte du modèle Galgal4 - Additional file 12

Additional File 12. Characteristics of the 54 neogenes derived from DNA transposons in the human genome and that of RJF (typed in blue).

Transposon family on origin	Related DNA transposon	Neogene names & synonyms in human	Protein size (aa)*	Acc. Number in the NBRF database	Chromosomal location in human	Chicken orthologs (NCBI or ENSEMBL)	Protein size (aa)*	Chromosomal location in chicken or taxonomic presence
Ginger	*Ginger1*	GIN1 (C5orf30)	522	NP_060146.2	5q21.1	XP_004949361	526	Z
	Ginger2	KRBA2	492	NP_998762.1	17p13.1			P only in Eutheria
	Ginger2-hobo	SCAND3 (ZBED9; Buster4 ; Charlie 10; ZNF452)	1325	NP_443155	6p21.33			P only in Theria
	hobo	ZBED1 (hDREF)	694	NP_004720	Xp22.33 & Yp11	XP_004938502	695	1
	hobo	ZBED2	218	NP_078784	3q13.2			P only in Eutheria
	hobo	ZBED3	234	NP_115743	5q13.3			P only in Eutheria
	hobo	ZBED4	1171	NP_055653	22q13.33	NM_001199541.1	1172	1
	hobo	ZBED5 (Charlie 1 ; Buster 1)	693	NP_001137139.1	11p15.3			P in Amniota, L in Sauropsides
	hobo	ZBED6	979	NP_001167579.1	1q32			P in Amniota, L in Birds
	hobo	GTF2IRD2 (Charlie 8)	949	NP_001003795.1	7q11.23	XP_004946491	1062	19
	hobo	ZBED8 (Buster3)	594	NP_071373.2	5q33.3			P in Chordata; L in Birds
IS630-Tc1-mariner	*Pogo*	Centromer protein B (CENP-B)	599	NP_001801.1	20p13			P only in Theria
	Pogo	Jerky (JRK)	568	NP_003715.2	8q24.3			P in Sauropsides, L in Birds
	Pogo	Jerky-like (JRKL	524	NP_001248762.1	11q21			P in Tetraopoda, L in Sauropsides
	Pogo	Tigger TE-derived 1 (TIGD1)	591	NP_663748	2q37.1			P in Boreoeutheria, L in Birds
	Pogo	Tigger TE-derived 2 (TIGD2)	525	NP_663761	4q22.1			P in Tetrapoda, L in Birds
	Pogo	Tigger TE-derived 3 (TIGD3)	471	NP_663771.1	11q13.1	ENSGALT00000043192	431	2
	Pogo	Tigger TE-derived 4 (TIGD4)	512	NP_663772	4q31.3	ENSTGUG00000005046		4
	Pogo	Tigger TE-derived 5 (TIGD5)	593	NP_116251.3	8q24.3	XP_004940149	544	2
	Pogo	Tigger TE-derived 6 (TIGD6)	521	NP_112215.1	5q32			P in Tetrapoda, L in Sauropsides
	Pogo	Tigger TE-derived 7 (TIGD7)	549	NP_149985	16p13.3			P only in Theria
	Tc2	POGK	609	NP_060012.3	1q24.1			Gene fossil in Chromosome 3 [GG4: 125200322 - 125201393]
	Tc2	POGZ	1410	NP_055915.2	1q21.3	XP_004948432.1	1386	25
	Hsmar1	SETMAR (Metnase)	684	NP_006506.3	3p26.1			Own to Primate
P	*P*	THAP0 (PRKRIR; DAP4)	761	NP_004696.2	11q13.5	XP_425679	891	1
	P	THAP1	213	NP_060575.1	8p11.21	XP_004949398	244	Z
	P	THAP2	228	NP_113623.1	12q21.1			P in Euteleostomi , L in Birds
	P	THAP3	175	NP_612359	1p36.31			P in Euteleostomi , L in Birds
	P	THAP4	577	NP_057047.3	2q37.3	XP_004937062	202	9
	P	THAP5	395	NP_001123947	7q31.1	XP_416027	413	1
	P	THAP6	222	NP_653322	4q21.1			P in Euteleostomi , L in Birds
	P	THAP7	309	NP_001008695.1	22q11.2	XP_001235650	267	Scaffold AADN03026676.1 [GG4: 79-2997]
	P	THAP8	274	NP_689871	19q13.12			P in Amniota, L in Birds

	P	THAP9 (Phsa)	903	NP_078948.3	4q21.22	XP_420555.4	857	4
	P	THAP10	257	NP_064532	15q23			P in Eutheria
	P	THAP11	314	NP_065190	16q22.1	ENSACAT00000001443	277	11
PIF-Harbinger	*Harbinger*	HARBI1	349	NP_776172	11p11.2	XP_421117	348	5
	Harbinger	nuclear apoptosis-inducing factor 1 (NAIF1)	327	NP_931045	9q34.11	XP_415512	330	17
piggyBac	*piggyBac*	PiggyBac-derived 1 (PGBD1)	809	NP_115896.1	6p22.1			P only in Theria
	piggyBac	PiggyBac-derived 2 (PGBD2)	341	NP_001017434.1	1q44			P only in Theria
	piggyBac	PiggyBac-derived 3 (PGBD3)	593	NP_736609	10q11			P only in Strepsirrhini
	piggyBac	PiggyBac-derived 4 (PGBD4)	585	NP_689808.2	15q14			P in Chordata; L in Sauria
	piggyBac	PiggyBac-derived 5 (PGBD5)	554	NP_078830.2	1q42.13	XP_004935545	409	3
Polinton-Marverik	*Polinton*	KRBA (KRAB-A1)	1030	NP_115923.2	7q36			P only in Theria
Transib	*Transib*	RAG1	1043	NP_000439.1	11p13	XP_421090	1041	5
Transib	*Transib*	RAG2	527	NP_001230715.1	11p13	AAA49052.1	528	5
Transposon family on origin	**Related DNA transposon**	**Neogene names & synonyms in human**	**Protein size (aa)***	**Acc. Number in the NBRF database**	**Chromosomal location in human**	**Chicken orthologs (NCBI or ENSEMBL)**	**Protein size (aa)***	**Chromosomal location in chicken or taxonomic presence**
Crypton	*Crypton*	KCTD1	865	NP_001136202	18q11.2	XP_004939868	267	2
	Crypton	KIAA1958a	744	NP_597722	9q32	XP_004949571	733	Z
	Crypton	QRICH1	776	NP_060200	3p21.31	XP_003642053	773	12
	Crypton	ZMYM2	1377	NP_003444	13q11-q12	XP_004938872	732	1
	Crypton	ZMYM3	495	NP_001164634	Xq13.1	XP_420201	1433	4
	Crypton	ZMYM4	1548	NP_005086	1p32-p34	XP_004947829	1514	23
	Crypton	ATF7IP	1270	NP_060649	12p13.1	XP_416205	1085	1
Zirupton	*Zisupton*	HMGXB3	1106	NP_002600	5q33.1	XP_414598	1453	13

*, sizes in amino acids (a.a.) were calculated from the largest ORF found in mRNA; **, DBD = DNA binding domain and CatD = catalytic domain. ?, located domains with no definitely characterized function; ***, the presence of neogenic orthologues in databases was checked using BLASP and genome browsers at http://blast.ncbi.nlm.nih.gov.gate1.inist.fr/Blast.cgi?PROGRAM=blastp&BLAST_PROGRAMS=blastp&PAGE_TYPE=BlastSearch&SHOW_DEFAULTS=on&LINK_LOC=blasthome,http://genome.ucsc.edu/, http://www.dyogen.ens.fr/genomicus-67.01/cgi-bin/search.pl. Related neogenes are listed in families and alternately shaded in grey or unshaded. In the last Column, P =Presence and L : Lost.

Annexe 26 : Ré-annotation et re-découverte du modèle Galgal4 - Additional file 13

Additional File 13 : Inventory of the number RM annotations that had no equivalent in the [TE+DM] annotation

RepeatMasker consensuses highlighted in yellow were consensus sequences corresponding to repeated genes coding structural RNA. Those highlighted in grey and blue corresponded to Satellite DNA repeats and consensuses belonging to one of the 24 TE models calculated with REPET, respectively. The 171 remaining consensuses that were not highlighted had no equivalent in our 34 models. In lane 246 of the table are indicated in red the averages for all consensuses unique to the Repbase and ISB libraries.

Repbase and ISB consensuses	Number of copies	Size of the smallest element	Size of the largest element	Averaged percentage of sequence divergence of each annotation with its consensus
AmnSINE1	1528	1	510	27
AmnSINE2	129	45	265	27
CR1-11_Crp	2751	1	1361	28
CR1-12_AMi	85	1	608	27
CR1-13_AMi	1015	1	674	26
CR1-16_AMi	349	1	659	26
CR1-1_Amn	410	1	476	26
CR1-3_Croc	608	1	554	29
CR1-8_Crp	564	1	1007	27
CR1-B	145	1	4951	11
CR1-B2	398	1	1062	11
CR1-C	629	1	3052	13
CR1-C2	313	1	2169	12
CR1-C3	640	1	1845	15
CR1-C4	1855	1	3170	17
CR1-D	551	1	1961	14
CR1-D2	1372	1	955	15
CR1-E	1000	1	2019	15
CR1-F	366	1	3105	19
CR1-F0	316	1	975	19
CR1-F2	2039	1	1107	20
CR1-G	327	1	2559	14
CR1-H	492	1	2646	13
CR1-H2	450	1	980	16
CR1-L3A_Croc	2195	1	1651	30

CR1-L3B_Croc	3582	1	2197	28
CR1-X	1158	1	1183	18
CR1-X1	950	1	561	16
CR1-X2	2029	1	806	19
CR1-Y	400	1	2310	21
CR1-Y1_Aves	2601	1	1333	22
CR1-Y2	455	1	1816	26
CR1-Y2_Aves	4118	1	1480	27
CR1-Y3	423	1	866	14
CR1-Y4	1148	1	3576	18
CR1AVI	13	2	122	20
CRP1	95	10	234	28
Char12Mar1_GG	488	1	1398	16
Charlie12_GG	97	1	748	18
Charlie12_GGa	36	1	914	18
Charlie7_Aves	82	5	1780	22
Charlie7a_Aves	82	18	319	20
Chompy-11C_Croc	291	2	289	30
Chompy-11I_Croc	97	29	302	31
Chompy-1_Croc	788	2	87	22
Chompy-2353_AMi	252	11	255	25
Chompy-2_Croc	575	5	80	23
Chompy-334_AMi	57	38	276	31
Chompy-3_Croc	344	6	131	24
Chompy-5_Croc	145	42	112	26
Chompy-6_Croc	745	1	556	30
Chompy-7_Croc	347	11	406	30
DIRS-1_AMi	44	31	348	23
DNA-12_Crp	14	15	342	25
EAVHP_I-int	24	2	1768	15
ENS1-LTR	34	1	1013	6
ENS1-int	25	4	1346	11
ENS1B-LTR	50	1	934	11
ENS3-int	120	1	1871	19
Eulor1	69	35	337	25
Eulor10	50	38	281	26
Eulor11	127	8	502	25
Eulor12	57	51	173	27
Eulor2A	73	55	305	26
Eulor2B	96	14	279	25
Eulor2C	68	47	173	27
Eulor3	48	61	272	26
Eulor4	41	46	469	26
Eulor5A	155	15	318	27
Eulor5B	54	42	191	26
Eulor6A	28	55	240	26
Eulor6B	117	13	352	25
Eulor6C	24	19	187	26

