Investigating the impact of repetitive DNA on genome dynamics : Towards a focus on mosquito genomes

Yasmine Mansour

THÈSE POUR OBTENIR LE GRADE DE DOCTEUR DE L'UNIVERSITE DE MONTPELLIER

En Informatique

École doctorale : Information, Structures, Systèmes

Unité de recherche LIRMM - ISEM

Investigating the impact of repetitive DNA on genome dynamics

Towards a focus on mosquito genomes

Présentée par Yasmine MANSOUR

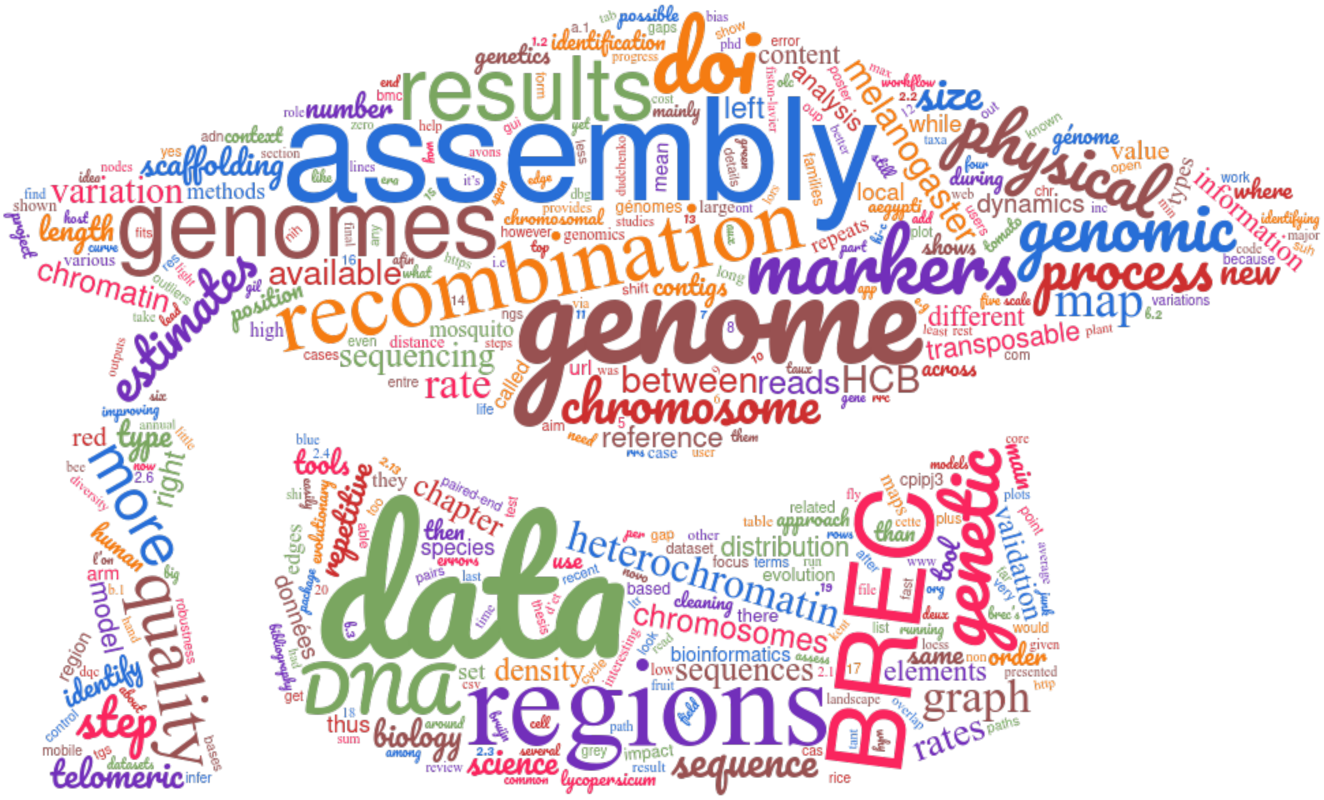
Le 10 Décembre 2021

Sous la direction de Annie CHATEAU et Anna-Sophie FISTON-LAVIER

Devant le jury composé de

Pr. Thérèse COMMES, Professeur des Universités, IRMB, Univ. Montpellier, France	Présidente du jury
Dr. Gabriel MARAIS, Directeur de Recherches, CIBIO-BIOPOLIS, Univ. Porto, Portugal	Rapporteur
Dr. Dominique LAVENIER, Directeur de Recherches, CNRS, IRISA-INRIA, Univ. Rennes 1, France	Rapporteur
Dr. Carène RIZZON, Maîtresse de Conférences, LaMME, Univ. Évry Val d'Essonne, France	Examinatrice
Dr. Annie CHATEAU, Maîtresse de Conférences HDR, LIRMM, Univ. Montpellier, France	Directrice
Dr. Anna-Sophie FISTON-LAVIER, Maîtresse de Conférences, ISEM, Univ. Montpellier, France	Co-encadrante

UNIVERSITÉ DE MONTPELLIER


assembly
genomes
genome
recombination
data
regions
BREC
markers
results
doi
physical
genomic
process
chromosome
heterochromatin
quality
more
DNA
genetic
estimates
melanogaster
map
reads
HCB
between
sequences
chromosomes
graph
step
telomeric
sequence
rates
transposable
elements
distribution
validation
variation
chromatin
available
sequencing
rate
scaffolding
methods
length
size
left
content
analysis
information
new
type
right
model
identify
density
chapter
species
tools
repetitive
biology
science
mosquito
reference
dynamics
different
local
value
contigs
number
genetics
identification
possible
human
bioinformatics
evolution
approach
order
same
cleaning
lycopersicum
données

Abstract

Genomic variation is induced by numerous factors simultaneously, which results in a set of genomic behaviours related to its structure, architecture, expression, evolution, etc, which could be referred to as genome dynamics. During my thesis project, we chose to focus on three major players impacting genome dynamics:

- **Chromatin structure** unevenly compacted along chromosomes;
- **Meiotic recombination landscape** reflecting the frequency variations of exchanging DNA fragments during cell division;
- **Repetitive DNA** mainly **T**ransposable **E**lements (**TEs**) inducing genome assembly errors.

Firstly, We propose an automated computational tool, based on the Marey maps method, allowing to identify heterochromatin boundaries along chromosomes and estimating local recombination rates. Our method, called **BREC** (heterochromatin **B**oundaries and **REC**ombination rate estimates) is non-genome-specific, running even on non-model genomes as long as genetic and physical maps are available. BREC is a statistic-based data-driven tool. Therefore, a data pre-processing module (data quality control and cleaning) is provided. BREC results would allow conducting more broadly an analysis with a comparative genomics approach on their identified heterochromatin regions in terms of recombination landscape, TE density, etc.

Secondly, in order to address the genome assembly process which is strongly impacted by the TE abundance, one type of repeats, we chose to focus on the scaffolding step with the aim of enhancing the assembly quality by exploiting the analysis of repeated regions and proposing a pipeline of improvement. We present an encouraging preliminary result towards this goal.

To conclude this thesis manuscript, we present an opening concerning genomes dynamics with respect to the different aspects addressed. Then, we present the conceptual, application, and technical limits identified by our experimental design. Finally, we suggest a few perspectives on the scope of our contributions beyond my PhD project.

Keywords

Bioinformatics, Genome dynamics, DNA repeats,
Transposable elements, Eu-heterochromatin regions,
Recombination rate, Genome assembly.

Résumé

Les variations du génome sont induites par de nombreux facteurs simultanément, ce qui se traduit par un ensemble de comportements génomiques liés à sa structure, son architecture, son expression, son évolution, etc., que l'on pourrait appeler la dynamique du génome. Au cours de mon projet de thèse, nous avons choisi de nous concentrer sur trois acteurs majeurs impactant la dynamique du génome :

- **Structure de la chromatine** inégalement compactée le long des chromosomes;
- **Paysage de la recombinaison méiotique** reflétant les variations de fréquences d'échange de fragments d'ADN lors de la division cellulaire ;
- **ADN répétitif** notamment les éléments transposables (**ET**) induisant des erreurs d'assemblage du génome.

En premier lieu, nous proposons un outil de calcul automatisé, basé sur la méthode des cartes de Marey, permettant d'identifier les limites d'hétérochromatine le long des chromosomes et d'estimer les taux de recombinaison locale. Notre méthode, appelée **BREC** (heterochromatin **B**oundaries and **REC**ombination rate estimates) n'est pas spécifique au génome, et s'exécute même sur des génomes non modèles tant que des cartes génétiques et physiques sont disponibles. BREC est basé sur des statistiques et axé sur les données, ce qui implique qu'une bonne qualité des données d'entrée reste une exigence forte. Par conséquent, un module de pré-traitement des données (nettoyage et contrôle de la qualité des données) est fourni. Les résultats de BREC permettent de mener une approche de génomique comparative sur les régions hétérochromatiques identifiées en terme de paysage de recombinaison, de densité de ET, etc.

En second lieu, afin d'aborder le processus d'assemblage du génome qui est fortement impacté par l'abondance des ET, nous avons choisi de nous concentrer sur l'étape d'échafaudage et d'améliorer la qualité de l'assemblage en exploitant l'analyse des régions répétées sous la forme d'un pipeline. Nous présentons un travail préliminaire encourageant dans cette perspective.

Pour conclure ce manuscrit de thèse, nous présentons une ouverture concernant la dynamique des génomes par rapport aux différents aspects abordés. Ensuite, nous présentons les limites conceptuelles, applicatives et techniques identifiées par notre modèle expérimental. Enfin, nous proposons quelques perspectives sur la portée de nos contributions au-delà de mon projet doctoral.

Mots-clés

Bioinformatique, Dynamique du génome, Répétitions d'ADN,
Éléments transposables, Régions eu-hétérochromatiques,
Taux de recombinaison, Assemblage des génomes.

ملخص الأطروحة

يتم حدوث التباين الجينومي من خلال العديد من العوامل في وقت واحد، مما يؤدي إلى مجموعة من السلوكيات الجينومية المتعلقة بهيكلها، وبنيتها، وتعبيرها، وتطورها، وما إلى ذلك، والتي يمكن الإشارة إليها باسم ديناميكيات الجينوم. خلال مشروع أطروحتي، اخترنا التركيز على ثلاثة لاعبين رئيسيين يؤثرون على ديناميكيات الجينوم:

- بنية الكروماتين: مضغوطة بشكل غير متساو على طول الكروموسومات.
- منظر لإعادة التركيب النصفي: يعكس الاختلافات في وتيرة تبادل قطع الحمض النووي أثناء انقسام الخلية.
- الحمض النووي المتكرر: على وجه الخصوص العناصر القابلة للنقل (TE) التي تسبب أخطاء في تجميع الجينوم.

أولاً، نقترح أداة حسابية آلية، بناءً على طريقة خرائط ماري، مما يسمح بتحديد حدود الهيتروكروماتين على طول الكروموسومات وتقدير معدلات إعادة التركيب المحلية. طريقتنا، المسماة BREC (تقديرات حدود الهيتروكروماتين وتقديرات معدل إعادة التركيب) غير محددة الجينوم، وتعمل حتى على الجينوم غير النموذجي طالما أن الخرائط الجينية والفيزيائية متوفرة. BREC هي أداة تعتمد على البيانات الإحصائية. لذلك، قمنا بتوفير وحدة المعالجة المسبقة للبيانات (مراقبة جودة البيانات والتنظيف). ستسمح نتائج BREC بإجراء تحليل على نطاق أوسع باستخدام نهج علم الجينوم المقارن على مناطق الهيتروكروماتين التي تم تحديدها من حيث مشهد إعادة التركيب، وكثافة TE ، وما إلى ذلك.

ثانيًا، من أجل معالجة عملية تجميع الجينوم التي تتأثر بشدة بوفرة TE ، نوع واحد من التكرارات، اخترنا التركيز على خطوة السقالات الجينومية بهدف تعزيز جودة التجميع من خلال استغلال تحليل المناطق المتكررة واقتراح خط أنابيب برمجي للتحسين. نقدم نتيجة أولية مشجعة نحو هذا الهدف.

لاختتام هده الأطروحة، نقدم تمهيدا يخص ديناميكيات الجينوم فيما يتعلق بالجوانب المختلفة التي تم تناولها. بعد ذلك، نشير إلى الحدود المفاهيمية والتطبيقية والتقنية التي حددها تصميمنا التجريبي. أخيرًا، نقترح بعض وجهات النظر حول نطاق مساهماتنا خارج مشروع الدكتوراه الخاص بي.

الكلمات المفتاحية:

المعلوماتية الحيوية،
علم الجينوم،
تكرارات الحمض النووي،
العناصر القابلة للنقل،
تجميع الجينوم.

Résumé étendu

Contexte scientifique : description du projet de thèse

Durant ce doctorat, nos travaux ont porté essentiellement sur des problématiques en lien avec la structuration du génome et son évolution. Le génome est une composante indispensable des individus, qui participe à leur identification, et propose un angle d'observation du vivant qui fluctue en terme de contenu à différentes échelles, que ce soit au niveau de l'espèce, des variétés ou des individus appartenant à une même espèce, le tout dans une dynamique qui est alimentée par les croisements entre les patrimoines génétiques des individus. Pour observer ces génomes et leur dynamique, nous disposons de plusieurs outils mêlant les modèles numériques et les données expérimentales. Durant les dernières décennies, l'avènement de la génomique a notamment été possible grâce aux avancées technologiques liées au séquençage haut-débit, rendant accessibles de plus en plus de génomes. Mais derrière la réalité de ce déluge de données, se cache la difficulté à traiter ces données de façon à pouvoir intégrer les informations qui s'y cachent. Loin d'être la seule source de données accessible pour observer les génomes, les données de séquençage sont souvent présentées comme une panacée permettant d'analyser les génomes, mais elles ne représentent qu'une facette des observations possibles. Au cours de la thèse, nous avons observé le génome non seulement en tant qu'entité dotée d'une structure, mais aussi en tant que modèle d'observation du vivant, imparfait et donc, perfectible. Nous avons focalisé notre attention sur l'observation des génomes eucaryotes, du point de vue de leur structure et dynamique, en relation avec les éléments répétés qu'ils contiennent. Chacun des paragraphes ci-dessous correspond à un chapitre du manuscrit[1].