Eulor6D	108	43	308	28
Eulor6E	39	47	274	28
Eulor7	17	45	197	25
Eulor8	197	52	533	28
Eulor9A	106	40	287	26
Eulor9B	13	57	168	28
Eulor9C	250	11	271	24
GGERV10_LTR	104	1	224	13
GGERV11_LTR	81	5	103	26
GGERV20_I-int	82	1	5385	12
GGERV21_LTR	436	1	318	17
GGERV22_LTR	910	1	349	18
GGERV28_I-int	175	1	3319	28
GGERV28_LTR	25	2	547	9
GGERV30_LTR	60	1	331	9
GGERVK1-int	42	2	2626	8
GGERVK10-int	49	1	3023	21
GGERVL-A-int	165	1	5991	16
GGERVL-B-int	175	1	5339	15
GGERVL-C-int	87	1	5059	16
GGERVL18-LTR	379	1	346	14
GGERVL18-int	123	1	5280	17
GGLTR1	5	22	241	13
GGLTR10A	3	270	290	3
GGLTR10B	4	69	248	1
GGLTR10C2	19	1	156	13
GGLTR10D-int	12	1	807	5
GGLTR10C-int	8	4	2611	7
GGLTR10C1	7	1	288	11
GGLTR10D	9	10	347	12
GGLTR11	143	1	539	4
GGLTR11-int	396	1	2698	17
GGLTR12A	7	1	782	20
GGLTR12B	21	1	216	9
GGLTR12C	66	1	674	15
GGLTR3A	43	1	582	15
GGLTR3B1	24	1	525	6
GGLTR3B2	11	1	570	9
GGLTR3B3	8	1	502	12
GGLTR3B4	34	1	551	11
GGLTR3C1	33	1	530	12
GGLTR3C2	45	1	620	16
GGLTR3D	73	1	664	15
GGLTR3E1	38	2	382	9
GGLTR3E2	38	2	582	14
GGLTR3E3	5	1	65	10
GGLTR3F1	23	1	571	10
GGLTR3F2	33	1	518	3

GGLTR3G1	37	1	517	6
GGLTR3G2	41	1	584	13
GGLTR4A	34	1	608	17
GGLTR4B	2177	1	346	15
GGLTR5A	521	1	328	12
GGLTR5B	456	1	590	22
GGLTR6	232	1	544	18
GGLTR7-int	165	1	1321	18
GGLTR7A	518	1	6483	15
GGLTR7B	147	1	621	7
GGLTR8A	62	1	597	18
GGLTR8B	1197	1	995	19
GGLTR9	2366	1	1023	23
Gypsy-5_AMi-I	611	1	363	17
Gypsy-5_AMi-LTR	23	54	561	29
HY1	1	111	111	13
HY3	1	100	100	6
Harbinger-N1_Croc	604	3	100	14
Harbinger-N276_AMi	333	29	132	26
Harbinger-N96_AMi	101	23	88	25
IS1	1	76	76	23
IS10	1	98	98	24
IS186	1	60	60	20
IS30	2	42	125	23
L2-1_AMi	404	2	951	20
L2-1_Crp	40	36	1006	28
L2-3_AMi	50	48	378	28
L2-3_Crp	670	1	880	24
LFSINE_Vert	1458	4	443	26
LmeSINE1c	3	80	176	17
MER121B	49	34	360	31
MER123	46	52	275	25
MER125	196	33	194	25
MER126	304	5	456	24
MER127	181	23	358	25
MER129	75	59	378	26
MER130	323	24	883	29
MER131	549	5	219	30
MER132	79	39	186	26
MER133A	78	48	120	27
MER133B	97	9	126	27
MER134	109	26	434	28
MER136	34	63	313	25
MIR1_Amn	1153	3	269	24
MIR_Aves1	1273	5	246	26
MIR_Aves2	747	14	248	26

MamRTE1	18	44	518	24
Mariner1_GG	123	1	1310	30
Mariner1b_GG	64	1	554	16
OldhAT1	478	8	1093	16
Penelope1_Vert	126	3	614	28
RSV-LTR	1	18	18	15
RSV-int	14	1	576	28
TguLTR5d	128	1	240	12
TguLTR5e	811	1	580	24
Tn1000	4	41	83	25
UCON1	373	2	491	14
UCON100	24	44	119	25
UCON103	1	36	36	22
UCON11	71	27	420	8
UCON12	112	51	418	23
UCON12A	43	53	353	23
UCON14	122	40	251	23
UCON15	56	3	295	25
UCON16	50	36	216	26
UCON17	42	57	320	24
UCON18	40	48	307	28
UCON19	32	56	312	25
UCON2	252	3	537	26
UCON20	53	55	501	27
UCON21	26	53	240	25
UCON22	74	45	279	28
UCON23	30	7	173	26
UCON24	26	38	370	22
UCON25	40	39	190	24
UCON26	277	9	309	27
UCON27	98	19	616	29
UCON28a	127	38	608	26
UCON28b	77	56	495	27
UCON28c	75	51	587	29
UCON29	723	4	424	26
UCON31	188	16	395	27
UCON4	168	3	306	27
UCON49	108	49	356	24
UCON5	145	26	390	25
UCON51	73	17	251	25
UCON56	19	35	144	26
UCON57	31	37	88	23
UCON6	147	12	385	21
UCON60	6	87	116	26
UCON61	14	64	139	25
UCON62	29	50	181	19
UCON63	12	55	156	27
UCON64	302	38	513	23

UCON65	38	70	337	28
UCON66	97	5	432	30
UCON67	27	31	119	29
UCON68	13	36	92	19
UCON69	445	9	347	16
UCON7	138	30	347	26
UCON70	53	14	247	24
UCON71	4	66	177	23
UCON71_Crp	32	44	271	24
UCON75	11	99	191	24
UCON78	252	10	428	20
UCON8	398	12	810	25
UCON80	35	32	137	27
UCON80_AMi	70	24	194	25
UCON84	139	32	600	25
UCON86	64	40	541	26
UCON89	74	2	271	27
UCON9	26	4	243	26
UCON92	12	8	130	25
UCON96	15	18	150	23
UCON97	66	40	297	22
UCON99	56	40	303	25
X13_LINE	81	14	536	26
X1_LINE	119	29	585	27
X2_LINE	335	1	553	26
X5B_LINE	13	59	197	28
X7A_LINE	17	13	215	24
X7B_LINE	29	42	196	28
X7C_LINE	34	2	203	25
X8_LINE	28	74	327	29
X9_LINE	71	63	507	25
hAT-16_Crp	18	30	798	29
hAT6-N1_Croc	841	4	356	23
Averages	**302,53**	**21,09**	**755,15**	**21,19**

GGCAN	54	3	354	9
REP131	4498	1	1577	26
5S	23	9	121	27
LSU-rRNA_Cel	3	21	70	28
LSU-rRNA_Hsa	102	1	1498	14
SSU-rRNA_Hsa	29	10	1161	20
U1	42	3	165	18
U13	1	93	93	21
U14	3	91	98	15
U17	1	202	202	14
U2	35	3	189	20
U3	4	33	214	17
U4	7	57	144	26

U5	13	7	116	17
U6	27	5	107	17
tRNA-Ala-GCA	1	42	42	28
tRNA-Ala-GCG	4	51	73	19
tRNA-Ala-GCY	23	26	74	8
tRNA-Ala-GCY_	2	64	73	7
tRNA-Arg-AGA	3	69	94	13
tRNA-Arg-AGG	7	2	72	6
tRNA-Arg-CGA	1	72	72	7
tRNA-Arg-CGG	1	73	73	8
tRNA-Arg-CGY_	5	42	73	4
tRNA-Asn-AAC	10	40	75	5
tRNA-Asp-GAY	10	43	75	2
tRNA-Cys-TGY	11	72	75	6
tRNA-Gln-CAA	2	55	74	4
tRNA-Gln-CAG	5	72	72	11
tRNA-Glu-GAA	1	72	72	6
tRNA-Glu-GAG_	15	3	72	1
tRNA-Gly-GGA	7	56	72	10
tRNA-Gly-GGY	8	71	74	2
tRNA-His-CAY_	10	67	74	2
tRNA-Ile-ATA	2	96	96	10
tRNA-Ile-ATC	1	63	63	0
tRNA-Ile-ATT	7	2	27	22
tRNA-Leu-CTA	1	82	82	3
tRNA-Leu-CTG	7	14	83	0
tRNA-Leu-CTY	4	82	85	11
tRNA-Leu-TTA	2	83	83	2
tRNA-Leu-TTG	5	36	106	14
tRNA-Lys-AAA	6	73	75	5
tRNA-Lys-AAG	11	2	74	6
tRNA-Met	2	73	76	9
tRNA-Met-i	2	73	73	7
tRNA-Met_	13	9	75	7
tRNA-Phe-TTY	5	73	73	10
tRNA-Pro-CCA	5	42	74	8
tRNA-Pro-CCG	5	39	74	5
tRNA-Pro-CCY	8	32	73	7
tRNA-SeC(e)-TGA	3	76	86	6
tRNA-Ser-AGY	7	34	85	16
tRNA-Ser-TCA	2	67	82	6
tRNA-Ser-TCA_	1	83	83	14
tRNA-Ser-TCG	5	82	82	26
tRNA-Ser-TCY	7	6	82	9
tRNA-Thr-ACA	5	73	75	8
tRNA-Thr-ACG	1	72	72	4
tRNA-Thr-ACG_	1	52	52	6
tRNA-Thr-ACY	3	74	74	17

tRNA-Trp-TGG	6	72	75	10
tRNA-Tyr-TAC	13	11	90	5
tRNA-Val-GTA	1	73	73	2
tRNA-Val-GTG	7	66	76	2
tRNA-Val-GTY	6	73	76	8
tRNA-Val-GTY	54	3	354	4

Annexe 27 : Ré-annotation et re-découverte du modèle Galgal4 - Additional file 14

Additional file 14: Origins of differences between ISB and [TE+DM] annotations.

171 Repbase and ISB TE consensuses involved in the ISB annotation had no equivalent in our TE models. The origins of these differences were investigated in order to: i) verify how loci were annotated in both annotations and ii) analyze the quality of the ISB annotations that were devoid of an annotation by our procedure.

The identities and differences between the [TE+DM] and the ISB annotations were first investigated locus by locus using bedtools. A large majority (Figure 7a, 85.4%) of the most abundant previously characterized TEs (CR1, Galluhop, Kronos, Charlie, BIRDDAWG, etc … [60]), the [TE+DM] annotations and the ISB annotations were identical. Nevertheless, 7.4% of ISB annotations did not match with those of the corresponding model in the [TE+DM] annotations (Figure S4a) and 7.3% of the ISB annotations had no corresponding matches in the de novo [TE+DM] annotations and were annotated with one of the 171 Repbase and ISB TE consensuses (Figure S4a).

Loci with differing annotations were enumerated (72,090 loci) and 40 loci were selected for manual annotation. Three types of result were obtained, five examples are presented in Figures S4b to d. The analysis of loci summarized in Figures S4b1, b2 and b3 illustrate 3 cases where the RM annotations did not match those of the corresponding model in the de novo [TE+DM] annotation. These cases are representatives of the majority (75 %) of those that were sampled. These surveys revealed that the ISB annotations were less precise for describing TE copies than those calculated in the de novo [TE+DM] annotations. Indeed, they did not take into account small TEs or pieces of TE inserted into larger TEs. Although more rare, annotations that were erroneous and/or ambiguous were also found in the [TE+DM] annotations. Figure 8C illustrates a case of erroneously annotated loci in the de novo [TE+DM] annotation. Because the ENS3-Int consensus described a non repeated sequence it was not identified by REPET. Figure 8D describes a locus that appeared similar to the case described in Figure S4c but for which a detailed analysis of sequence similarities with the involved consensuses did not allow us to make a decision. Therefore, our investigations indicated that the library dependent steps in both approaches (RM versus TEannot) remained

a delicate step for which the quality of the annotation fully depended on the supplied sequence library.

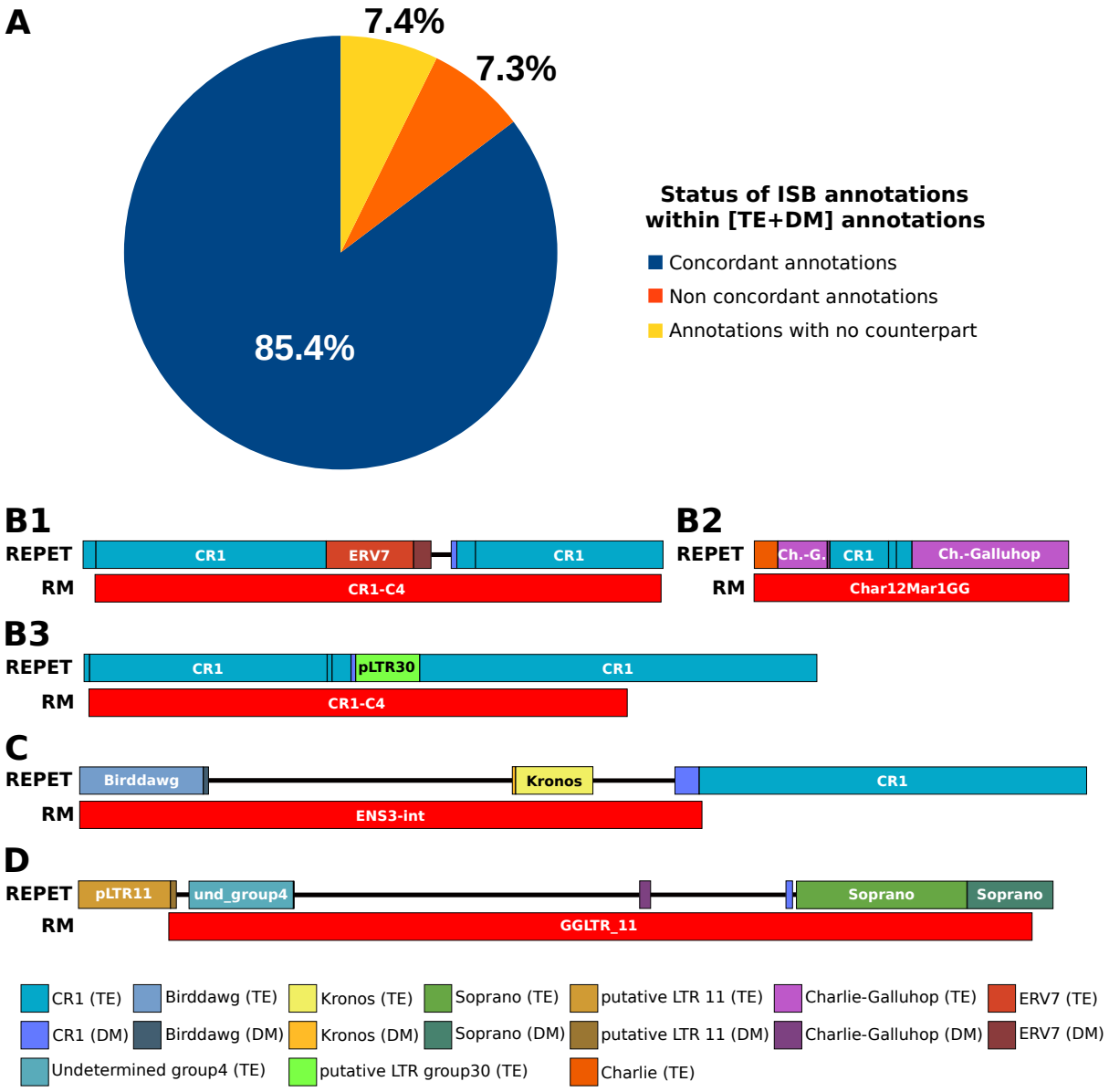


Fig. S4 Quality of [TE+DM] and ISB annotations. A, pie chart representing the correspondences of the ISB annotations within the [TE+DM] annotations. Slices corresponded to coverage percentages of ISB annotations that were identical to those of the [TE+DM] annotation (blue), did not matched with those of the corresponding model in the [TE+DM] annotations (orange), or had no corresponding annotation in the the [TE+DM] annotations (yellow) and were annotated with one of the 171 Repbase and ISB TE consensuses own to the ISB annotation (yellow). B to D, Examples of five loci differentially annotated in the de novo TE+DM and ISB annotations. For each locus, the positions were supplied for the REPET and RM annotations: B1, chr2:54915707..54918245 / chr2:54915760..54918231, B2, chr2:87534124..87535508 /

chr2:87534124..87535507, B3, chr1:157232263..157234613 / chr1:157232241..157235461, C, chr8:8963651..8968051 / chr8:8963650..8966367, D, chrZ: chrZ:29482798..29486975 / 29483193..29486961. Annotations resulting from REPET annotations and the DM annotations were respectively indicated by (TE) and (DM). This allowed visualizing to which extend the DM annotations enlarged those made by REPET in these five examples. Back vertical lines in CR1 REPET annotation in B1, B2 and B3 indicated that the different regions were annotated by different consensuses belonging to the CR1 model.

The [TE+DM] and ISB annotations share 8.94% of coverage and 0.8 % are specific of the ISB and resulted from annotations performed with one of the 171 Repbase and ISB TE consensuses. Among them, ISB annotations for genes coding structural RNA covered 0.04%. The remaining 0.76 % corresponded to annotations of TEs such as LINEs, LTR, Eulor, UCON, MER, SINEs, IS and DNA transposon that were not detected by REPET as being repeats, because of their sequence divergence (sequence divergence with their Repbase and ISB consensuses ranged from 20 to 40%; Additional file 13) or their size (311 RM annotation are smaller than 20bp). Forty percent of TE RM annotations have sequence divergence from their consensus of 20% or more (124,555 loci ranging from 10 to 5625bp). To create a consensus REPET needs at least 3 similar copies (>95 %) and it tends to be less efficient with repeats inactivated for mobility for a long time.

The consistency of the RM annotations made from 171 Repbase and ISB TE consensuses was investigated and focussed on a sampling that gathered consensuses annotating SINE element (AmnSINE1 (574 bp), AMnSINE2 (358 bp), LMESINE1c (404 bp), MER129 (469 bp), MER130 (475 bp) and MER131 (173 bp)), Harbinger transposons (Harbinger-N1_Croc (88 pb), Harbinger-N276_Ami (110 bp), Harbinger-N96_Ami (79 bp)), and Kolobok transposons (UCON29 (437 bp) and UCON2 (280 bp)). Annotated copies with each of these consensuses were extracted into a single file into which their Repbase consensus were added to help drive their alignment using Muscle. The result of these analyses was that it was not possible to consistently align the extracted copies because numerous copies were too small in size (Additional file 13) and were at least 30 to 50 % divergent from each other (although they were all 20 to 30 % similar to their consensus). Together, the analysis of these cases (4298 RM annotations) led us conclude that a significant part of these annotations are likely artefacts inherent to library-based annotation methods. Regarding SINEs, the conclusion of our investigations agreed with recent results that support avian CR1 mediated transposition involved specific nucleic structure in CR1 which near perfectly hampered the