Concepts fondamentaux

Dans le règne du vivant, un organisme est considéré comme une espèce présente dans l'arbre de la vie. Chez les eucaryotes, organismes dont les cellules possèdent un noyau, celui-ci contient le matériel génétique, le génome. Le génome est organisé en *chromosomes*, chaque chromosome faisant intervenir deux parties appelées les *chromatides*, solidarisées au niveau du *centromère*. Un *gène* est une portion d'*ADN* (Acide DésoxyriboNucléique) menant à la production d'une *protéine* assurant une fonction donnée dans l'organisme. La séquence d'ADN est constituée d'une succession de *nucléotides*, également appelés *bases*, de quatre types, représentées par les lettres A (Adénine), C (Cytosine), G (Guanine) et T (Thymine). Et puisque la molécule d'ADN est constituée d'une double-hélice formée de deux *brins* complémentaires, chacune de ces lettres est associée par *complémentarité* à une autre, A avec T, C avec G. Cette

[1] Nous avons fait le choix de ne pas faire apparaître les citations dans ce résumé long, leur densité risquant d'en gêner la lecture.

séquence de nucléotides s'organise en une structure appelée *chromatine* qui est constituée d'une succession de *nucléosomes* eux-mêmes formés à partir de nucléotides et de protéines : les *histones*. Ces nucléosomes organisent la *compaction* de la molécule d'ADN. Cette compaction de l'ADN joue un rôle dans les différents mécanismes en lien avec le génome, notamment la *réplication*, la *réparation* et la *recombinaison*, sur laquelle nous reviendrons dans le chapitre 1. La chromatine se divise en deux catégories : l'*euchromatine* et l'*hétérochromatine*. L'hétérochromatine ne change pas d'état de condensation au cours du cycle cellulaire, à l'inverse de l'euchromatine.

En-dehors des gènes codant l'information sur les protéines, il existe de nombreuses régions *non-codantes* dans l'ADN, que l'on a longtemps qualifiées d'*ADN poubelle*. Dans ces régions, qui sont loin d'avoir révélées tout leur mystère, des séquences existent en plusieurs occurrences dans le génome, que l'on appelle *régions répétées*. Ces régions répétées sont de plusieurs types : les *répétitions en tandem*, qui sont de courtes séquences répétées consécutivement ; les répétitions de grande taille, aussi appelées *duplications* et qui peuvent concerner par exemple des gènes ; et enfin, les *éléments transposables* (ET), qui se distinguent par leurs grande diversité et dispersé le long des génomes. Ces ET ont révolutionné le champ de la génétique, et sont de plus en plus utilisés pour étudier la dynamique des génomes. En effet, hautement répétés, inégalement répartis le long des génomes, et se décomposant en plusieurs classes facilement identifiables grâce à leur structure, ces éléments peuvent représenter unr grande part des génomes et contribuer à la structuration des génomes. La composition des génomes en ET est très variable en fonction des espèces (de <1% à plus de 90% des génomes de plantes). Ils sont connus pour être impliqués dans différents processus biologiques comme les réarrangements chromosomiques, ou la modification de l'expression des gènes. Contrairement aux deux autres catégories de répétitions, les ET sont des répétitions mobiles: Ils peuvent se déplacer (ou se transposer) d'une position à l'autre le long du génome en suivant deux mécanismes différents : Les éléments de la classe 1 (rétrotransposons) utilisent le mécanisme copier-coller à travers un ARN (Acide RiboNucléique) intermédiaire. De cette façon, la séquence d'ADN du site donneur conserve l'ET d'origine pendant que sa copie trouve une séquence cible pour s'y insérer. D'autre part, les éléments de la classe 2 (transposons) utilisent le mécanisme de couper-coller pour sauter du site donneur à la séquence cible. Dans ce cas, la séquence du donneur subit une rupture d'ADN qui sera réparée soit en joignant les deux extrémités du gap, soit en recevant une nouvelle insertion d'ET. Tous ces événements donnent lieu à une très grande diversité (taille, nombre de copies, ...) d'ET, qui sont ensuite classés en sous-classes, super-familles et familles en fonction des caractéristiques de leurs séquences. Afin de mieux décrire l'impact des ET sur la structure et l'évolution de ces génomes, nous avons besoin de différents types d'informations. Par exemple, l'abondance et la distribution des ET, quelles familles d'ET sont principalement présentes, est-ce qu'il existe une corrélation entre les ET et d'autres caractéristiques génomiques telles que la densité des gènes et le taux de recombinaison, entre autres.

Étant donné que l'étude des ET nécessite de prendre en compte toutes leurs caractéristiques et tous leurs comportements, dans ce projet, nous avons choisi de tester les approches développées avec les génomes de moustiques. Les moustiques sont des vecteurs de maladies infectieuses chez l'homme (paludisme, Zika, fièvre jaune, etc.). Leurs génomes présentent un modèle unique d'évolution dont l'adaptation est l'un des processus évolutifs les mieux connus. Elle permet aux moustiques de développer une résistance aux insecticides et de survivre dans les environnements

extrêmes et aux changements climatiques brutaux. Certaines familles d'ET sont impliquées dans un tel processus adaptatif. Nous nous sommes intéressés à trois espèces de moustiques: *Aedes aegypti* (1,3 Gb), *Anopheles gambiae* (278 Mb) et *Culex pipiens* (579 Mb), qui avaient été reassemblées au début de ma thèse. Nous nous référons à l'espèce *Drosophila melanogaster* (mouche du vinaigre) comme une référence hors groupe. Ces espèces phylogénétiquement très proches présentent une grande variabilité en termes de taille de génome et de contenu de séquences répétées, y compris les ET.

La *dynamique* des génomes, que l'on peut définir comme la capacité des génomes à évoluer dans leur structure, architecture ou expression, à différentes échelles, est l'objet d'étude en filigrane de la thèse. Loin de balayer les nombreux aspects liés à cette dynamique, nous avons concentré notre attention sur trois acteurs majeurs de la dynamique des génomes:

- la structure chromatinienne, notamment la répartition entre euchromatine et hétérochromatine, qui est inégale le long des génomes;
- la recombinaison méiotique qui permet l'échange de fragments d'ADN au cours de la division cellulaire;
- l'ADN répété, et particulièrement les ET, qui sont notamment l'un des facteurs perturbant la reconstruction des séquences génomiques après leur séquençage.

BREC, un outil d'observation de la structure chromatinienne et du taux de recombinaison le long des chromosomes

Dans les travaux qui suivent, nous nous intéressons aux génomes à l'échelle du génome complet, en tant que représentant d'un organisme donné. Les événements qui vont nous intéresser se produisent donc à une échelle suffisamment grande pour lisser ou occulter les variations ponctuelles que l'on peut observer entre les individus. Nous nous focaliserons ainsi sur la structure globale du génome, avec le premier problème de l'identification de zones structurellement non homogènes que sont les régions euchromatiques et hétérochromatiques. Pour déterminer ces zones avec acuité, il est intéressant de s'intéresser aux variations de taux de recombinaison le long des chromosomes. L'estimation de ce taux de recombinaison fait intervenir un premier type de données, les cartes de recombinaison, sur lesquelles nous nous sommes appuyées pour mettre au point une méthode d'analyse et de visualisation, détaillée ci-desssous et dans le chapitre 2.

Afin de déterminer ces frontières entre les régions le long des chromosomes, nous nous sommes intéressées à une donnée qui est fortement correlée à la nature chromatinienne, à savoir le taux de recombinaison. Ce taux est fortement variable le long des chromosomes, et il est possible de l'étudier avec divers procédés et à des échelles différentes. L'échelle qui nous intéresse ici est une échelle suffisamment grossière pour déterminer les tendances globales le long des génomes, et pouvoir les visualiser. Les méthodes d'inférence du taux de recombinaison sont de plusieurs natures et comptent notamment :

- les approches fondées sur l'étude des populations (groupe d'individus d'une même espèce et localisés sur une même zone géographique). Ces approches

nécessitent un très grand nombre de données mais produisent une estimation assez fine du taux de recombinaison ;

- les approches qui observent les gamètes et vont chercher les recombinaisons au coeur de la constitution de celles-ci. Mais elles ne sont applicables que sur les mâles et ne peuvent atteindre que des portions limitées du génome;
- les méthodes fondées sur le pédigrée, qui nécessitent d'étudier les recombinaisons observées entre les parents et leurs descendants. Ces méthodes sont moins précises, mais nécessitent moins de données que les précédentes. Elle sont fondées sur les données génomiques telles que les cartes génétiques et physiques. C'est le cas de l'approche des cartes de Marey, sur laquelle nous nous appuyons dans cette thèse. En effet, les données sont plus accessibles et moins coûteuses, et l'estimation proposée par cette méthode est suffisante pour atteindre le but recherché, à savoir l'observation globale de la structure chromatinienne.

Dans les cartes de Marey, notamment exploitées dans l'outil MareyMapOnline, on trouve deux types de données : les cartes physiques, qui représentent la cartographie de marqueurs donnés le long des chromosomes, avec des distances exprimées en paires de bases, et les cartes génétiques, qui se fondent sur l'observation des recombinaisons durant la méiose, à travers des liaisons statistiquement surreprésentées dans les triades parents-enfants. Ces dernières fournissent une distance statistique, exprimée en centiMorgan : la distance génétique entre deux marqueurs est le nombre moyen de crossing-overs entre les deux marqueurs par méiose. En croisant ces deux informations pour un certain nombre de marqueurs, il est possible d'inférer le taux de recombinaison le long du chromosome.

Afin de compléter les approches existantes, nous avons conçu un outil, BREC (**B**oundaries and **REC**ombination rate estimates), qui se base sur ces données et propose une solution automatique, générique et ergonomique pour :

- estimer les bornes entre les régions euchromatiques et hétérochromatiques,
- estimer les taux de recombinaison localement,
- et ajuster les taux de recombinaison dans les régions où la structure de la chromatine est instable.

La méthode sous-jacente se déploie en une étape préliminaire et six étapes principales. L'étape préliminaire constitue une vérification des données d'entrée en terme de qualité, car la densité des marqueurs et leur distribution sont des facteurs importants pour obtenir une estimation de qualité. Voici les étapes principales :

1. Estimation du taux de recombinaison local en utilisant les cartes de Marey
2. Identification du type de chromosome
3. Préparation de l'identification des bornes hétérochromatiques (calcul d'un facteur d'adéquation entre le taux de recombinaison estimé et les données, et test local à l'aide d'une fenêtre glissante).
4. Identification des bornes du centromère (s'il existe). Cette estimation se fonde sur l'extension de la zone où le taux de recombinaison est le plus faible, en tenant compte du caractère télocentrique ou atélocentrique du chromosome.

5. Identification des bornes des télomères. Cette estimation identifie une chute significative dans les courbes d'adéquation.
6. Extrapolation du taux de recombinaison local et affichage du résultat.

L'implémentation de BREC consiste en un paquet `R`, utilisant le module d'interface `Rshiny`.

Afin de valider la méthode, nous l'avons appliquée à différents jeux de données, qui sont intégrés à l'outil par défaut. Nous en présentons quelques uns dans ce manuscrit, et nous avons concentré l'analyse sur le génome de la drosophile *D. melanogaster* et de la tomate *Solanum lycopersicum*, étudié l'influence des paramètres principaux, et utilisé des données simulées pour étudier la robustesse vis-à-vis de la qualité des données. Enfin, nous avons observé les résultats de BREC sur les génomes de moustiques.

Amélioration des assemblages de génomes

L'assemblage du génome est le processus consistant à rassembler des données de séquençage, les lectures, qui sont de petits fragments d'ADN, dans le but de produire, de la manière la plus proche possible, la forme originale du génome entier. Les régions répétées perturbent de façon importante le processus d'assemblage, qui se fonde essentiellement sur les chevauchements entre les fragments d'ADN lus lors du séquençage. Différentes technologies de séquençage existent, qui proposent des lectures essentiellement de deux types : les lectures courtes, de l'ordre de la centaine de paires de bases, et les lectures longues, atteignant plusieurs milliers ou dizaines de milliers de paires de bases. L'immense majorité des séquences génomiques disponibles dans les bases de données publiques provient de séquençage en lectures courtes, et sont malheureusement les données les plus sensibles au répétitions lors du processus d'assemblage. Ainsi, peu de génomes qualifiés de "complets" proposent des séquences à l'échelle du chromosome, et sont le plus souvent constitués de centaines, voire de milliers de séquences différentes.

Lors de la mise au point de BREC, il nous est apparu qu'il était indispensable, pour observer les phénomènes à l'échelle du génome complet, de disposer de séquences complètes et de bonne qualité pour ces génomes. L'observation d'une part des différences entre les différentes versions des génomes, et d'autre part de la fragmentation très importante des génomes dans les bases de données, nous a conduit sur la piste de l'amélioration des séquences génomiques existantes. L'objectif n'est pas ici de proposer un outil supplémentaire d'assemblage ou d'échafaudage de génomes, mais de considérer le problème posé par les régions répétées lors de l'assemblage, non plus comme un problème, mais également une solution potentielle. C'est l'objet du chapitre 3.

Dans cette partie, nous nous concentrons donc sur l'étape d'échafaudage de génomes, qui représente le produit fini accessible dans les bases de données, afin de déterminer, dans le cadre de données issues de séquençage en lectures courtes, s'il est possible d'exploiter favorablement la connaissance que l'on peut avoir des répétitions.

L'idée de la méthode proposée est de s'appuyer sur le graphe d'échafaudage construit à partir des séquences assemblées, les contigs, et de l'alignement de courtes lectures appariées sur ces contigs. Le graphhe est défini comme suit :

- les sommets du graphe représentent les extrémités des contigs (il y a donc deux sommets par contig)
- les arêtes du graphe sont de deux types : des arêtes dites "de contig" rejoignent les deux extrémités correspondant à un contig donné (ces arêtes forment un couplage parfait du graphe), et des arêtes inter-contig, qui représentent les liens observés entre les contigs grâce à l'alignement des lectures sur les contigs. Ces dernières sont porteuses d'un poids correspondant au nombre de paires de lectures qui relient ces extrémités.