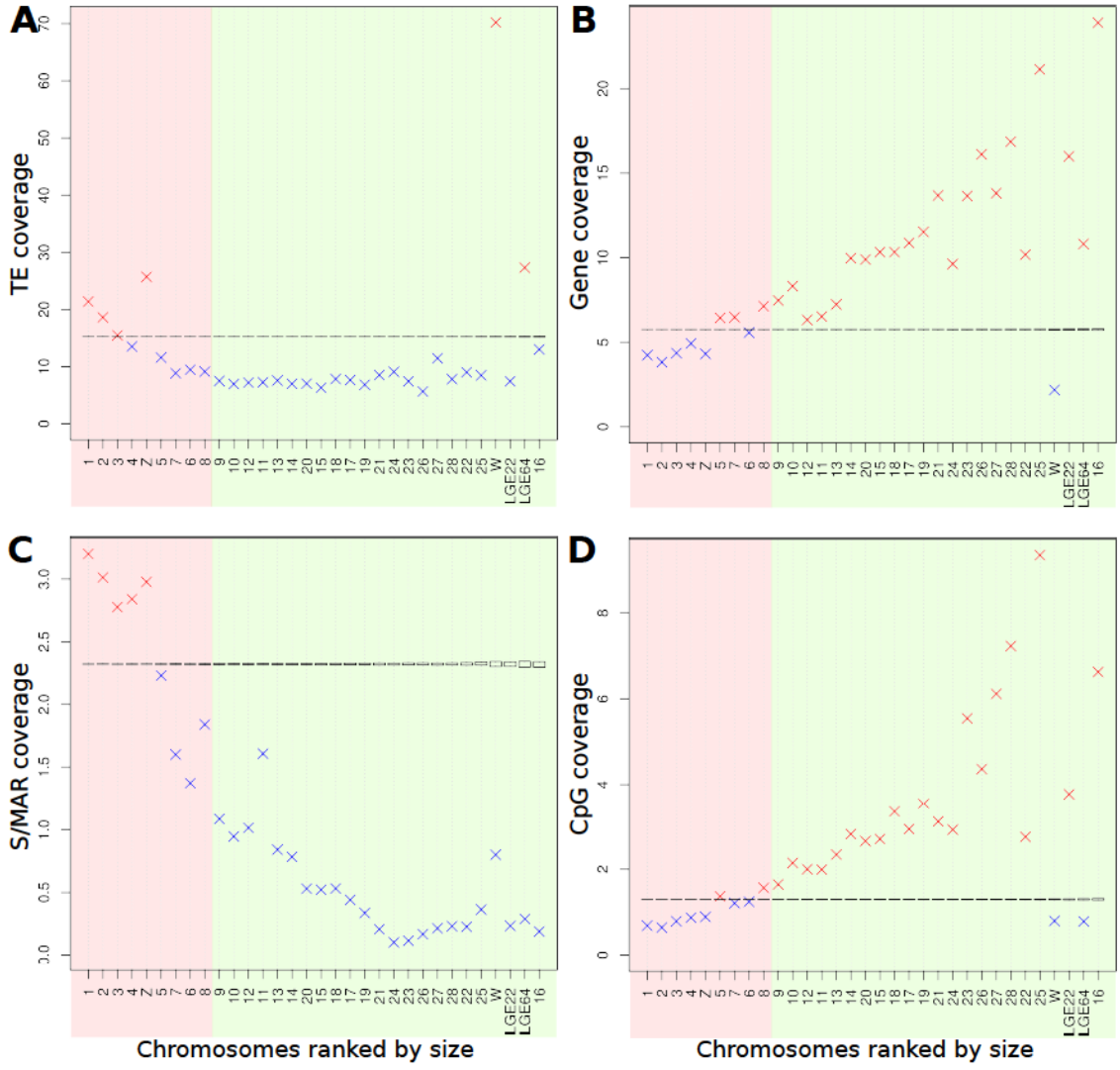
retrotransposition of pseudogenes[32]. Another issue was that a part of the consensuses (in our sampling LFSINE_vert (SINE) and Penelope1_Vert (non-LTR retrotransposons) corresponding to those defined by the ISB were not publicly accessible. Therefore, the quality of their annotations could not be verified.

32 Suh A. The Specific Requirements for CR1 Retrotransposition Explain the Scarcity of Retrogenes in Birds. J Mol Evol. 2015;81:18-20.

Annexe 28 : Ré-annotation et re-découverte du modèle Galgal4 - Additional file 15

Additional File 15: Graph showing the expected and observed coverages of TEs (A), genes (B), S/MAR (C) and CpG islands (D) in galGal 4 chromosomes.

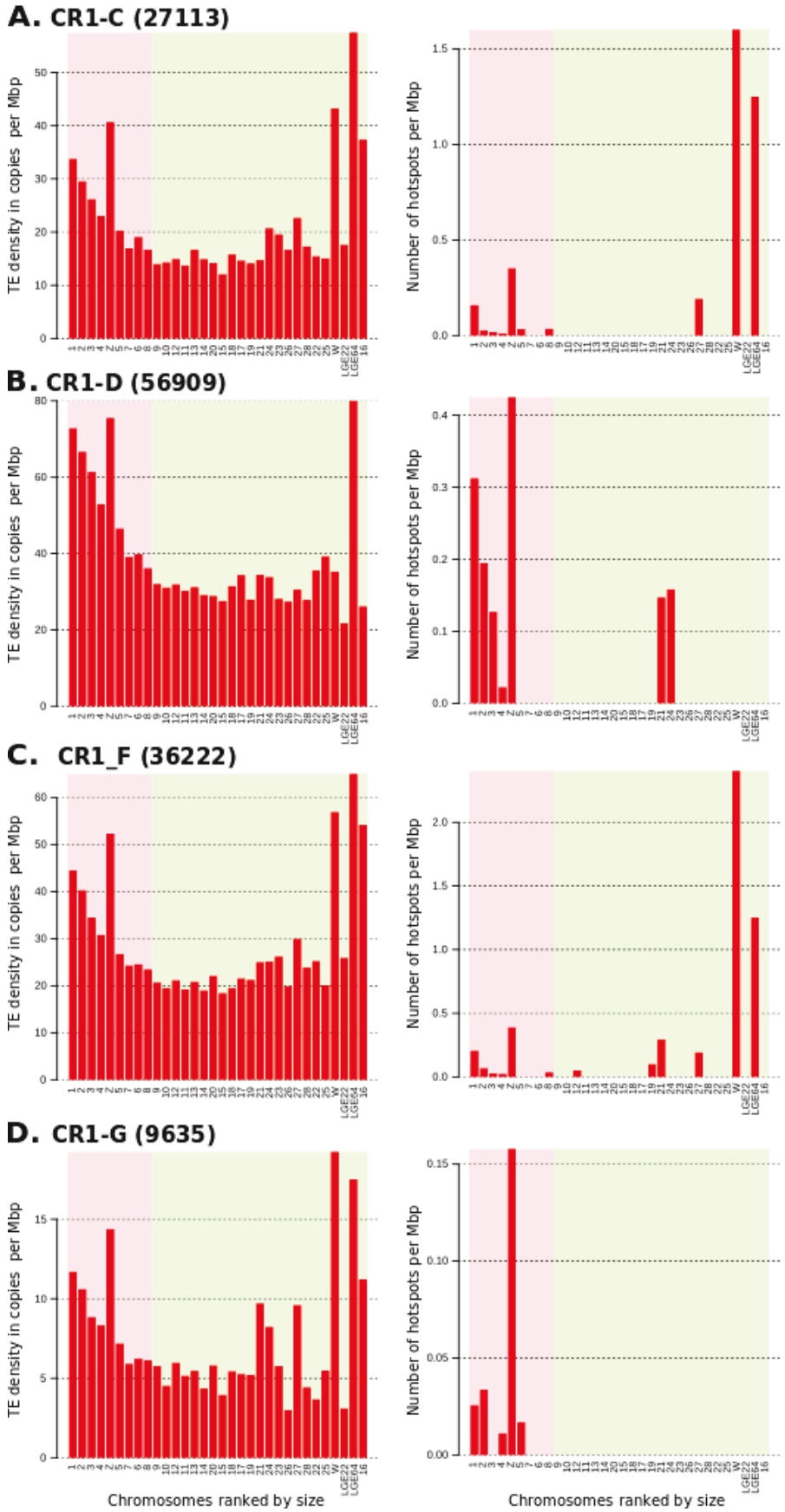
Each box was calculated from 1000 permutations and represents 98% of distributions obtained per chance. Red crosses above the boxes indicate over-representation of the element in the chromosome and blue crosses an under-representation (p>99% in both cases). Backgrounds in pink indicate macrochromosomes and those in green indicate microchromosomes. All galGal4 chromosomes, except chromosome 32 (only 1028 bases) were included in the analysis.

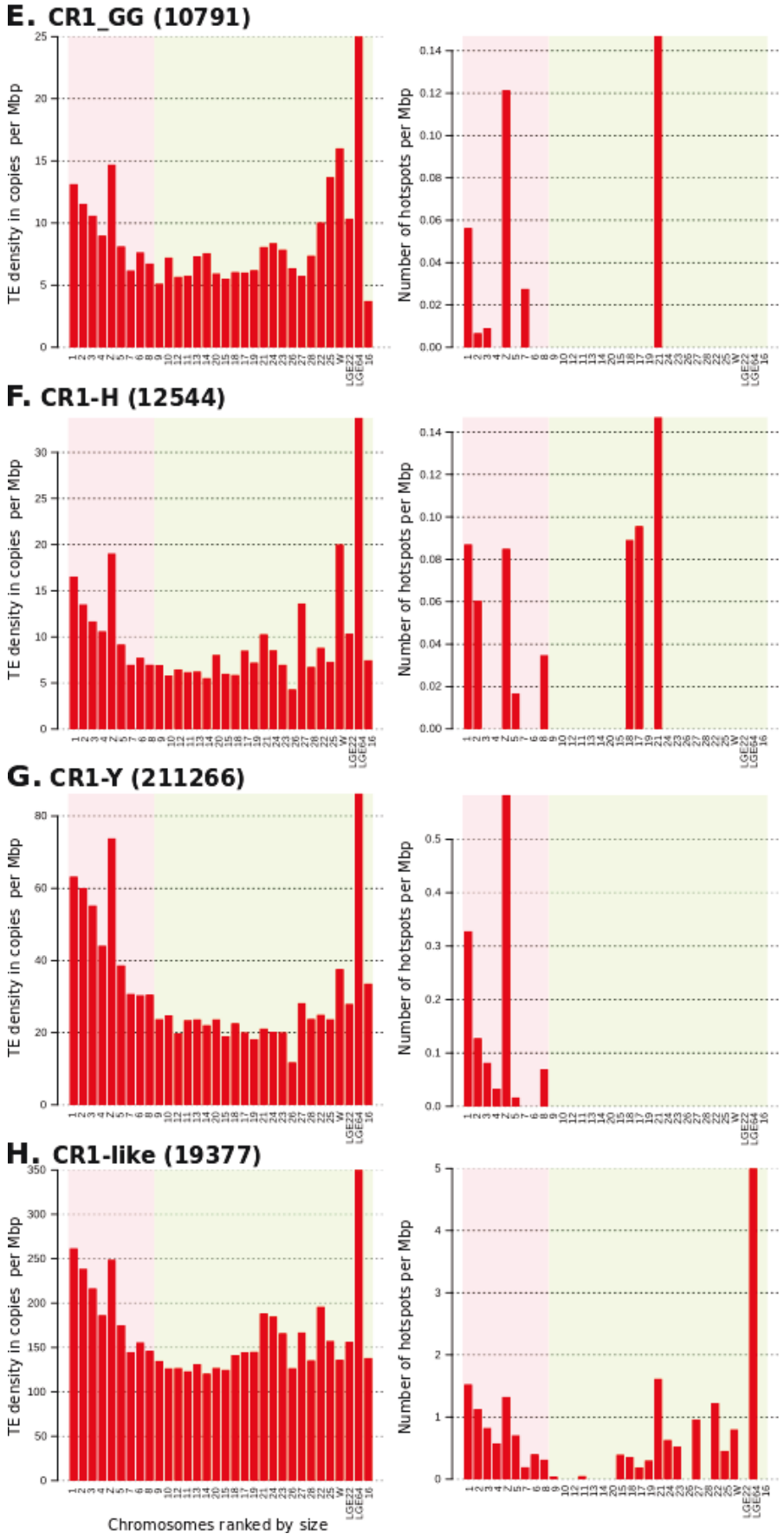


Annexe 29 : Ré-annotation et re-découverte du modèle Galgal4 - Additional file 16

Additional File 16: Histograms showing the densities of TEs and TE hot spots in galGal4 chromosomes for the 8 sub-families of CR1 elements.