Dans ce contexte, le problème de l'échafaudage de génome correspond à la recherche de chemins optimaux dans ce graphe, ce qui est un problème difficile dû au grand nombre de chemins possibles (autrement dit un problème NP-complet). Nous ne cherchons pas ici à résoudre ce problème, mais à améliorer le graphe d'échafaudage en amont de la résolution. La connaissance préalable que l'on peut avoir des répétitions peut ainsi guider l'élimination ou le renforcement de certaines arêtes dans ce graphe, ce qui a une incidence sur la résolution ensuite.

Nous avons mis au point un pipeline de traitement intégrant ces informations de répétitions, en les extrayant d'une base de données, les alignant sur les contigs afin d'étiqueter ces derniers, et comparer les arêtes du graphe d'échafaudage aux étiquettes des contigs qu'ils relient. En invalidant certaines arêtes dont les informations présentent des incohérences avec l'information des étiquettes, et en renforçant celles qui au contraire présentent une concordance entre les deux types d'informations, on obtient un graphe d'échafaudage modifié, prêt à être résolu.

Nous avons testé la méthode sur deux génomes de référence considérés comme de bonne qualité, le génome de la drosophile *D. melanogaster* et le génome du nématode *Caenorhabditis Elegans*. Le premier est connu pour comporter des répétitions et notamment des ET, le second est moins riche en répétitions. Plusieurs méthodes d'assemblage et d'alignement ont été testées. Nous avons notamment observé le nombre de discordances d'assemblage par rapport à la référence. Les résultats se révèlent encourageant pour le génome de la drosophile pour un type d'assembleur, et à retravailler pour *C. elegans*. Nous avons ensuite analysé ces résultats sous le prisme des ET et constaté qu'ils sont très largement impliqués dans les discordances restantes.

Conclusion et Perspectives, discussion sur le rôle des ET

Dans un dernier chapitre conclusif, le chapitre 4, nous revenons sur les améliorations possibles de l'outil BREC, et proposons des perspectives complémentaires questionnant le rôle des ETs dans l'étude des génomes complets. Nous revenons sur leur rôle dans la dynamique des génomes, phénomène de plus en plus observé et étudié. Le contenu en gènes étant généralement conservé, les ET sont-ils responsables de cette expansion de la taille du génome ? Si c'est le cas, quel type d'ET influence le plus ces génomes ? Les ET influencent-ils toutes les régions chromosomiques de la même manière ou bien existe-t-il des régions spécifiques plus touchées que d'autres, telles que l'hétérochromatine (ADN compact) ou plus précisément les centromères ?

Contents

Abstract v

Résumé vii

Arabic abstract xi

Résumé étendu xiii

1 Introduction 1
- 1.1 Context: on the interface of computer science and evolutionary genomics . . . 1
  - 1.1.1 From Computer Science to Bioinformatics: Data science for life sciences . . . 2
  - 1.1.2 From Bioinformatics to Genomics: Computational biology for the study of whole genomes . . . 3
  - 1.1.3 From Genomics to Evolution: Mosquito research interests . . . 5
- 1.2 Fundamental concepts . . . 6
  - 1.2.1 Genome architecture . . . 6
  - 1.2.2 Transposable Elements: one type of repetitive DNA . . . 7
    - TEs discovery has revolutionized the genetics field . . . 7
    - TEs characteristics . . . 8
    - TEs are ubiquitous and not evenly distributed across eukaryotic taxa . . . 9
- 1.3 Research scope: Genome dynamics . . . 11
  - 1.3.1 Chromatin regions: one chromosome, different genomic profiles 11
  - 1.3.2 Recombination rate: one genomic feature, different landscapes 12
  - 1.3.3 Genome assembly: one genomic goal, different computational challenges . . . 15
- 1.4 Mosquitoes: an interesting model to aim for . . . 16
- 1.5 Thesis overview . . . 20

2 Recombination and heterochromatin regions 23
- 2.1 Context and motivation . . . 24
  - 2.1.1 Approaches for estimating recombination rate variation . . . 24
  - 2.1.2 An approach to estimate quickly and easily the recombination rate along chromosomes . . . 28
- 2.2 New Approach: BREC . . . 31
  - 2.2.1 Step 0 - Apply data pre-processing . . . 36
    - Data quality control . . . 36
    - Data cleaning . . . 36
  - 2.2.2 Step 1 - Estimate Marey Map-based local recombination rates . 37
  - 2.2.3 Step 2 - Identify chromosome type . . . 37
  - 2.2.4 Step 3 - Prepare the HCB identification . . . 39

2.2.5 Step 4 - Identify centromeric boundaries . . . 39
2.2.6 Step 5 - Identify telomeric boundaries . . . 40
2.2.7 Step 6 - Extrapolate the local recombination rate estimates and generate interactive plot . . . 40
2.3 Validation process . . . 41
2.3.1 Validation data . . . 41
Fruit fly genome *D.melanogaster* . . . 43
Tomato genome *S. lycopersicum* . . . 43
2.3.2 Simulated data for quality control testing . . . 45
2.3.3 Validation metrics . . . 50
2.3.4 Implementation and Analysis . . . 50
2.3.5 Description of main components of the Shiny app . . . 51
Build-in dataset . . . 51
GUI input options . . . 53
2.4 BREC results . . . 56
2.4.1 Approximate, yet congruent HCB . . . 59
Fruit fly genome *D.melanogaster* . . . 61
Tomato genome *S. lycopersicum* . . . 61
2.4.2 Consistency despite the low data quality . . . 64
BREC handles low markers density . . . 64
BREC handles heterogeneous distribution . . . 66
2.4.3 Accurate local recombination rate estimates . . . 69
2.4.4 BREC is non-genome-specific . . . 73
2.4.5 Easy, fast and accessible tool *via* an R-package and a Shiny app 73
2.5 Applying BREC to identify chromatin regions along the mosquito genome: *Ae. aegypti* . . . 73
2.6 Discussion and Conclusion . . . 76

3 Improving the quality of genome assembly using repeats 81
3.1 Context and motivation . . . 81
3.1.1 Genome assembly overview . . . 82
Genome assembly: Reconstructing the reference genome . . . 82
Sequencing data: the inputs for genome assembly approaches . 84
To what extent the current reference genomes are complete? . . 86
3.1.2 *de novo* assembly approaches . . . 87
Greedy algorithms . . . 87
Overlap-Layer-Consensus . . . 87
De Bruijn graphs . . . 88
Repeats: a big challenge facing the assembly process . . . 91
Assemblers . . . 93
3.1.3 Genome scaffolding . . . 94
Assembling contigs into scaffolds . . . 94
3.1.4 Correcting short read assembly errors . . . 97
3.2 New approach: From classic to enhanced scaffolding . . . 98
3.2.1 Method description . . . 98
Data production . . . 98
Repeated Regions analysis . . . 99
After the RR analysis: solving and quality analysis . . . 101
3.2.2 Data simulated . . . 101
3.3 Results . . . 101
3.3.1 Impact on the assembly quality . . . 101

D. melanogaster . . . 101
C. elegans . . . 102
3.3.2 RR within the misassemblies . . . 103
3.4 Discussion and conclusion . . . 105

4 Conclusion and Perspectives **107**
4.1 BREC : A user-friendly tool for accurate recombination rate and chromatin boundary estimates . . . 107
4.1.1 BREC's limitations . . . 108
4.1.2 Ongoing deployment of BREC for an install-free web access . . . 108
4.1.3 BREC 2.0 is on the way . . . 109
4.1.4 How can BREC serve the community? . . . 110
4.2 A new assembly pipeline to improve the assembly of repeat rich regions 111
4.3 Conclusion/Discussion : Towards mosquito genomes . . . 112

A Communications **115**
A.1 Conferences . . . 117
A.1.1 Automatic identification of heterochromatin boundaries through recombination rate estimates . . . 117
A.1.2 How transposable elements shape mosquito genomes . . . 117
A.1.3 Organization of insect genomes driven by some transposable element families . . . 117
A.1.4 How to improve genome assembly using repetitive elements . 117
A.1.5 In-depth analysis of the impact of transposable elements on genome assembly quality . . . 118
A.2 Journal Articles / Conference supplements . . . 118
A.2.1 Delorme, Q., Costa, R., Mansour, Y., Fiston-Lavier, AS., and Chateau, A. Involving Repetitive Regions in Scaffolding Improvement. To appear in Journal of Bioinformatics and Computational Biology special issue on RECOMB-CG2021. . . . 118
A.2.2 Mansour, Y., Chateau, A. and Fiston-Lavier, AS. BREC: an R package/Shiny app for automatically identifying heterochromatin boundaries and estimating local recombination rates along chromosomes. BMC Bioinformatics 22, 396 (2021). . . . 118

B Grants awarded and peripheral scientific activities **177**
B.1 The journey of my PhD towards bioinformatics . . . 177
B.2 Scholarships, fellowships and merit grants awarded . . . 178
B.3 Doctoral training > 200 hours . . . 179
B.4 Co-supervision experience . . . 179
B.5 Organizing and chairing experience . . . 179

C Case study **181**
C.1 First identification of chromatin regions in *Cx. pipiens* genome . . . 181
C.2 Analysis of the distribution of TEs along the *Cx. pipiens* genome . . . 181
C.2.1 *Cx. pipiens*: a genome enriched in DNA elements . . . 183
C.2.2 Centromeres are enriched in one type of TEs: LINE elements . . . 183

Bibliography **187**

List of Figures

1.1 Overview of data science ecosystem . . . 2
1.2 Sequencing cost per human genome . . . 3
1.3 Human sequences in GenBank . . . 4
1.4 The 5V model . . . 4
1.5 Geographic distribution of the West Nile virus . . . 5
1.6 Highlighting the multiple areas benefiting from mosquito genomics . . . 6
1.7 From a living cell to it's DNA sequence . . . 7
1.8 Barbara McClintock and TEs in Maize . . . 8
1.9 TE classification . . . 9
1.10 Distribution of TEs across the eukaryote phylogeny . . . 10
1.11 Crossing-event during meiosis . . . 13
1.12 Recombination variation . . . 14
1.13 From sequencing data to sequence assembly . . . 16
1.14 Different types of repeat-related assembly errors . . . 16
1.17 TEs distribution in mosquitoes and *D. melanogaster* . . . 17
1.15 Conservation of the content of chromosome arms across mosquitoes . . . 18
1.16 Genome size and TE content . . . 19

2.1 Summary of the three genomic-based approaches to infer the recombination landscape . . . 25
2.2 From inference to landscape . . . 27
2.3 Physical and genetic maps . . . 28
2.4 Marey map . . . 29
2.5 MareyMapOnline plots . . . 29
2.6 Comparison of recombination rate estimates between Marey map and population genetics based methods . . . 30
2.7 RRC for *D. melanogaster* chromosome 2 arms . . . 31
2.8 BREC workflow (overview) . . . 33
2.9 BREC workflow (detailed) . . . 35
2.10 Data cleaning process . . . 37
2.11 A schematic description of the chromosome type identification process implemented within BREC . . . 38
2.12 BREC result on different species . . . 42
2.13 Distribution simulations . . . 46
2.14 Variations of markers local density per 1-Mb bins along *D. melanogaster* Release 5 chromosomal arms . . . 48
2.15 Variations of markers local density per 5-Mb bins along the tomato genome *S. lycopersicum* 12 chromosomes . . . 49
2.16 Download, install and launch BREC . . . 51
2.17 Screenshots of BREC web application . . . 54
2.18 Screenshots of BREC web application - Genomic data web pages . . . 56

2.19 BREC workflow steps applied on chromosomal arm 2L of *D. melanogaster* Release 5 . . . 58
2.20 Plots representing results of BREC and reference HCB on the *D. melanogaster* genome . . . 60
2.21 Plots representing results of BREC and reference HCB on the *S. lycopersicum* genome . . . 63
2.22 The impact of decreasing markers density on the resolution of BREC's HCB expressed by the shift value . . . 65
2.23 Low density simulations . . . 68
2.24 Genomic features (right) and BREC results (left) for the *D. melanogaster R6* genome . . . 69
2.25 Comparison of regression models for recombination rate estimates along the five chromosomes (X, 2L, 2R, 3L, 3R) of *D. melanogaster* Release 5 . . . 71
2.26 Comparison of BREC *vs.* FlyBase recombination rate recombination rates along the five chromosomal arms (X, 2L, 2R, 3L, 3R) of *D. melanogaster* Release 5 . . . 72
2.27 Comparison of *AaegL4* and *CpipJ3* with genetic maps . . . 74
2.28 Genomic features (right) and BREC results (left) for the *Ae. aegypti AaegL4* genome . . . 75
2.29 The relationship between genetic (cM) and physical map (Mb) positions and estimated local recombination rates across the three chromosomes of a previous version of the *Ae. aegypti* genome . . . 76

3.1 Milestones in genome assembly . . . 82
3.2 Overview of the genome assembly process . . . 83
3.3 General workflow of the de novo assembly of a whole genome . . . 84
3.4 Currently available genomics technologies . . . 85
3.5 Overlap–layout–consensus genome assembly algorithm . . . 88
3.6 Two strategies for genome assembly: from Hamiltonian cycles to Eulerian cycles . . . 90
3.7 Example of a wrong assembly of a repetitive region. . . . 91
3.8 Repetitive sequences can lead to loops in a de Bruijn graph. . . . 91
3.9 Sequence errors and repeats lead to more complex k-mer graphs. . . . 92
3.10 Graph simplification techniques . . . 92
3.11 The difference to represent repeats in OLC and DBG graphs. . . . 93
3.12 Use of pairwise linkage information for scaffolding. . . . 95
3.13 Illustration of the difference between contigs and scaffolds in genome assemblies . . . 96
3.14 A scaffold graph . . . 97
3.15 Overview of the pipeline. . . . 98
3.16 RRs detection and characterization. . . . 100
3.17 PE Edge validation for case with only one contig carrying RR. Validation depends on position of RR within the contig. . . . 100
3.18 Analysis of the number of repeats on the extremities of *misassemblies* along the 2R chromosomal arm of *D. melanogaster* . . . 103
3.19 Number of TEs related to gaps, classified by type for *D. melanogaster R1 vs. R6* . . . 104