Histograms of TE model density (left column) and TE hot spot density (right column) were calculated for all galGal4 chromosomes, except chromosome 32 (too small; 1028 bp). The number of copies for each dataset are indicated in parentheses.

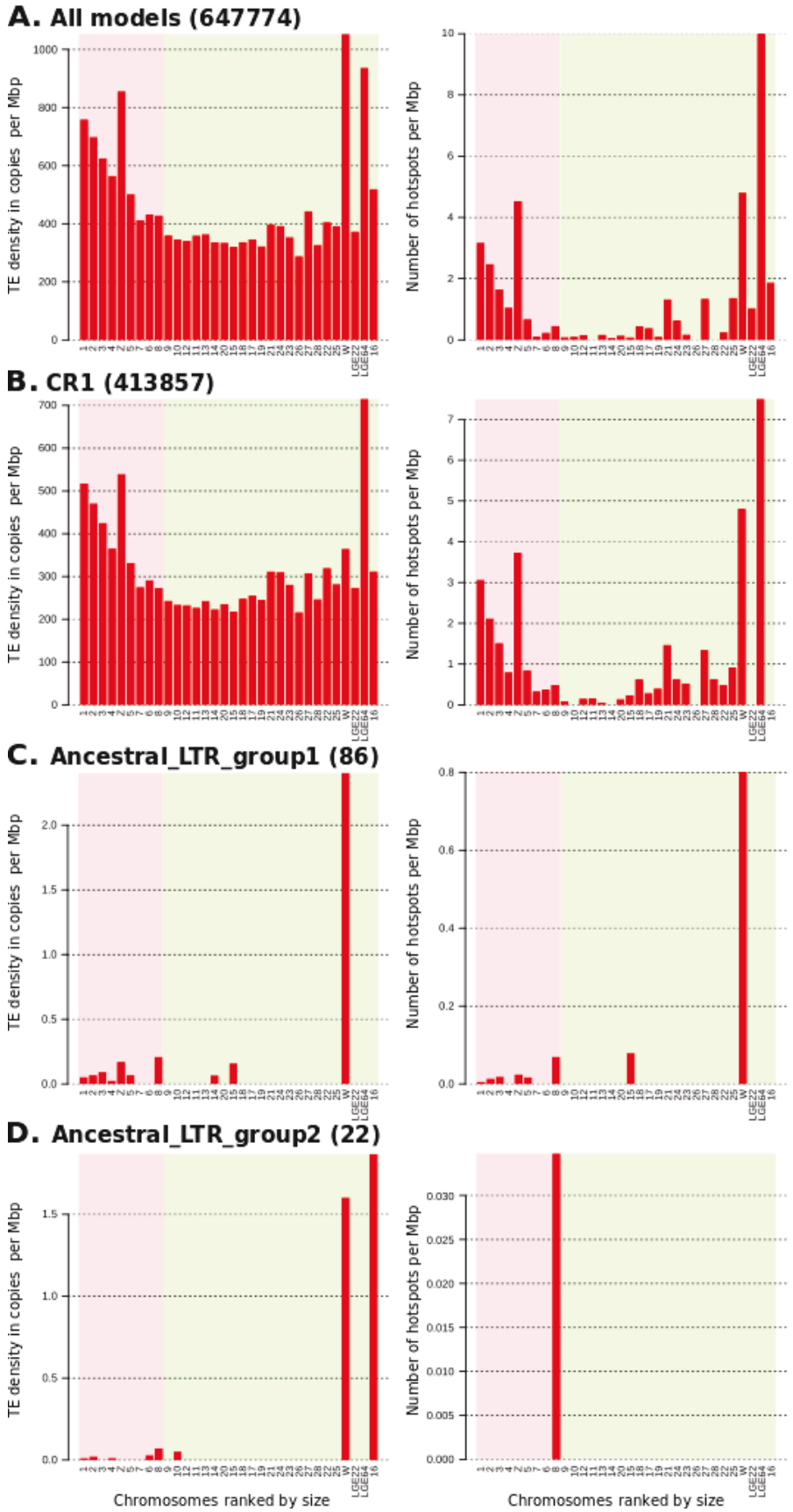

A. CR1-C (27113)
B. CR1-D (56909)
C. CR1_F (36222)
D. CR1-G (9635)
TE density in copies per Mbp
Number of hotspots per Mbp
Chromosomes ranked by size
Chromosomes ranked by size

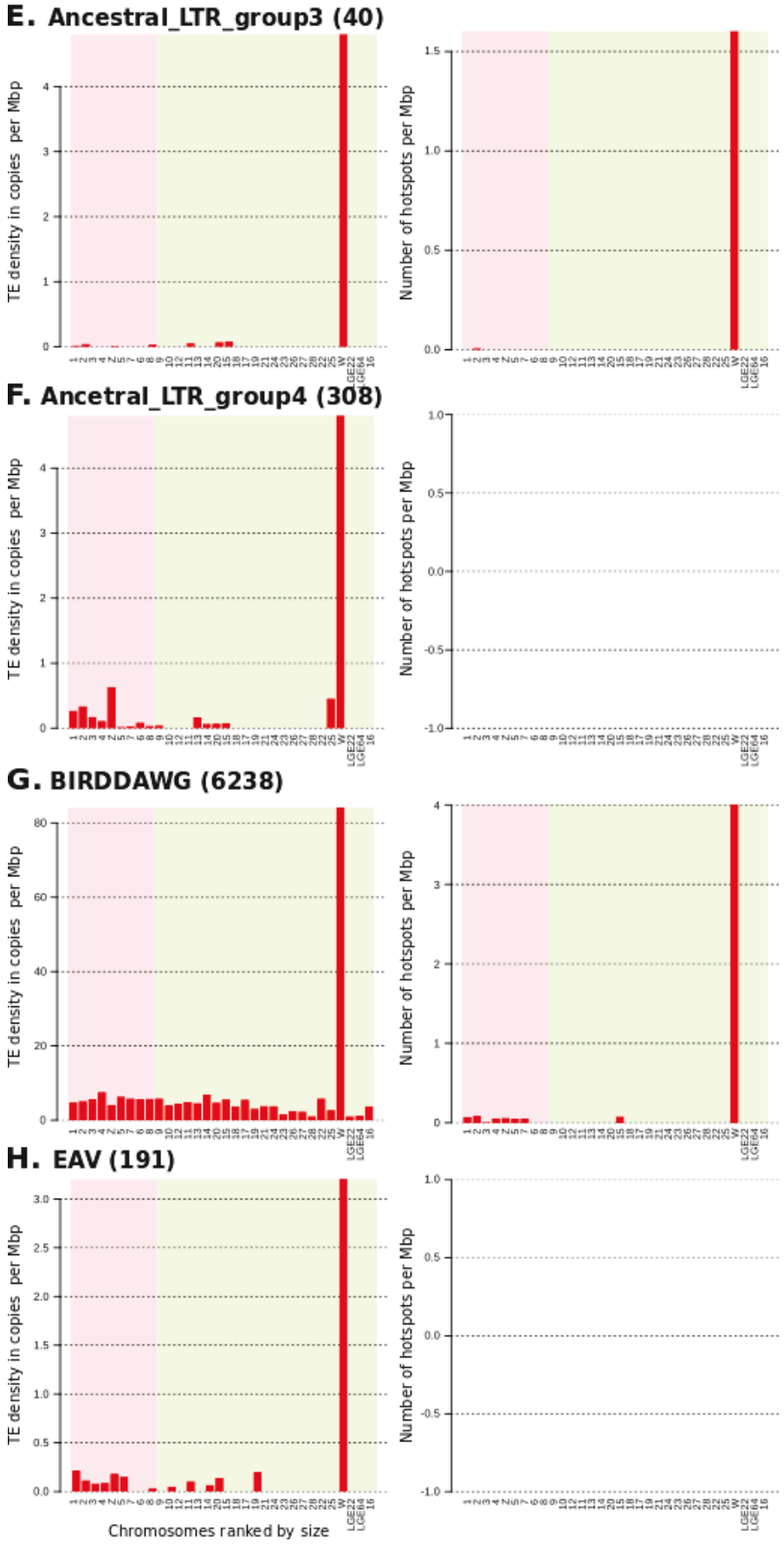

E. CR1_GG (10791)
F. CR1-H (12544)
G. CR1-Y (211266)
H. CR1-like (19377)
TE density in copies per Mbp
Number of hotspots per Mbp
Chromosomes ranked by size

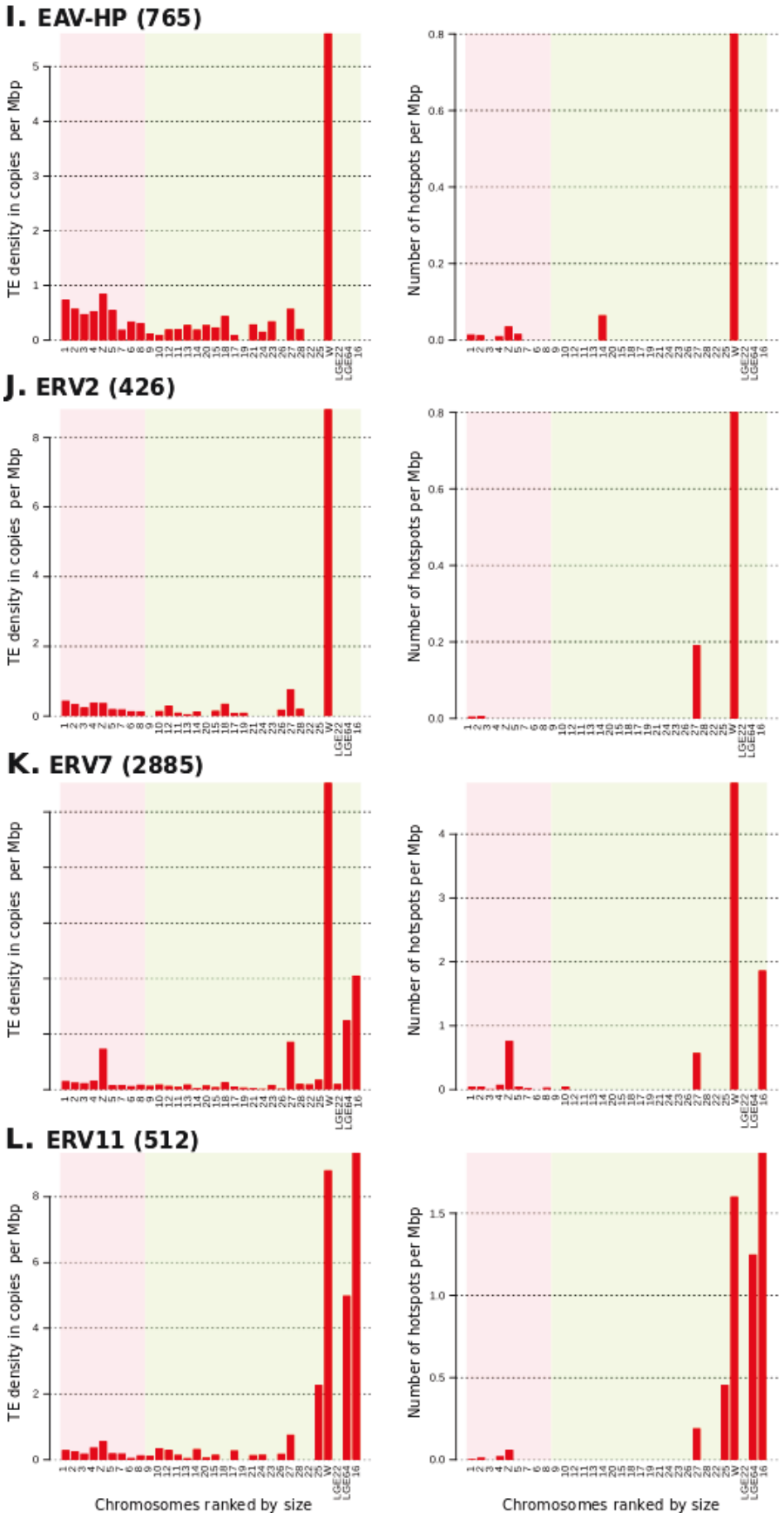
Annexe 30 : Ré-annotation et re-découverte du modèle Galgal4 - Additional file 17