4.1 Screenshot of the ongoing deployment of BREC on the shinyapps.io platform. . . . 109

4.2 Screenshot of current BREC interface status. 110
4.3 Recombination landscape related questions 111

A.1 Vertical timeline for the main scientific communications related to this thesis project 116

B.1 Timeline from the scholarship contest to the steady progress in bioinformatics 177
B.2 Timeline for the grants awarded 178

C.1 Genomic features (right) and BREC results (left) for the *Cx. pipiens CpipJ3* genome. 182
C.2 *Culex* TE database construction pipeline. 183
C.3 Distribution of transposable elements by class in the *Cx. pipiens* genome. 184
C.4 Correlation of TEs distribution and heterochromatin boundaries in *CpipJ3* - chromosome 2. 185
C.5 LINEs distribution along *Cx. pipiens* chromosomes. 186
C.6 MITEs distribution along *Cx. pipiens* chromosomes. 186

List of Tables

1.1 Main features associated to the different chromatin domains in higher eukaryotes . . . 12
1.2 Statistics on the TE content of four diptera genomes . . . 17

2.1 Genomic features and BREC running time for the *D. melanogaster* Release 5 genome . . . 43
2.2 Genomic features and BREC running time for the *S. lycopersicum* . . . 44
2.3 BREC's built-in dataset of genomic data . . . 52
2.4 BREC HCB compared to reference boundaries from the reference genome of *D. melanogaster* . . . 59
2.5 Results of BREC and reference HCB on the genome of *S. lycopersicum* . . . 62
2.6 Comparing BREC with similar widely used tools . . . 78

3.1 Comparison of the read lengths, error rates, and costs of various DNA Sequencing Technologies. [table from (Bansal and Boucher, 2019), costs from (Logsdon, Vollger, and Eichler, 2020)] . . . 86
3.2 Quantification of missing DNA in the reference genomes of three model organisms . . . 86
3.3 Results on the *D. melanogaster* dataset . . . 102
3.4 Result on the *C. elegans* dataset (whole genome) . . . 102
3.5 Number of *misassemblies* on the *C. elegans* dataset, chromosome per chromosome . . . 103

C.1 Summary of resulting boundaries estimates with default model. . . . 181

List of Abbreviations

DNA	**D**eoxyribo**N**ucleic **A**cid
RNA	**R**ibo**N**ucleic **A**cid
NGS	**N**ext **G**eneration **S**equencing
TGS	**T**hird **G**eneration **S**equencing
TE	**T**ransposable **E**lement
SINE	**S**hort **I**nterspersed **N**uclear **E**lement
LINE	**L**ong **I**nterspersed **N**uclear **E**lement
TIR	**T**erminal **I**nverted **R**epeat
LTR	**L**ong **T**erminal **R**epeat
RR	***R**epetitive **R**egion*
HCB	**H**etero**C**hromatin **B**oundaries
BREC	heterochromatin **B**oundaries and **REC**ombination rate estimates
bp	***b**ase **p**air*
KB	***K**ilo **b**ase pair*
MB	***M**ega **b**ase pair*
GB	***G**iga **b**ase pair*
cM	***c**enti**M**organ*
D. melanogaster	***D**rosophila **melanogaster*** (Fruitfly)
Ae. aegypti	*A**e**des **aegypti*** (Yellow fever mosquito)
Cx. pipiens	***Culex pipiens*** (Common house mosquito)
An. gambiae	*A**n**opheles **gambiae*** (African malaria mosquito)
H. sapiens	***H**omo **sapiens*** (Human)
C. elegans	*C**a**enorhabditis **elegans*** (Worm)
S. lycopersicum	*S**o**lanum **lycopersicum*** (Domesticated tomato)
E. coli	***E**scherichia **coli*** (Bacterium)

Chapter 1

Introduction

Contents

1.1 Context: on the interface of computer science and evolutionary genomics . . . 1
1.1.1 From Computer Science to Bioinformatics: Data science for life sciences . . . 2
1.1.2 From Bioinformatics to Genomics: Computational biology for the study of whole genomes . . . 3
1.1.3 From Genomics to Evolution: Mosquito research interests . . . 5
1.2 Fundamental concepts . . . 6
1.2.1 Genome architecture . . . 6
1.2.2 Transposable Elements: one type of repetitive DNA . . . 7
1.3 Research scope: Genome dynamics . . . 11
1.3.1 Chromatin regions: one chromosome, different genomic profiles . . . 11
1.3.2 Recombination rate: one genomic feature, different landscapes . . . 12
1.3.3 Genome assembly: one genomic goal, different computational challenges . . . 15
1.4 Mosquitoes: an interesting model to aim for . . . 16
1.5 Thesis overview . . . 20

1.1 Context: on the interface of computer science and evolutionary genomics

Bioinformatics is a very recent research field, compared to biology, computer science and statistics. This discipline saw the light when genomic studies were unable to catch up to the enormous advances in whole genome sequencing technologies. For the last 20 years, bioinformatics gathers skillset from across multiple backgrounds. Mainly, there is molecular biology, computer science, data engineering, mathematics and statistics, and reaching further to chemistry, physics, electronics, among others. Figure 1.1 presents an overview highlighting computation-related skills. Therefore, being at the interface of such a wide variety of expertise is one of the major challenges facing this field as well as it's actors.

1.1.1 From Computer Science to Bioinformatics: Data science for life sciences

During this thesis project, we focus on approaching bioinformatics upon primarily three large aspects shown in Figure 1.1, customized to our study needs, as follows:

1. **Domain science: Biology** is the leading research interest, representing the concern to address, and providing the original data source as a starting point.
2. **Tool building: Computer Science** is the core layer. It consists of the conceptual and formal modeling of the biological problem, and the related computational resources allowing to develop a solution. Such complex tasks are achieved by a set of technical steps. To mention no to limit, the designed model goes through the process of implementing, testing, validating, experimenting, visualizing, and open sharing with the community, *via* easy and accessible automated tools.
3. **Data science: Statistics** is the theoretical design and formalism allowing to (1) infer the biological data into the computational model, (2) qualify and quantify the input data features, as well as interpreting the intermediate and the final solution outcomes. After all, it allows to evaluate and readjust the model parameter as per the quality of available data.

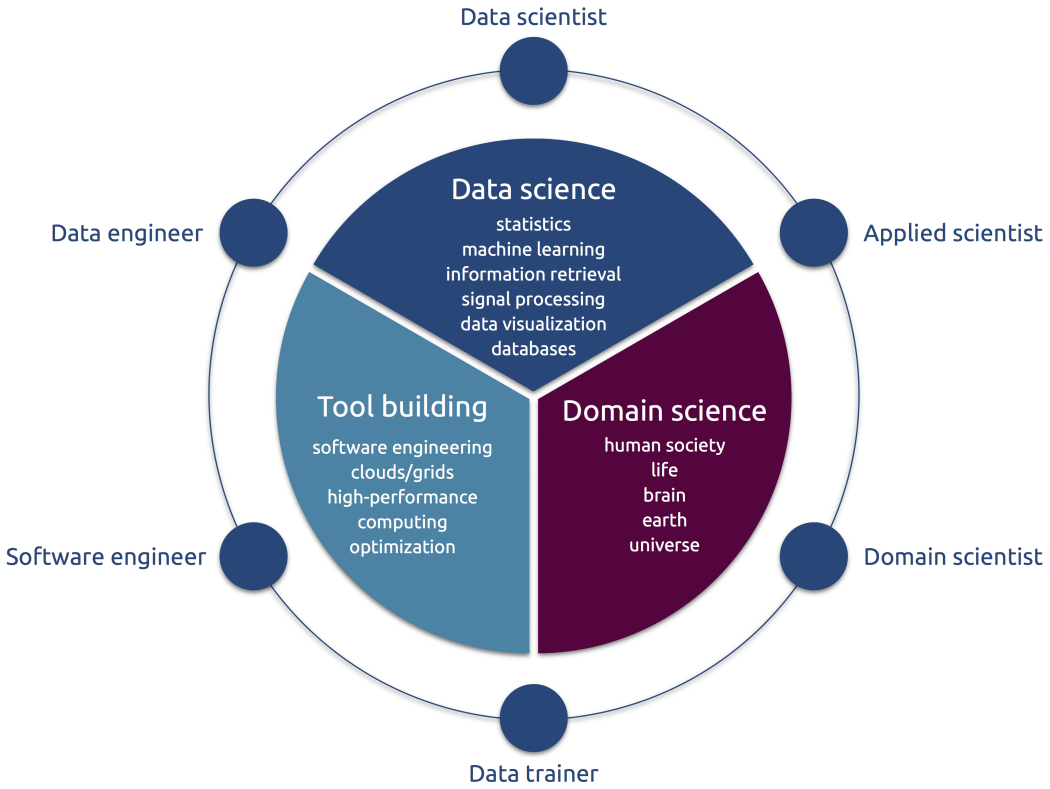


FIGURE 1.1: An overview of a representative data science ecosystem. [from http://www.datascience-paris-saclay.fr/data-science/]

It is important to note that throughout this project, we are not concerned with how data are generated nor with generating our customized input datasets. Instead, we will approach our biology research interests with a forward vision of how to develop a solution based on the already existing genomic data.

1.1.2 From Bioinformatics to Genomics: Computational biology for the study of whole genomes

Aiming to decipher the code of life, DNA, genomic studies across all three life domains (Archaea, Bacteria, and Eukarya) are becoming more and more accessible thanks to the next generation sequencing (NGS) platforms (Rice and Green, 2019). Over the past two decades, the biotechnology industry has revolutionized whole genome sequencing. In the case of the human genome, the chart representing the sequencing cost exhibits a significant drop since 2007, as reported by the NIH (National Institute of Health) in Figure 1.2). The human genome was first sequenced in 2001 and had cost around US$100 billion, while today it's for only less than US$1000. This was due to the emerging of Illumina® short-read sequencing technology (reads < 250 bp), which is still the most used since then. This technology has revolutionized the field since it consists of sequencing of several genomics fragments in parallel.

Both of these factors, the technology breakthrough along with the cost decrease, has resulted in huge amounts of genomic data ready to be analyzed, in order to help further the understanding of biology processes. However, most of the new genomes assembled are still in a draft phase as the assembly process of short-reads is a huge challenge, specifically for repeated regions (Consortium, 2001) (see Figures 1.3 and 1.2).

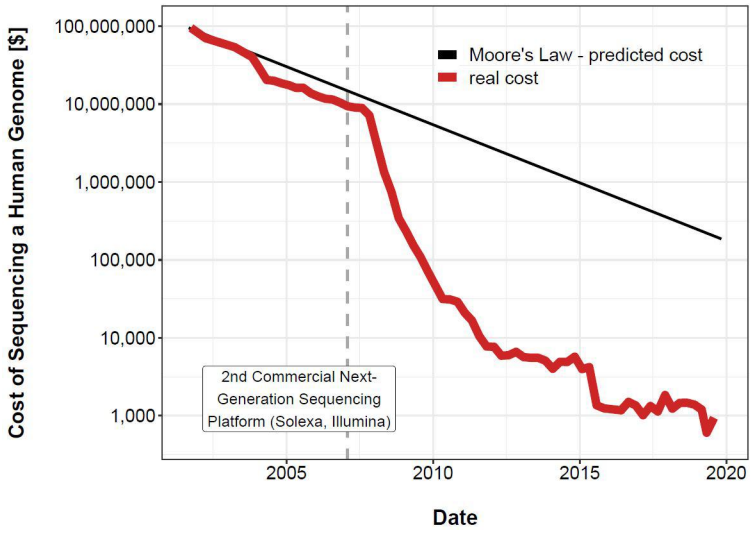


FIGURE 1.2: Sequencing cost per human genome -source: NIH, May 2020- [from https://www.reddit.com/r/singularity/comments/hi9rok/oc_the_cost_of_sequencing_the_human_genome/]

However, the field of bioinformatics is still unable to catch up to the data explosion, due to the lack of computational resources. Not only in terms of skillset and human expertise, but mainly in terms of developing automated solutions and tools (algorithms, programs, reproducible pipelines, experimental design, data visualization, user-friendly interfaces,...etc). Now, and more than ever before, life sciences are urgently in need of data science actors to handle the different challenges of big data. This latter is a tremendous sub-field of computer science that has been applied in almost all life areas. Its primary complications derive from the multitude of data facets. Figure 1.4 presents the simplest model of 5-Vs: volume, variety, value, velocity and veracity.

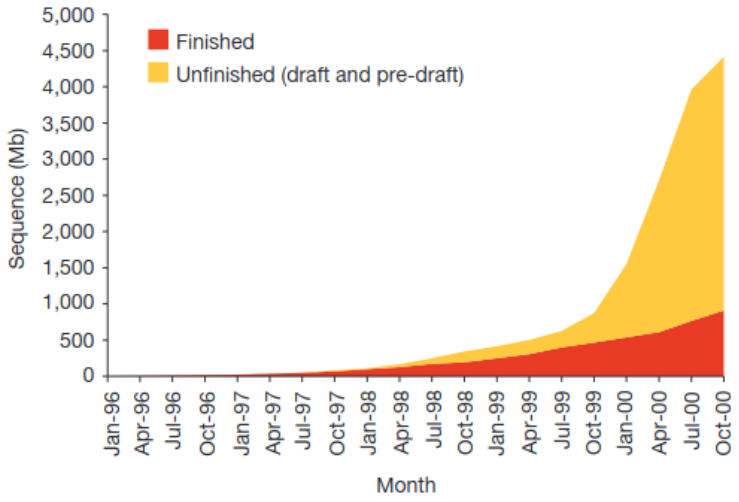


FIGURE 1.3: Total amount of human sequence in the High Throughput Genome Sequence (HTGS) division of GenBank. The total is the sum of finished sequence (red) and unfinished (draft plus predraft) sequence (yellow). [Figure by (Consortium, 2001)]

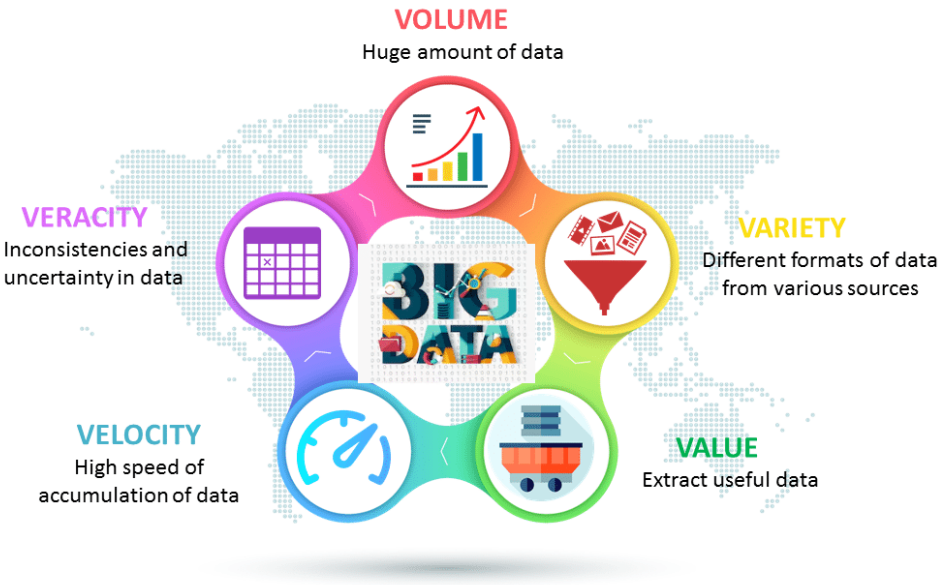


FIGURE 1.4: Data is the most valuable fuel that powers the 21st century world. Here is the 5Vs model distinguishing some of the challenging facets of Big Data. [from https://www.edureka.co/blog/big-data-characteristics/]

1.1.3 From Genomics to Evolution: Mosquito research interests

Since the rise of genomics, besides the most studied biological species, the fruit fly *Drosophila melanogaster*, and among the very interesting model organisms (human, mouse, thale cress, etc.), mosquitoes are increasingly catching scientists attention. Not only because of the early availability of the first draft genome, sequenced in 2002 (Holt et al., 2002), but mainly because of their capability to survive and adapt in a variety of different environments.

One of the best known evolutionary processes is *adaptation*. For example, an adaptation response allows mosquitoes to develop insecticide resistance as well as to survive extreme environments and brutal climate changes. As vectors of infectious human diseases (Malaria, Zika, yellow fever, dengue, chikungunya, ...), mosquitoes present highly complex biology challenges, in terms of their tremendous diversity, as well as the very rapid evolutionary processes observed in their genomes.

The latest mosquito enumeration has been reported in the *Medical and Veterinary Entomology* book (Foster and Walker, 2019) as follows: "Culicidae, the mosquito family, is comprised of 41 recognized genera incorporating about 3,500 species, many of which are vectors of disease pathogens that have afflicted humans and domestic animals for centuries, with devastating consequences for tens of millions of people". And to better understand the impact of a mosquito-borne pandemic, the world map in figure 1.5 emphasises the geographic distribution of the West Nile virus, transmitted by the *Culex* mosquito species.

Furthermore, mosquito genomics has proven interesting in numerous other fields of application. Figure 1.6 clarifies few examples where the contribution of human, plant and animal genomics in general, and mosquitoes in particular, has become fundamental.

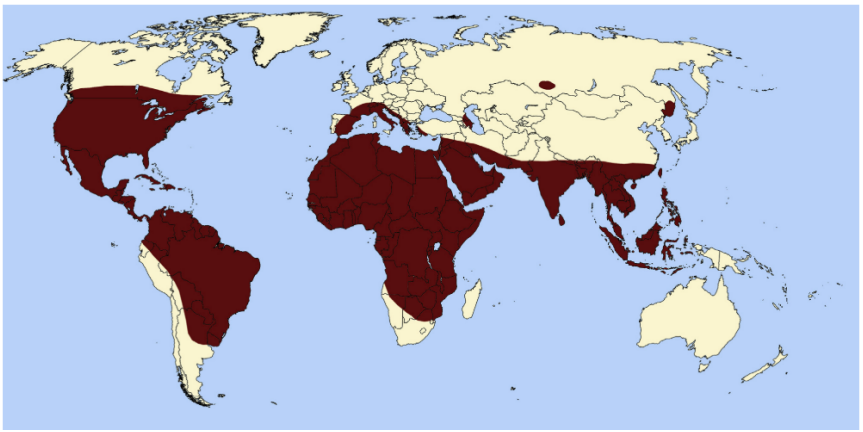


FIGURE 1.5: Geographic distribution of the *Culex* mosquito-borne West Nile virus, first appeared in 1937. Based on data from the Centers for Disease Control and Prevention, USA. (Foster and Walker, 2019)

More precisely, mosquito related research focuses on three main species: the host (human or vertebrate), the pathogen inducing the decease, and the mosquito being the vector responsible of transmitting the pathogen between host organisms.

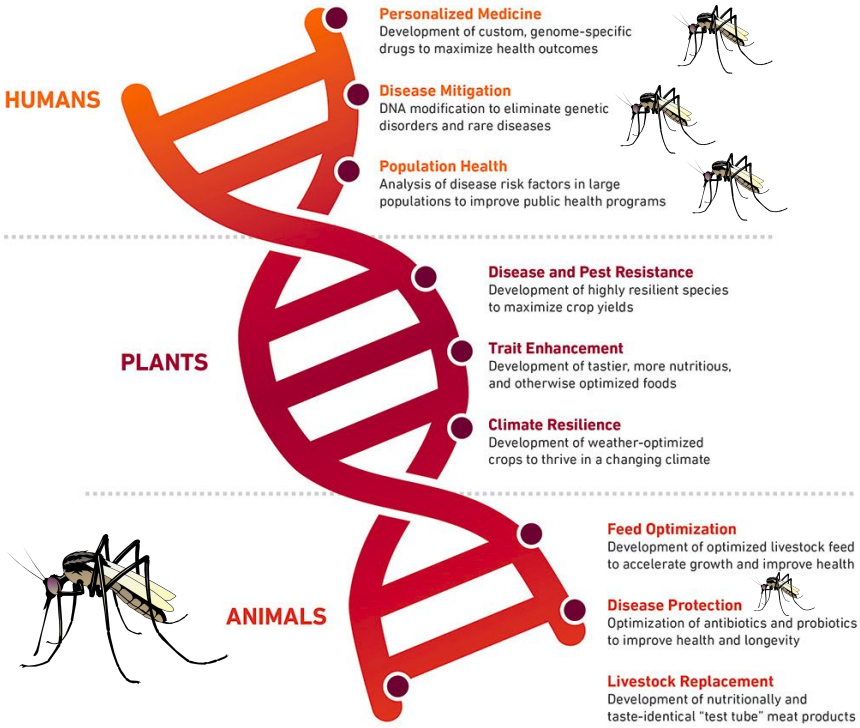


FIGURE 1.6: Highlighting the multiple areas benefiting from mosquito genomics. [Adapted from https://www.hardingloevner.com/big-data-infects-life-itself/]

In conclusion, the interdisciplinary research context of this thesis project asserts the urgent need of skillset and expertise from computation and data science backgrounds, in an attempt to close the gap, as fast as humanly possible, between the enormously available genomic data and the insightful knowledge that might be extracted from it, to the service of all life science domains: biology, health care, medicine, agriculture, biodiversity, environment, etc.

1.2 Fundamental concepts

1.2.1 Genome architecture

In evolution, an organism is a living entity representing a species on the tree of life (also known as the tree of species), such as a human, a mouse, a plant, an insect, etc. In eukaryotes, each organism is composed of a set of organs, like the brain in the human body, which is itself composed of cells, like the neurons of the nervous system in this case. The cell is a miraculous machinery that encompasses all the components able to make the (human) body functions properly (or not, when there is a disorder). The nucleus is the core part where all the genetic material is conserved, in other words: the genome (see Figure 1.7).

A genome is organised into a set of chromosomes. As represented in Figure 1.7, each chromosome in a eukaryotic organism consists of two parts, called *sister chromatids*, on which genes are carried. Sister chromatids are glued together thanks to a genomic component called a *centromere*, that often is located on the center of a chromosome. A *gene* is a piece of DNA that codes for a functional protein (exp. an

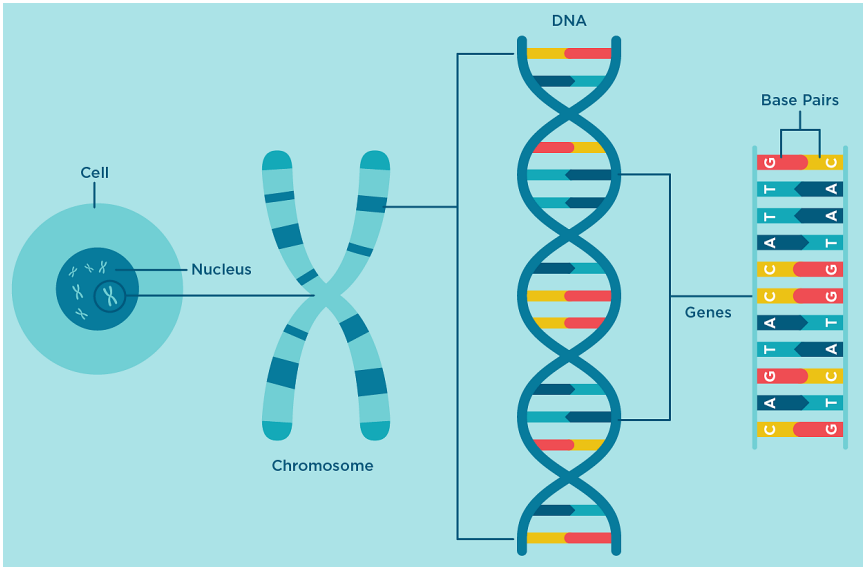


FIGURE 1.7: From a living cell to it's DNA sequence: a look at the scales of genomic data [from https://www.color.com/genetics-101-understanding-the-basics-of-genetics-891dc6b733be]

enzyme). The DNA sequence is basically built on 4 nucleotides referred to as the 4 alphabet letters A (Adenine), C (Cytosine), G (Guanine) and T (Thymine). And since the DNA double helix is formed of two complementary strands, each letter is exclusively associated with another one: A-T and C-G, also known as *base pairs*.

On top of coding DNA, there is non-coding DNA, that usually does not code for a protein. For a long time, non-coding DNA sequences was called as "junk DNA". It is mainly represented by repeated sequences (or repeats) of different sizes scattered between (or within) the genes, for which, recent studies are increasingly demonstrating its interest and function in the genome (Bernardi, 2021). A repetitive sequence can be considered as a substring that can be found in several occurrences in the main string on the alphabet $\{A, C, G, T\}$.

1.2.2 Transposable Elements: one type of repetitive DNA

Whole genome sequencing has revealed the importance of DNA repeats in terms of their impact on the structure and evolution of almost all genomes.

Among DNA repeats, the literature distinguishes three main categories. Firstly, the tandem repeats known as *satellites*. Secondly, the DNA sequence, encompassing any or several types of DNA elements (repeats and genes), which are duplicated within the genome and so called *duplications*. They could even represent a duplicate copy of the entire genome. Thirdly, the interspersed repeats, called *transposable elements* (TEs), and which are the focus of this thesis.

TEs discovery has revolutionized the genetics field

It was the observation of pigmentation in maize kernels that shed the light on the possibility of the existence of novel genetic elements which are responsible for such unusual coloration, as shown in Figure 1.8. Also called mobile DNA, TEs made

an outstanding leap in science, that they were worth the Nobel Prize in medicine or physiology in 1983. TEs were first discovered in maize by ***Barbara McClintock***, affiliated at the time of the award to Cold Spring Harbor Laboratory in New York, USA (McClintock, 1950; Ravindran, 2012).

FIGURE 1.8: Barbara McClintock, laureate of the 1983 Nobel Prize in medicine or physiology for her discovery of TEs in maize. [from https://www.nobelprize.org/prizes/medicine/1983/summary/]

In the editorial paper entitled *"The mobile world of transposable elements"*, (Navarro, 2017) describes the historical reversal of TEs value, from when they were first discovered and up till currently, as follows:

> "It has been almost 70 years since Barbara McClintock first suggested that elements exist that have the capacity to move and reshape the genome, and that these elements could potentially control gene expression. At first met with skepticism and considered to be 'junk' or 'selfish' pieces of DNA, TEs have now been shown to be major components of the genome with the ability to influence genome evolution and function. Today, TEs have been shown not only to regulate host gene expression but are often co-opted by the host to serve new cellular functions."

TEs characteristics

Mobile DNA consists of dispersed, diverse, and highly repetitive sequences. TEs are classified in two main classes based on their transposition mechanism. Class I elements also called *RNA retrotransposons* transpose *via* a copy-and-paste mechanism while the Class II elements also called *DNA transposons* transpose *via* a cut-and-paste mechanism. Within each class, one can classify them into Subclasses, Superfamilies and Families based on their sequence features. There are two subclasses in the class I : retrotransposons with and without LTR (Long-Terminal Repeats). within the non-LTR elements, one can distinguish the LINE elements for *Long Interspersed Nuclear Elements* that are autonomous and the SINE elements for *Short Interspersed Nuclear Elements* that are non-autonomuos. Among the DNA elements, we can also identified

autonomous elements called TIR for *Terminal Inverted Repeats* and non-autonomous elements called MITE for *Miniature Inverted-repeats TEs* (see Figure 1.9).