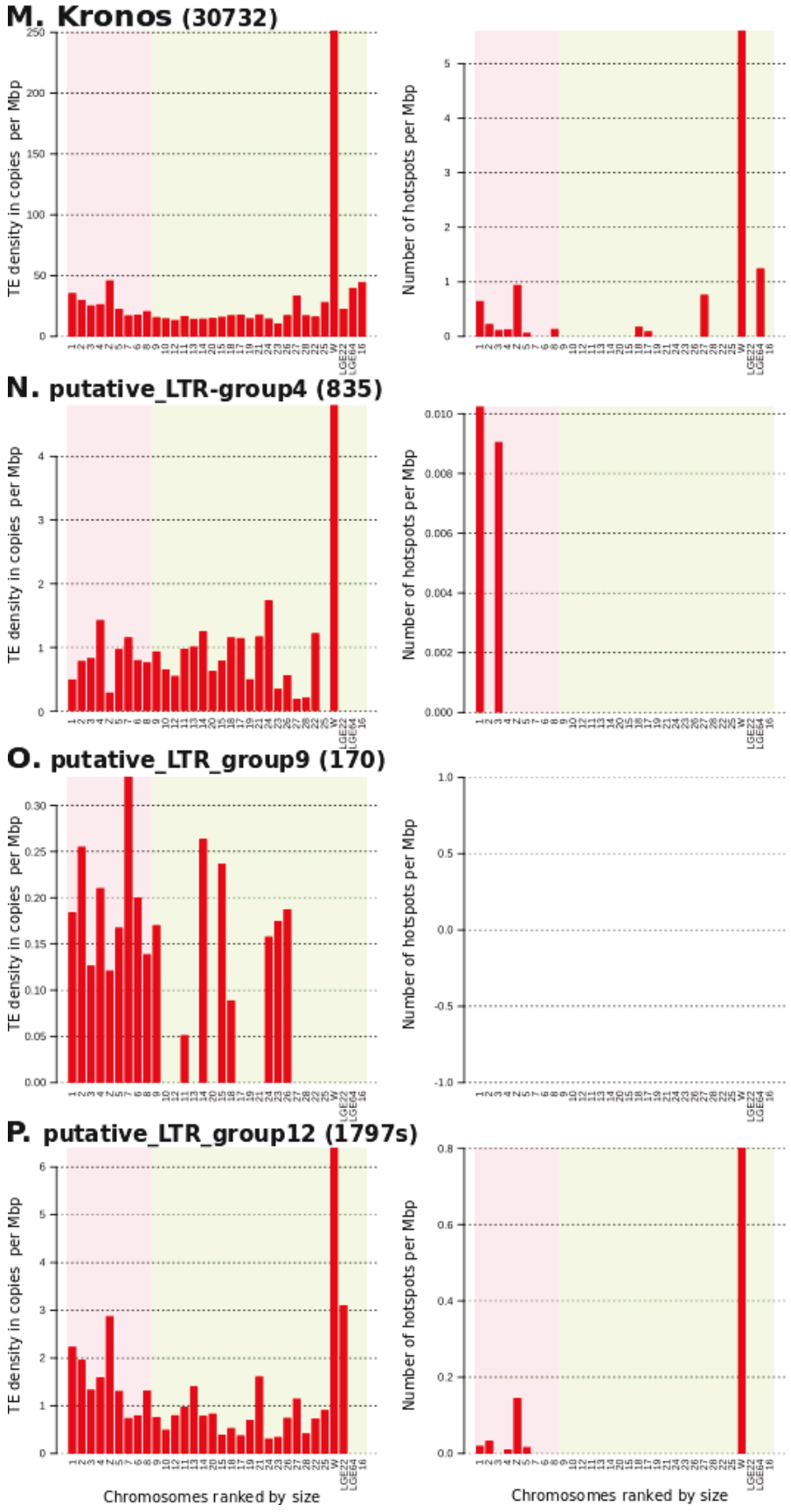
Additional File 17: Histograms showing the densities of TEs (left column) and TE hot spots in galGal4 chromosomes for all TEs plus each of the 34 TE models.

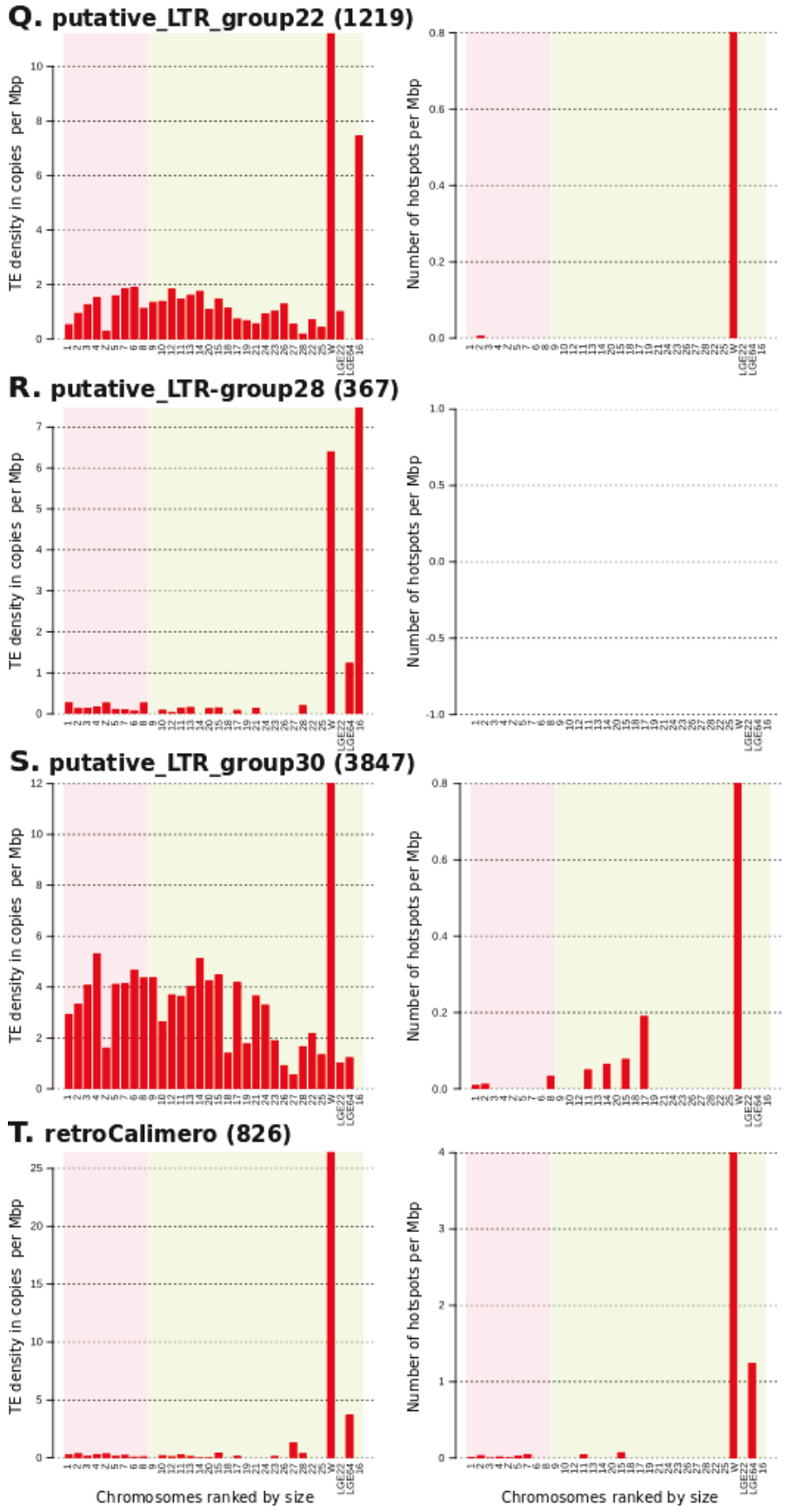
Histograms of TE model density (left column) and TE hot spot density (right column) were calculated for all galGal4 chromosomes, except chromosome 32 (1028 bp). The number of copies for each dataset are indicated in brackets.