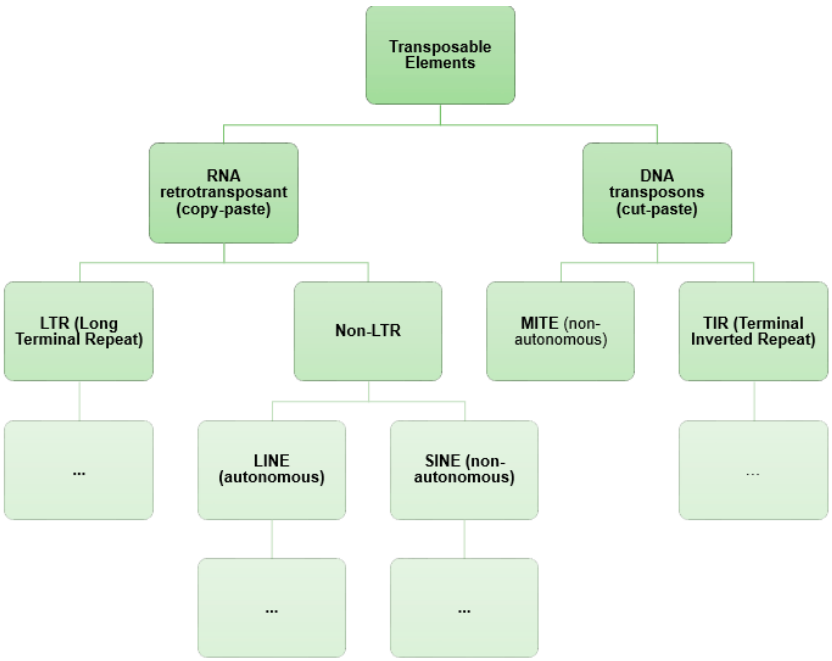


FIGURE 1.9: TE classification - Adapted from (McCullers and Steiniger, 2017).

TEs are ubiquitous and not evenly distributed across eukaryotic taxa

Advances of genome sequencing technologies have allowed to identify and analyze TEs more easily and accurately across a wide range of eukaryotic taxa. So far, TEs have been detected within the large majority of sequenced genomes suggesting their role in genome dynamics. Figure 1.10 by (Wells and Feschotte, 2020) gives an overview of TEs in the main genomes across eukaryotic taxa (human, animals, and plants). TE abundance can vary drastically from one species to another one, often associated with the genome size (*e.g.* from 85% in maize *Zea mays* (Anderson et al., 2019) to 4.25% in honey bee *Apis mellifera* (Petersen et al., 2019)).

The proportion of TE types also varies between taxa. For example, the maize genome is composed of almost 70% of LTR elements while in zebrafish *Danio rerio* more than 25% of the genome is composed of DNA elements (Figure 1.10).

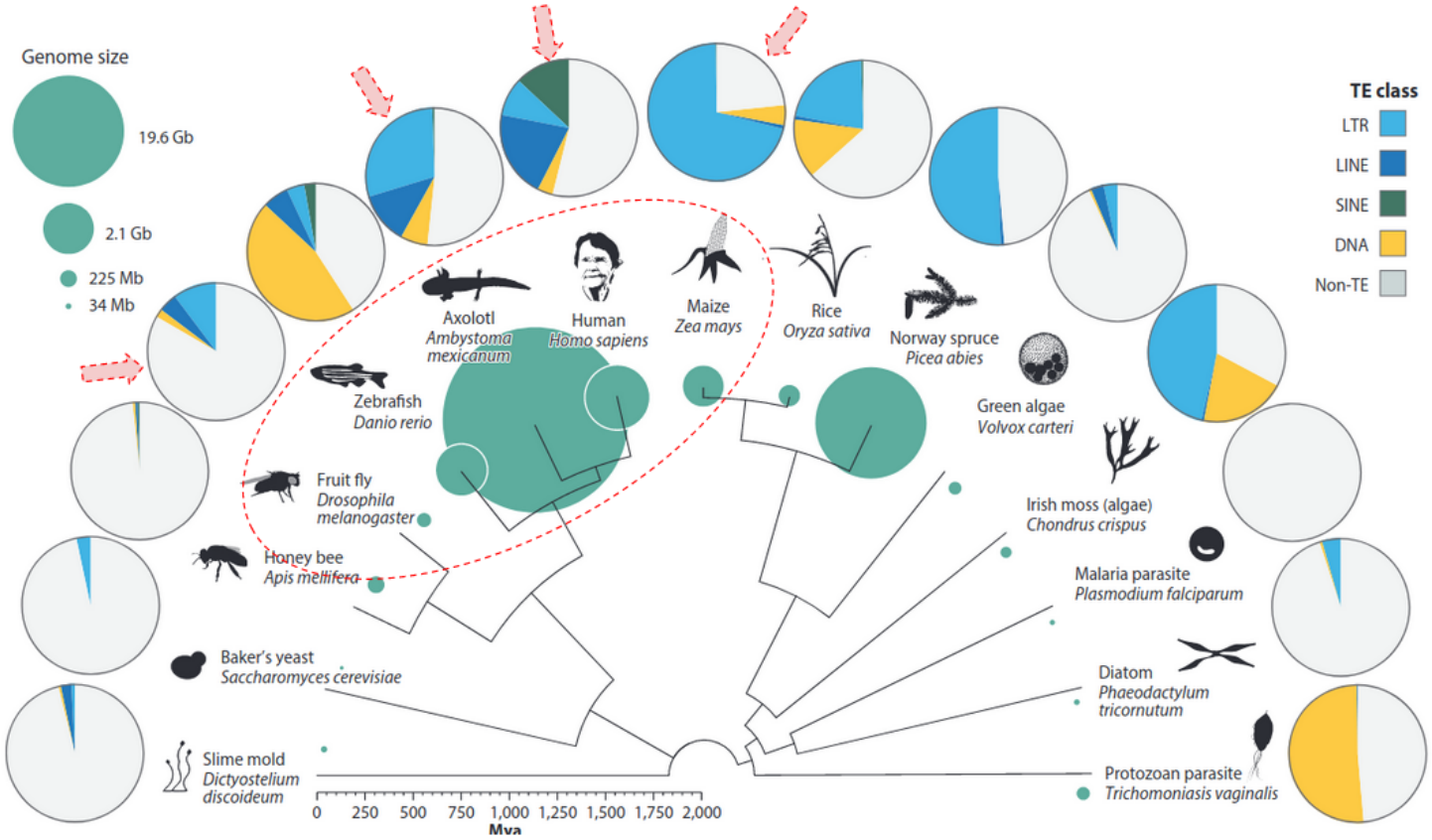


FIGURE 1.10: Distribution of TEs across the eukaryote phylogeny. Reference genome size (sea green circles) varies dramatically across eukaryotes and is loosely correlated with TE content. Here, the honey bee TE content is likely an underestimate, as approximately 3% of the genome derives from unusual large retrotransposon derivatives (Elsik et al., 2014). For ease of visualization, YR retroelements have been included with LTRs and all class II elements are included under DNA. Data were acquired from genome RepeatMasker output files. Figure adapted with permission from (Huang, Burns, and Boeke, 2012); the Volvox characteristic silhouette was provided by Matt Crook. (Imperviously used abbreviation: YR, tyrosine recombinase). [Adapted from (Wells and Feschotte, 2020)]

TEs are increasingly demonstrated to be involved in numerous biological processes of the host organisms, and thus are becoming an interesting resource to exploit when investigating the dynamics of genomes within and between species. For example, some specific TE types play a role in inducing chromosomal rearrangements, while others have an impact on the expression or repression in genes (Bourque et al., 2018).

1.3 Research scope: Genome dynamics

A genome is far away from being a stable entity. It does not only vary between species, but also within species on multiple scales, such as between populations from different geographical locations, between individuals sharing, or not, the same ancestors, within the same genome, and even along one specific chromosome. Furthermore, genomic variation is induced by numerous factors simultaneously, which results in a set of genomic behaviours related to its structure, architecture, expression, evolution, etc, which could be referred to as *genome dynamics*.

Among the genetic factors impacting the majority of genome dynamics, there is a large set from which scientists choose to focus on according to each research project they conduct, and more specifically, with regard to the biological questions to be addressed.

During my thesis project, and in order to dive in the field of bioinformatics by bridging my background in computer science with my scientific interests in genomics, we chose to focus on three major players impacting genome dynamics:

1. Chromatin structure: unevenly compacted along chromosomes;
2. Meiotic recombination: exchanging DNA fragments during cell division;
3. Repetitive DNA (especially TEs): inducing genome assembly errors.

One of the major genomic factors interfering with TEs behavior is meiotic recombination (Kent, Uzunović, and Wright, 2017). The significant variation in recombination rates is strongly correlated with TEs distribution (Rizzon et al., 2002; Petrov et al., 2011; Kent, Uzunović, and Wright, 2017) and diversity within various genomes.

In order to efficiently address the cause-effect relationship between TEs and recombination, genome-wide local recombination rates are vital. However, recombination maps are not so often available.

1.3.1 Chromatin regions: one chromosome, different genomic profiles

Among the genomic features, one can distinguish two primary domains of chromatin. Table 1.1 highlights the main differences (Termolino et al., 2016), where:

1. Euchromatin, is lightly compact with a high gene density;
2. Heterochromatin, is highly compact due to specific proteins or chromatin modification, and with a paucity in genes.

Feature	Euchromatin	Heterochromatin
Structure	Loosely packed, open, accessible	Densely packed, closed, inaccessible
Composition	Mainly genes	Mainly repetitive elements
Activity	Expressed, active	Repressed, silent
DNA methylation	Hypomethylation	Hypermethylation

TABLE 1.1: Main features associated to the different chromatin domains in higher eukaryotes. [Cropped from (Termolino et al., 2016)]

The heterochromatin is represented in different chromosome regions mainly within the centromere and telomeres. Euchromatin and heterochromatin regions exhibit different behaviors in terms of genomic dynamics related to their biological function, such as the cell division process that ensures the organism viability. Consequently, easily distinguishing chromatin domains is necessary for conducting further studies in various research fields and to be able to address questions related to cellular processes such as meiosis, gene expression, epigenetics, DNA methylation, natural selection and evolution, genome architecture and dynamics, among others (Chan, Jenkins, and Song, 2012; Stapley et al., 2017; Morata et al., 2018).

1.3.2 Recombination rate: one genomic feature, different landscapes

Meiotic recombination is a major evolutionary force - Meiotic recombination is a vital biological process which guarantees the diversity of genetic material over generations. This process consists on the exchange of DNA fragments within and between chromosomes. Figure 1.11 illustrates the parental homologous chromosomes which duplicate during meiosis, and then recombine *via* a crossing-over event. This process increases the genetic diversity carried by new chromosomes in the descendants. Consequently, recombination plays an essential role in investigating genome-wide structural and functional dynamics. Recombination events are observed in almost all eukaryotic genomes. Recombination is a fundamental process that ensures genotypic and phenotypic diversity. Thereby, it is strongly related to various genomic features such as gene density, repetitive DNA, and thus also chromatin domains (Coop and Przeworski, 2007; Duret and Galtier, 2009; Auton and McVean, 2012).

Recombination rate variation highlights heterochromatin regions Recombination rate varies not only between species but also between populations, between sexes, between individuals, within individuals, as well as between and within chromosomes. Along chromosomes, different chromatin domains can be identified based on their recombination rate intensity (from low to high). This variation is a composite within-chromosome variation due to well-known genomic features such as open chromatin regions and crossing-over inference (see Figure 1.12).

Besides, the recombination landscape of numerous genomes exhibits an interestingly unique profile along heterochromatin regions, particularly, the telomeres, the ending parts of the chromosome which mainly protects the DNA sequence during cell division, and is directly associated to cell aging aspects, and the centromere, which connects both sister chromatids.

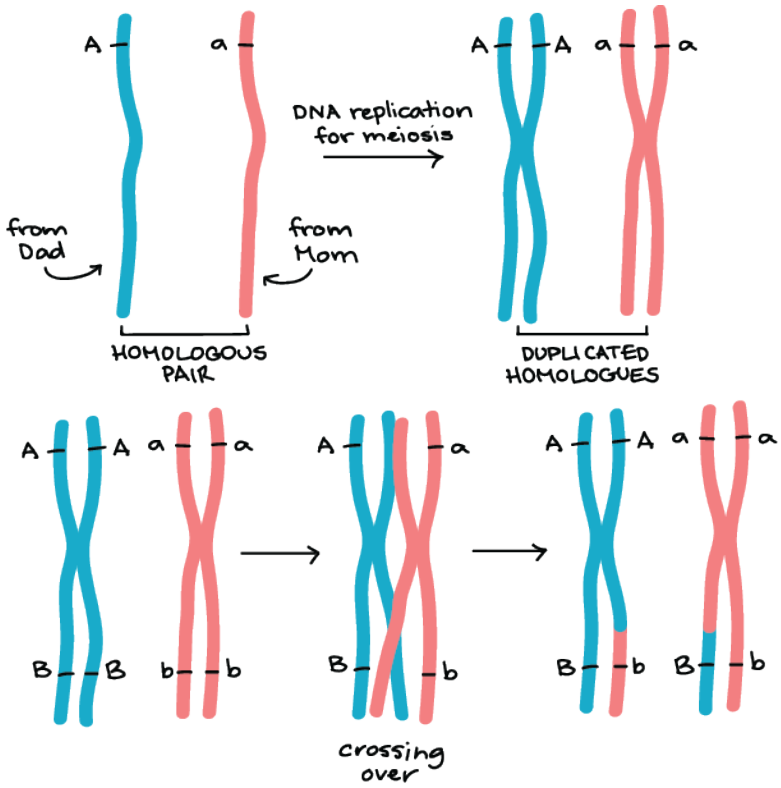


FIGURE 1.11: Illustration of the parental homologous chromosomes which duplicate during meiosis, and then recombine *via* a crossing-over event. [from https://www.khanacademy.org/science/ap-biology/heredity/non-mendelian-genetics/a/linkage-mapping]