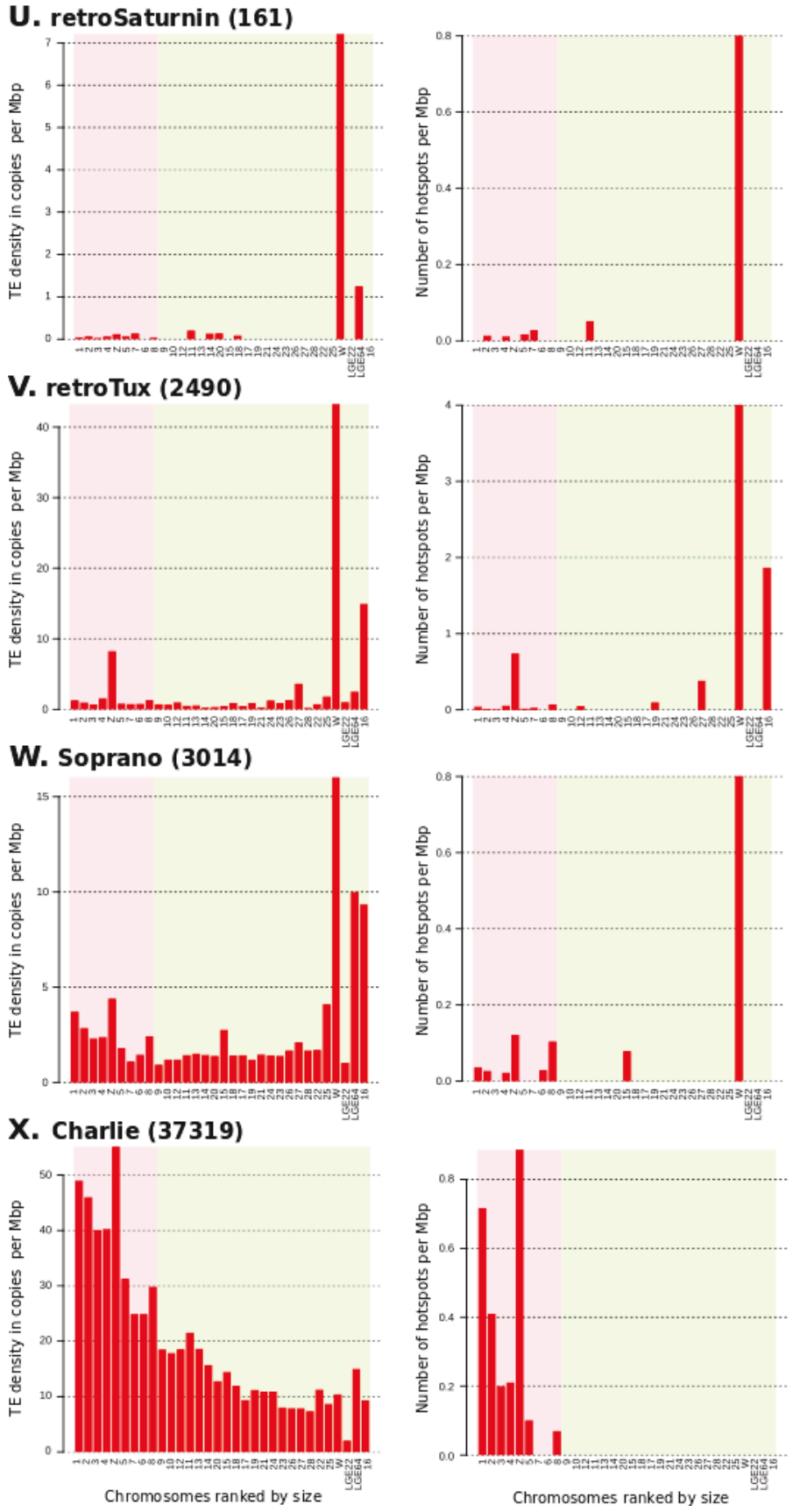

A. All models (647774)
B. CR1 (413857)
C. Ancestral_LTR_group1 (86)
D. Ancestral_LTR_group2 (22)
TE density in copies per Mbp
Number of hotspots per Mbp
Chromosomes ranked by size
Chromosomes ranked by size

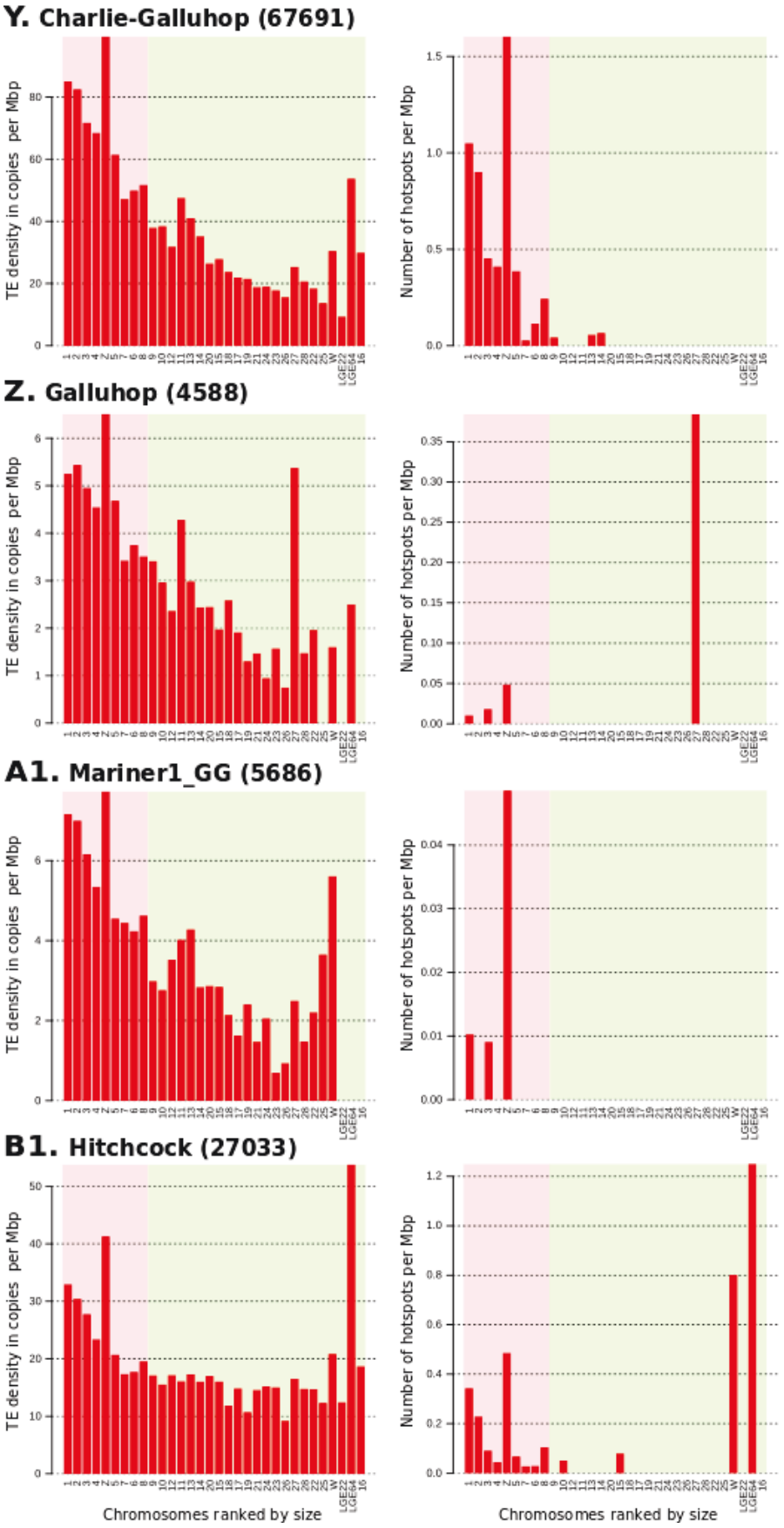

E. Ancestral_LTR_group3 (40)
F. Ancetral_LTR_group4 (308)
G. BIRDDAWG (6238)
H. EAV (191)
TE density in copies per Mbp
Number of hotspots per Mbp
Chromosomes ranked by size

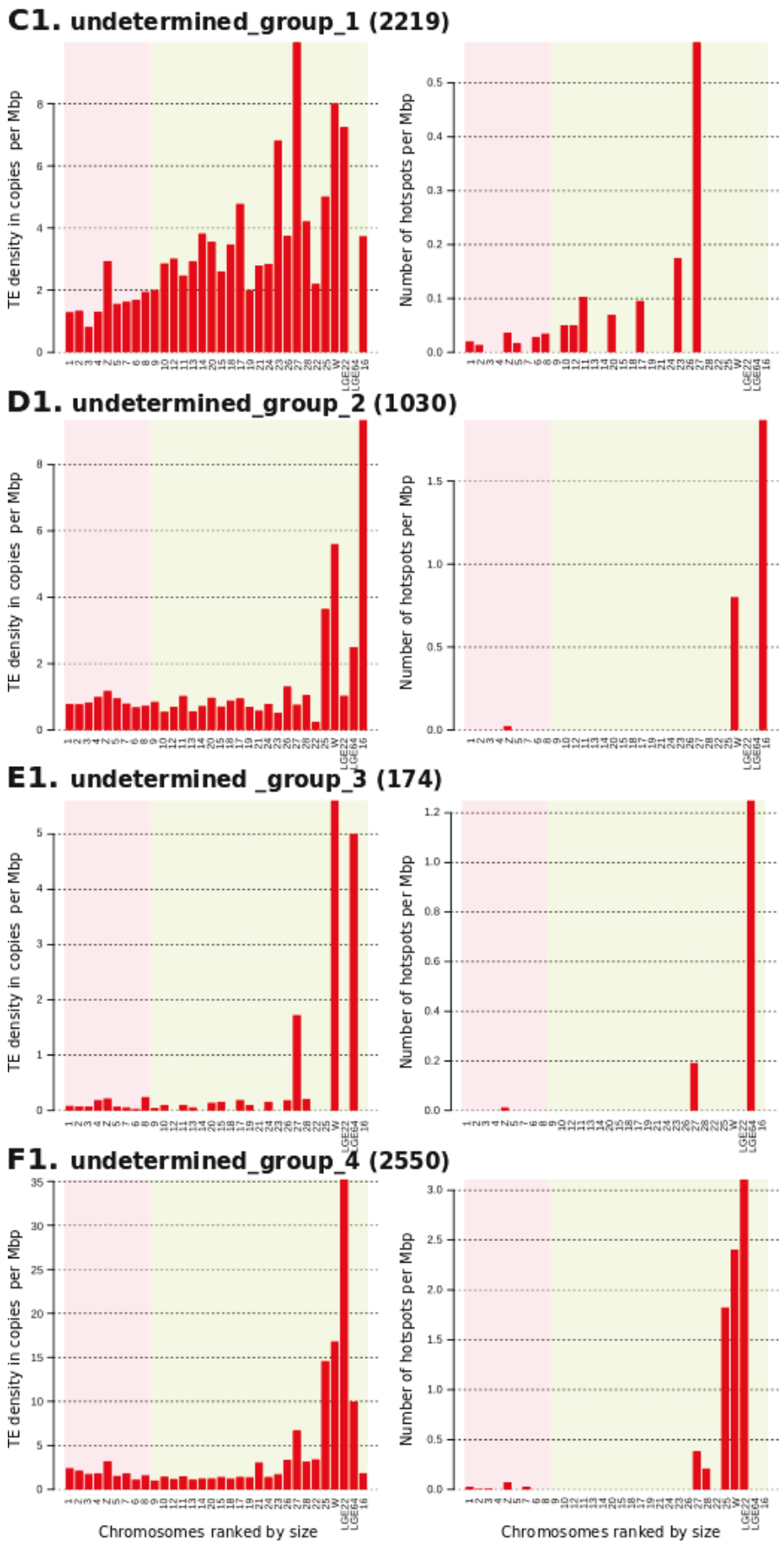

I. EAV-HP (765)
J. ERV2 (426)
K. ERV7 (2885)
L. ERV11 (512)
TE density in copies per Mbp
Number of hotspots per Mbp
Chromosomes ranked by size
Chromosomes ranked by size