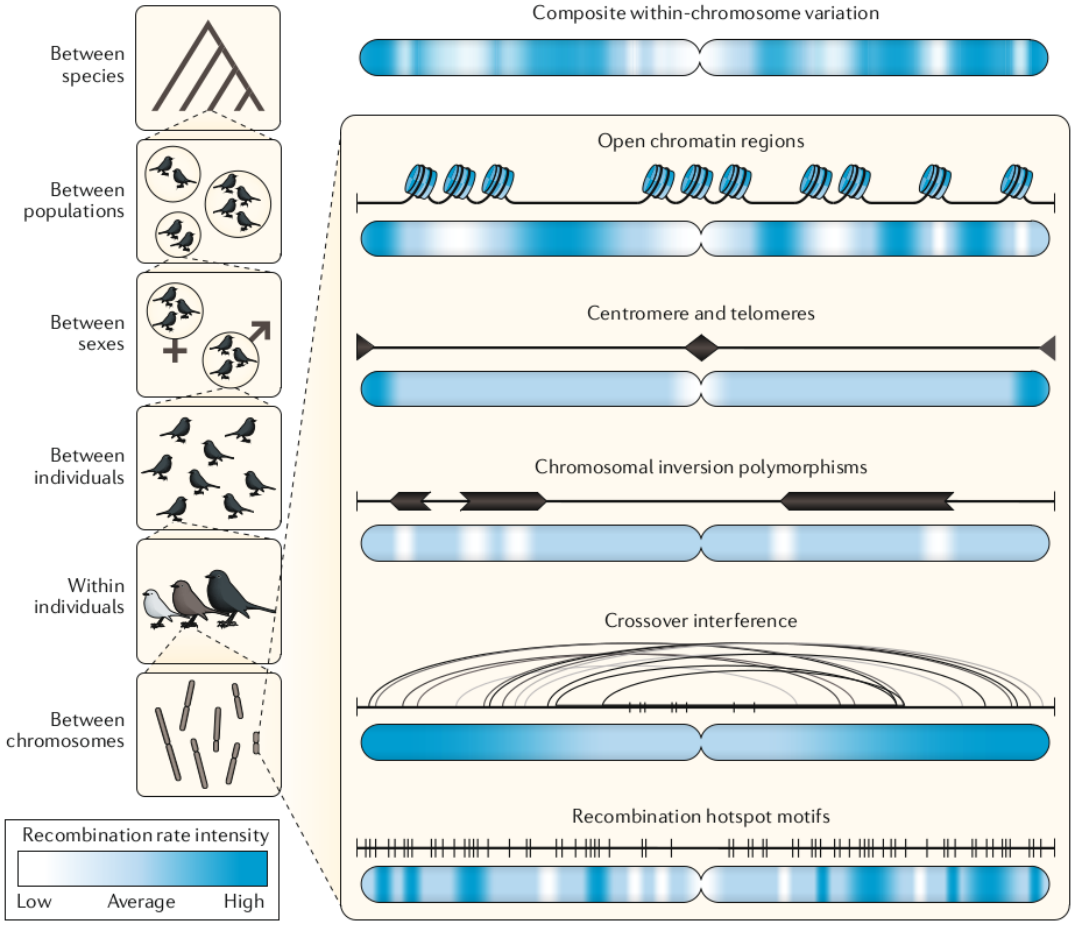


FIGURE 1.12: **Recombination variation.** The left part of the figure shows the different levels of biological organization within and between which recombination can vary. The right part shows some of the molecular mechanisms that affect within-chromosome variation. The chromosome below each mechanism depicts how that mechanism would affect the direction of recombination rate modification. The schematic on top depicts the overlaid recombination rate across the chromosome resulting from all processes. [From (Peñalba and Wolf, 2020)]

The understanding of centromeres structure, organization, and evolution is currently a hot research area. Besides, the highly diverse mechanisms of centromere positioning (Vanrobays, Thomas, and Tatout, 2017) and repositioning (Lu and He, 2019) remain a complicated obstacle in the face of fully understanding genome dynamics. Thus, generating high resolution genetic, physical, and recombination maps, as well as locating heterochromatin regions is increasingly attractive to the community across an extensive range of taxa (Schueler et al., 2001; Weinstock et al., 2006; Silva-Junior and Grattapaglia, 2015; Robert L. Nussbaum, McInnes, and Willard, 2015; Shen et al., 2017; Gui et al., 2018; Rowan et al., 2019). Despite the enormous advances offered by sequencing technologies, centromeres are still considered enigmas, mostly due to their enrichment in repeat DNA that prevents genome assembly algorithms to achieve more complete whole genome sequences (Muller, Gil, and Drinnenberg, 2019).

1.3.3 Genome assembly: one genomic goal, different computational challenges

Genome assembly is the process consisting of putting together DNA fragments from sequencing with the aim of reconstruct the original form of a genome.

Genome assembly goes through mainly two steps as shown on Figure 1.13:

1. DNA fragments, which are the sequencing data (i.e. reads), are brought together to form longer sequences, called *contigs*;
2. The contigs are then oriented, ordered and connected to form more complete sequences which may hopefully reach the chromosome-length, called **scaffolds**.

Genome assembly has become crucial for conducting genomic studies in various fields as environment, health, genetics, evolution and many more. Thus, recent studies has highlighted the impact of assembly quality on result interpretations, that could be biased due to low quality genomes (Chakraborty et al., 2018).

While the efficiency of bioinformatic tools used for assembly is increasing, errors of sequence construction persist. One of the most common sources of such errors is repeated regions, including TEs, as they are known for causing misassemblies (*i.e. assembly errors*. The presence of repeated elements can induce (1) chimeric contigs due to collapsed repeats and/or (2) assembly breaks (Treangen and Salzberg, 2011).

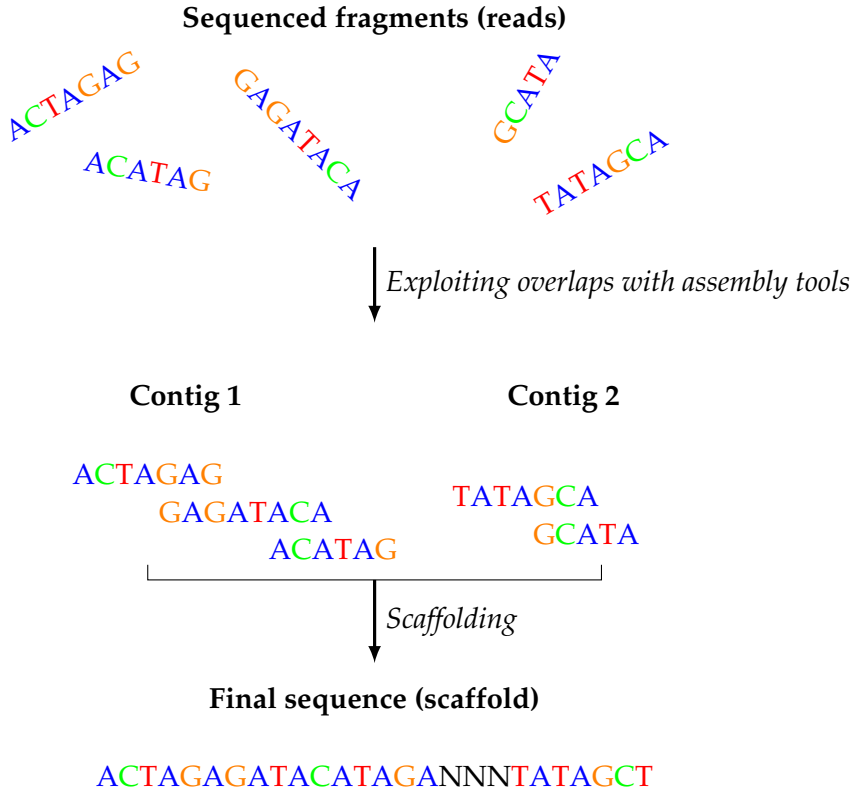


FIGURE 1.13: From sequencing data (reads) to sequence assembly (chromosome-length scaffold).

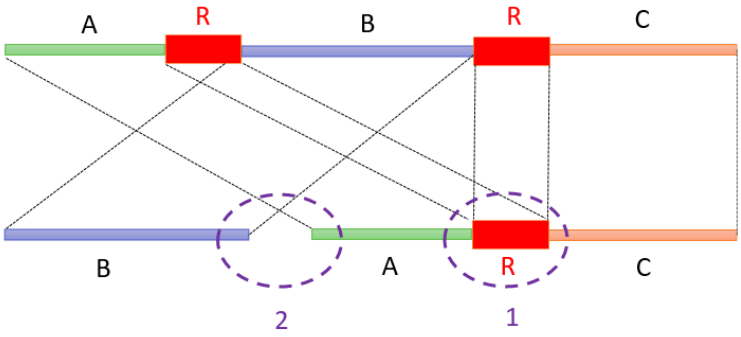


FIGURE 1.14: Different types of repeat-related assembly errors. This is a scenario of a DNA sequence composed of three contigs A, B, C and two copies of the same repeat (R).

1.4 Mosquitoes: an interesting model to aim for

Following one of the main research topics studied at ISE-M[1], my interest in mosquitoes has been driven by the remarkable adaptation response their genomes manifest. Previous studies highlighted a strong variation across different mosquito species in terms of genome size, while their gene content is conserved. Figure 1.15 by (Duddchenko et al., 2017) shows the phylogeny of four diptera genomes: the most famous

[1]Institut des Sciences de l'Évolution de Montpellier, France

three mosquito species, *Aedes aegypti* (1.3Gb), *Culex pipiens* (579Mb), *Anopheles gambiae* (278Mb), and the fruit fly *D. melanogaster* (180Mb).

Indeed, these closely related species exhibit a high variability in terms of genome size, but also their content of repeat sequences including TEs (Figure 1.16). TEs are suspected to have caused genome size expansion mainly in the heterochromatin regions (Morata et al., 2018). Meanwhile, such regions highlight reduced recombination rates. Thus, being a source of genomic diversity and novelty, TEs are good candidates to investigate the adaptive evolutionary process within genomes (Biémont, 2010).

Despite the correlation between the genome size and TEs coverage, TEs are not always present with the same types across the four genomes. Table 1.2 reports the annotated TE sequences (in Mb) for the DNA, LINE, LTR, SINE, and Unknown transposons. These numbers are represented by the Figure 1.17, which allow to clearly visualize the variation of TE density, per type, as well as the proportion to the total TE content within the species. Moreover, for this level of classification, we observe that TE types are identical between the three mosquitoes, but with different proportions. More importantly, these proportions strongly varies with the fruit fly genome, which does not include SINEs at all.

Species	Genome size	DNA	LINE	LTR	SINE	Unknown	Total	Coverage(%)
Ae. aegypti	1383.97	284.58	170.62	74.48	19.27	224.71	773.66	55.90
Cx. quinque.	579.04	148.83	19.23	12.53	10.42	82.12	273.13	47.17
An. gambiae	265.01	14.56	8.43	6.90	2.38	13.96	46.24	17.45
D. melanogaster	143.73	1.86	6.20	14.98	0.00	4.41	27.45	19.10

TABLE 1.2: Statistics on the TE content of four diptera genomes, listing the genome assembly size as well as the genome coverage of DNA, LINE, LTR, SINE, and Unknown transposons (in Mb). [Adapted from (Petersen et al., 2019)]

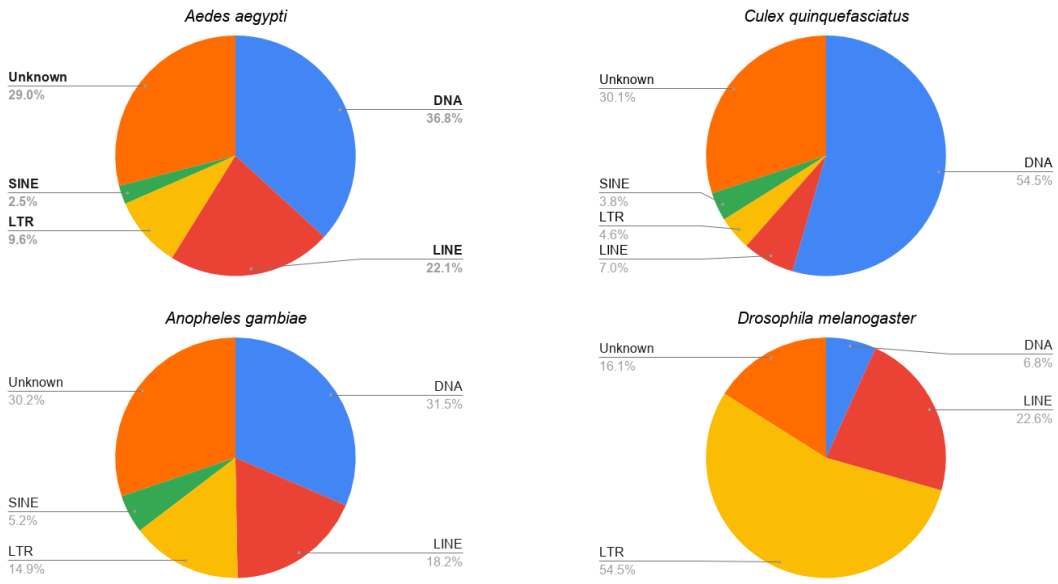


FIGURE 1.17: TEs distribution in mosquitoes and *D. melanogaster*: pie charts reproduced with data of Table 1.2 from (Petersen et al., 2019)