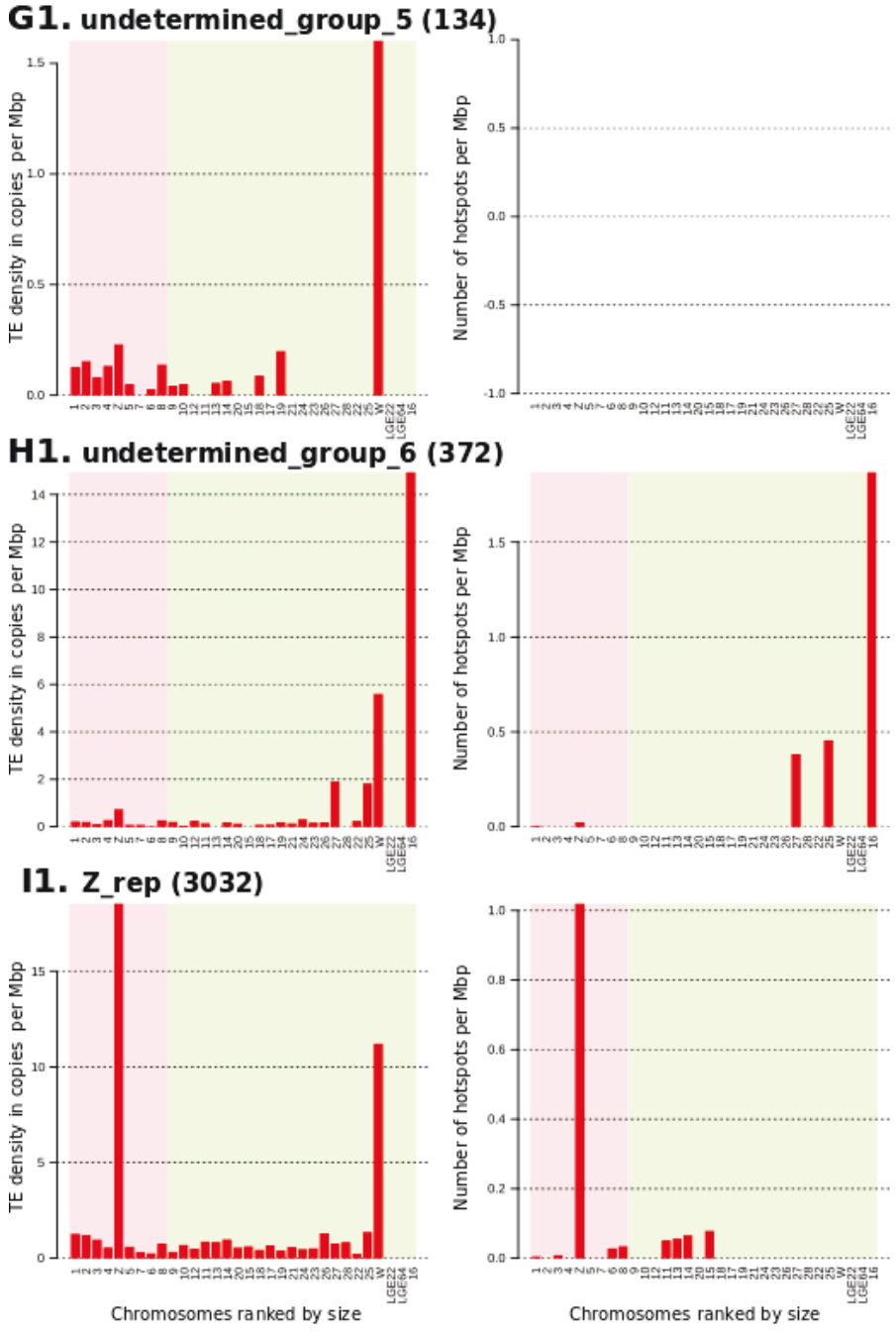

M. Kronos (30732)
TE density in copies per Mbp
Number of hotspots per Mbp
N. putative_LTR-group4 (835)
TE density in copies per Mbp
Number of hotspots per Mbp
O. putative_LTR_group9 (170)
TE density in copies per Mbp
Number of hotspots per Mbp
P. putative_LTR_group12 (1797s)
TE density in copies per Mbp
Number of hotspots per Mbp
Chromosomes ranked by size
Chromosomes ranked by size


Q. putative_LTR_group22 (1219)
R. putative_LTR-group28 (367)
S. putative_LTR_group30 (3847)
T. retroCalimero (826)
TE density in copies per Mbp
Number of hotspots per Mbp
Chromosomes ranked by size


U. retroSaturnin (161)
V. retroTux (2490)
W. Soprano (3014)
X. Charlie (37319)
TE density in copies per Mbp
Number of hotspots per Mbp
Chromosomes ranked by size
Chromosomes ranked by size


Y. Charlie-Galluhop (67691)
TE density in copies per Mbp
Number of hotspots per Mbp
Z. Galluhop (4588)
TE density in copies per Mbp
Number of hotspots per Mbp
A1. Mariner1_GG (5686)
TE density in copies per Mbp
Number of hotspots per Mbp
B1. Hitchcock (27033)
TE density in copies per Mbp
Number of hotspots per Mbp
Chromosomes ranked by size
Chromosomes ranked by size


C1. undetermined_group_1 (2219)
TE density in copies per Mbp
Number of hotspots per Mbp
D1. undetermined_group_2 (1030)
TE density in copies per Mbp
Number of hotspots per Mbp
E1. undetermined _group_3 (174)
TE density in copies per Mbp
Number of hotspots per Mbp
F1. undetermined_group_4 (2550)
TE density in copies per Mbp
Number of hotspots per Mbp
Chromosomes ranked by size
Chromosomes ranked by size

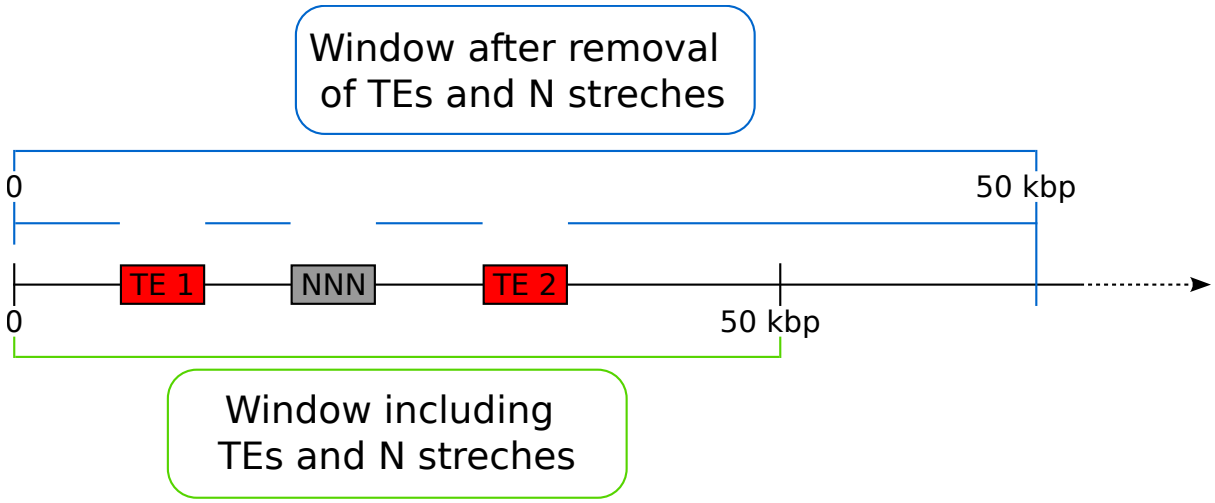

G1. undetermined_group_5 (134)
TE density in copies per Mbp
Number of hotspots per Mbp
H1. undetermined_group_6 (372)
TE density in copies per Mbp
Number of hotspots per Mbp
I1. Z_rep (3032)
TE density in copies per Mbp
Number of hotspots per Mbp
Chromosomes ranked by size
Chromosomes ranked by size

Annexe 31 : Ré-annotation et re-découverte du modèle Galgal4 - Additional file 18

Additional File 18: Graph showing thresholds calculated in permutation assays and windows calculated along chromosomes for permutation tests designed to inventory hot spots.

A. Thresholds

B. Features of windows calculated for permutation assays designed to inventory TE hotspots



Sébastien Guizard

Étude de l'organisation du génome de poulet à travers les séquences répétées

Résumé

Les génomes des espèces aviaires ont des caractéristiques particulières comme la structure des chromosomes et le contenu en séquences répétées. En effet, alors que dans les génomes vertébrés, la proportion de répétitions dans le génome varie de 30 à 55 %, dans les espèces aviaires, cette proportion est plus faible et varie de 8 à 10 %.

L'annotation du contenu répété est le plus souvent réalisée avec le programme RepeatMasker qui s'appuie généralement sur la banque de séquences répétées Repbase. Ce genre de méthode repose uniquement sur la séquence des éléments transposables connus. De fait, ce programme n'est pas en mesure de détecter de nouvelles séquences répétées, et la qualité de l'annotation sera donc dépendante de la banque de séquences d'éléments transposables utilisée.

De plus en plus d'études montrent que les éléments transposables jouent un rôle dans le fonctionnement du génome et peuvent influer sur l'expression des gènes. Il est donc primordial que l'annotation de ces séquences soit la plus complète possible.

Au cours de ma thèse a été mise en place une stratégie d'annotation des séquences répétées que nous avons élaborée et appliquée à un génome de grande taille, celui de la poule rouge de jungle. L'annotation ainsi obtenue m'a permis d'étudier l'organisation du génome de cette espèce au travers de ses séquences répétées et éléments transposables.

Mots-clés : poulet / ADN satellite / éléments transposables / Bio-informatique / benchmarking / répétitions

Résumé en anglais

The genomes of avian species have special features such as the structure of chromosomes or their content in repeated sequences. Indeed, compared to vertebrate genomes in which the amount of repetitions varies from 30 to 55%, it is lower in avian species and varies from 8 to 10%.

The annotation of repeated content is most often done with the RepeatMasker program that is generally use the Repbase database of repeated sequences. This kind of approach is based solely on the sequence of already known transposable elements. In fact, this program is not able to detect new repeats and in consequence produced annotations with a quality that depends on the sequences of transposable elements used.

More and more studies show that transposable elements play a role in the functioning of the genome and can influence gene expression. It is therefore essential that the annotation of these sequences is as complete as possible. There are many programs using methods for detecting de novo transposable elements, either by searching for characteristic structures, or by comparing the genome against itself. However, no standard strategy of annotation for repeated sequences have been defined yet.

My thesis aims to set-up a standard strategy of annotation for repeated sequences that was applied to a large genome, that of the red jungle fowl. The obtained annotation allowed me studying the genome organization in this species through its repeated sequences and transposable elements.

Keywords : chicken / satellite DNA / transposable elements / bioinformatics / benchmarking / repeat