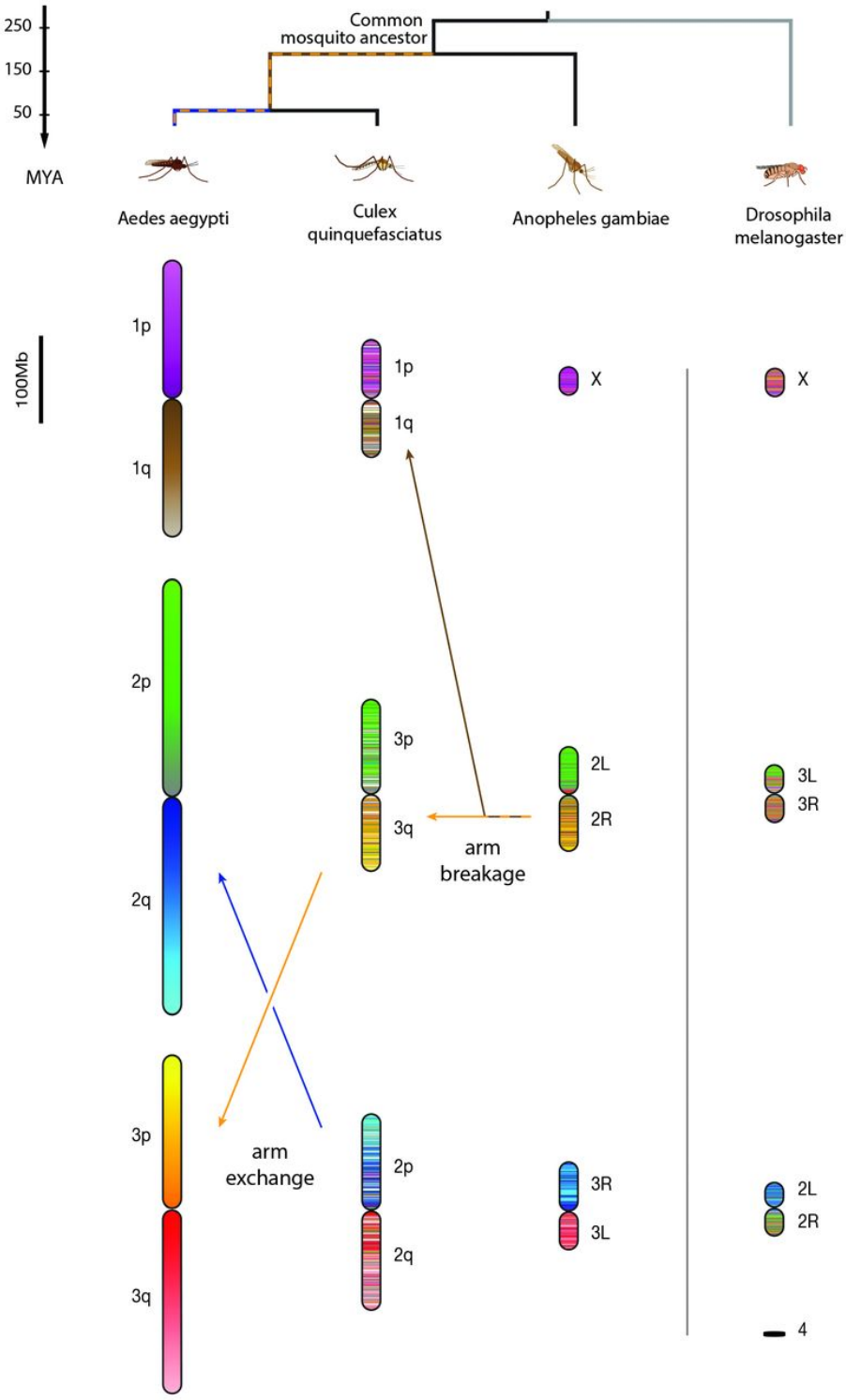


FIGURE 1.15: The content of chromosome arms is strongly conserved across mosquitoes. Here each 100-kb locus in *Ae. aegypti* is assigned a color. For the other species, each 100-kb locus is assigned a combination of the colors of the corresponding DNA sequences in *Ae. aegypti*, weighted by length. (MYA) million years ago. [Figure by (Dudchenko et al., 2017)].

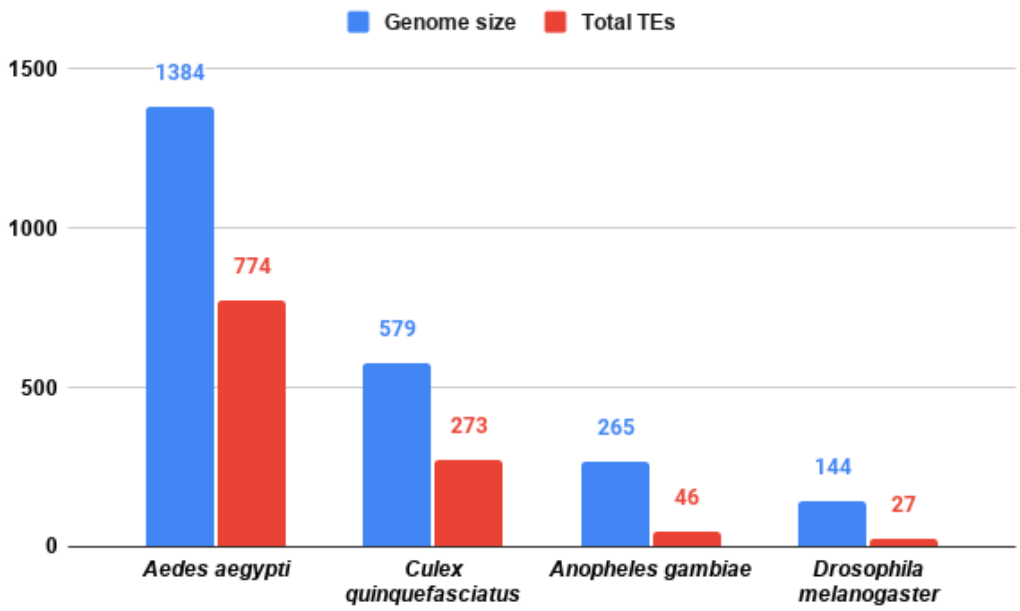


FIGURE 1.16: Genome size and TE content [data from (Petersen et al., 2019)].

Upon such observations, natural questions arise. Are TEs responsible for such genome size expansion? If so, which type of TEs is the most influencing these genomes? Are TEs influencing all chromosomes, within and between diptera species in the same way? What about on the same chromosome? Are there specific regions that are more affected than others, like in euchromatin *vs.* heterochromatin? More precisely, are centromeric and telomeric regions exhibiting any special TE-related patterns?

For example, what is the abundance and distribution of TEs? Which TE types, called families, are mostly present in a genome? Is there a correlation between TEs and other genomic features like gene density and recombination rate among others? By collecting enough knowledge on TEs organization and dynamics, the scientific community will be able to investigate their impact on genomes architecture and dynamics, as in chromosomal rearrangements.

Therefore, there is still plenty of issues to handle, such as the quality and completeness of genomes, which, among other factors, influence the quality of the TE annotations. Besides, the capacity of identifying the different genomics regions: chromatin domains.

1.5 Thesis overview

The rest of this PhD manuscript is organized in three chapters such as :

Chapter 2 - An automated computational tool that I develop to identify heterochromatin boundaries along chromosomes and estimating local recombination rates is presented. The tool based on the Marey maps method is called **BREC** for **B**oundaries and **REC**ombination rate estimates.

Chapter 3 - Focusing on the scaffolding step with the aim of enhancing the assembly quality, an approach that exploited the repeated regions had been proposed.

Chapter 4 To conclude my thesis project with a showcase of the previous results, we present an opening regarding the genome dynamics. We provide some insights on the perspectives of the work presented here and how it may be extended to further the understanding of the related research topics. Furthermore, we present a preliminary case study in Appendix C where we focus on the analysis of TEs distribution in mosquito genomes to raise few perspectives.

Additional content As part of my research activity, I had the opportunity to present my results in numerous scientific events including national and international conferences. The set of my publications consists of various posters, one talk and one journal published article (see details in Appendix A). Also, I list the grants I got awarded as well as the peripheral scientific activities I took part in (see Appendix B).

Chapter 2

Recombination and heterochromatin regions

Contents

2.1 Context and motivation 24
2.1.1 Approaches for estimating recombination rate variation . . 24
2.1.2 An approach to estimate quickly and easily the recombination rate along chromosomes 28
2.2 New Approach: BREC 31
2.2.1 Step 0 - Apply data pre-processing 36
2.2.2 Step 1 - Estimate Marey Map-based local recombination rates 37
2.2.3 Step 2 - Identify chromosome type 37
2.2.4 Step 3 - Prepare the HCB identification 39
2.2.5 Step 4 - Identify centromeric boundaries 39
2.2.6 Step 5 - Identify telomeric boundaries 40
2.2.7 Step 6 - Extrapolate the local recombination rate estimates and generate interactive plot 40
2.3 Validation process 41
2.3.1 Validation data 41
2.3.2 Simulated data for quality control testing 45
2.3.3 Validation metrics 50
2.3.4 Implementation and Analysis 50
2.3.5 Description of main components of the Shiny app 51
2.4 BREC results 56
2.4.1 Approximate, yet congruent HCB 59
2.4.2 Consistency despite the low data quality 64
2.4.3 Accurate local recombination rate estimates 69
2.4.4 BREC is non-genome-specific 73
2.4.5 Easy, fast and accessible tool *via* an R-package and a Shiny app 73
2.5 Applying BREC to identify chromatin regions along the mosquito genome: *Ae. aegypti* 73
2.6 Discussion and Conclusion 76

2.1 Context and motivation

In this chapter, we aim to address the issue of identifying the eu-heterochromatin boundaries, in order to distinguish the two main genomic regions, euchromatic and heterochromatic, and mainly for the latter, to localise the centromeric and telomeric regions along each chromosome. This aspect will allow us to address the various dynamics of such genomes, by conducting a deeper analysis according to the different chromatin domains. Besides, along chromosomes, and later on, investigate its correlation with the TE density and distribution.

Therefore, we chose to start by exploiting the previously available datasets, along with the existing grounds on recombination, in terms of fundamental knowledge, biological experimentation results, statistical analysis tools and computational implementation, in order to gather the various essential elements which will guide us towards a better understanding of genome dynamics in mosquitoes.

2.1.1 Approaches for estimating recombination rate variation

Numerous methods for estimating recombination rates exist. Genomics inference methods, covering population-based, pedigree-based , and gamete-based approaches (Auton and McVean, 2012; Peñalba and Wolf, 2020), are used to estimate the variation of recombination rates at different scales (see Figure 2.1):

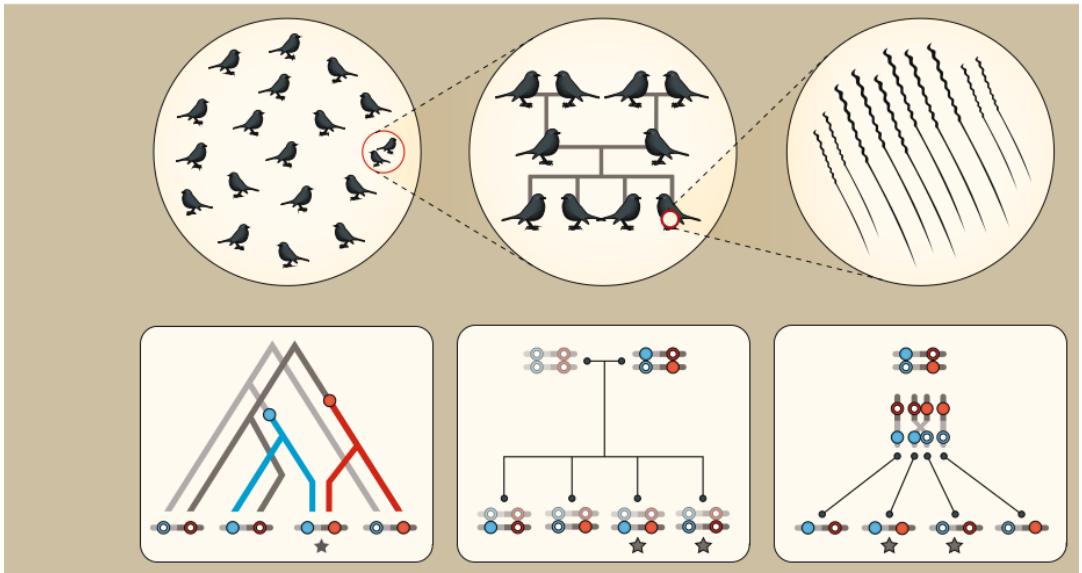
1. The population-based approach estimates the recombination rate within a population, *i.e.* a group of individuals of the same species living in the same geographical zone. This approach provides fine-scale genome-wide recombination estimates (1-5Kb). However, it requires at least 10000 generations analysed.
2. The pedigree-based approach estimates the recombination rate within one family of individuals, which are closely related, *i.e.* parents and their descendants. This approach provides average-scale genome-wide recombination estimates (10Kb-5Mb). But it requires only from 1 to 10 generations.
3. The gamete-based approach also known as the sperm typing method, because they are applicable on males only but they are limited to small regions of the genome.

Among the listed methods, population genetic-based methods (Stumpf and McVean, 2003) provide accurate fine-scale estimates. Nevertheless, these methods are costly, time-consuming, require substantial expertise, and most of all, do not apply to all kinds of organisms. Moreover, the sperm-typing method (Jeffreys, 2000), which is also extremely accurate, providing high-density recombination maps, is male-specific and is applicable only on limited genome regions. On the other hand, a purely statistical approach, the Marey Maps (Chakravarti, 1991), could avoid some of the above issues based on other available genomic data: the genetic and physical distances of genomic markers.

We have to compromise between the number of data analysed and the results resolution, which will drive the choice of the approach used.

Unfortunately, some data types are more rare than others or more challenging to obtain and generate to get the appropriate resolution, and that is one of the various

reasons that motivate our interest in developing an automated and user-friendly solution.



Parameter	Population-based	Pedigree-based	Gamete-based
Estimate	Population recombination rate	Genetic distance	Crossover frequency
Sample size	10–30	>100	≥1
Pros	Low to moderate sampling Moderate to high resolution Simple study design	Sex-specific recombination maps Unbiased by population level processes Variation between individuals and heritability can be measured Can improve genome assemblies	High resolution Low sampling required Variation between individuals can be compared
Cons	Average between sexes and through time Biased by effective population size, demographic history, selection and mutation Dependent on genome assembly quality	Large sampling required Complex and time-intensive study design required Low resolution (~1 Mb, 0.5–2 cM) Biased by SNP density in mapping population Only contemporary snapshot of recombination	Single-sex recombination map (usually males) Biased by individual SNP density Dependent on genome assembly quality Only contemporary snapshot of recombination

FIGURE 2.1: **Summary of the three genomic-based approaches to infer the recombination landscape**. In the schematic figure, grey stars indicate recombination events. cM, centiMorgans; SNP, single-nucleotide polymorphism. [from (Peñalba and Wolf, 2020)]

According to (Peñalba and Wolf, 2020), it would be interesting if the community aim for a unified approach as an attempt to include the complementary advantages of each of the three, and avoid their limitations as much as possible. The complementary Figure 2.2 sheds the light more closely on the difference in data types between the three approaches mentioned above. Despite the dissimilarities in terms of estimates accuracy, the recombination rate variation of the three different approaches (a), (b), and (c) converge towards the same recombination landscape. Furthermore,since population-based and gamete-based approaches present numerous limitations in terms of data availability, particularly for the non-model organisms where datasets are rare or difficult to access, we believe the most feasible solution

is the pedigree-based approach, shown in (b) of the same figure, with increasingly available data (Corbett-Detig, Hartl, and Sackton, 2015).

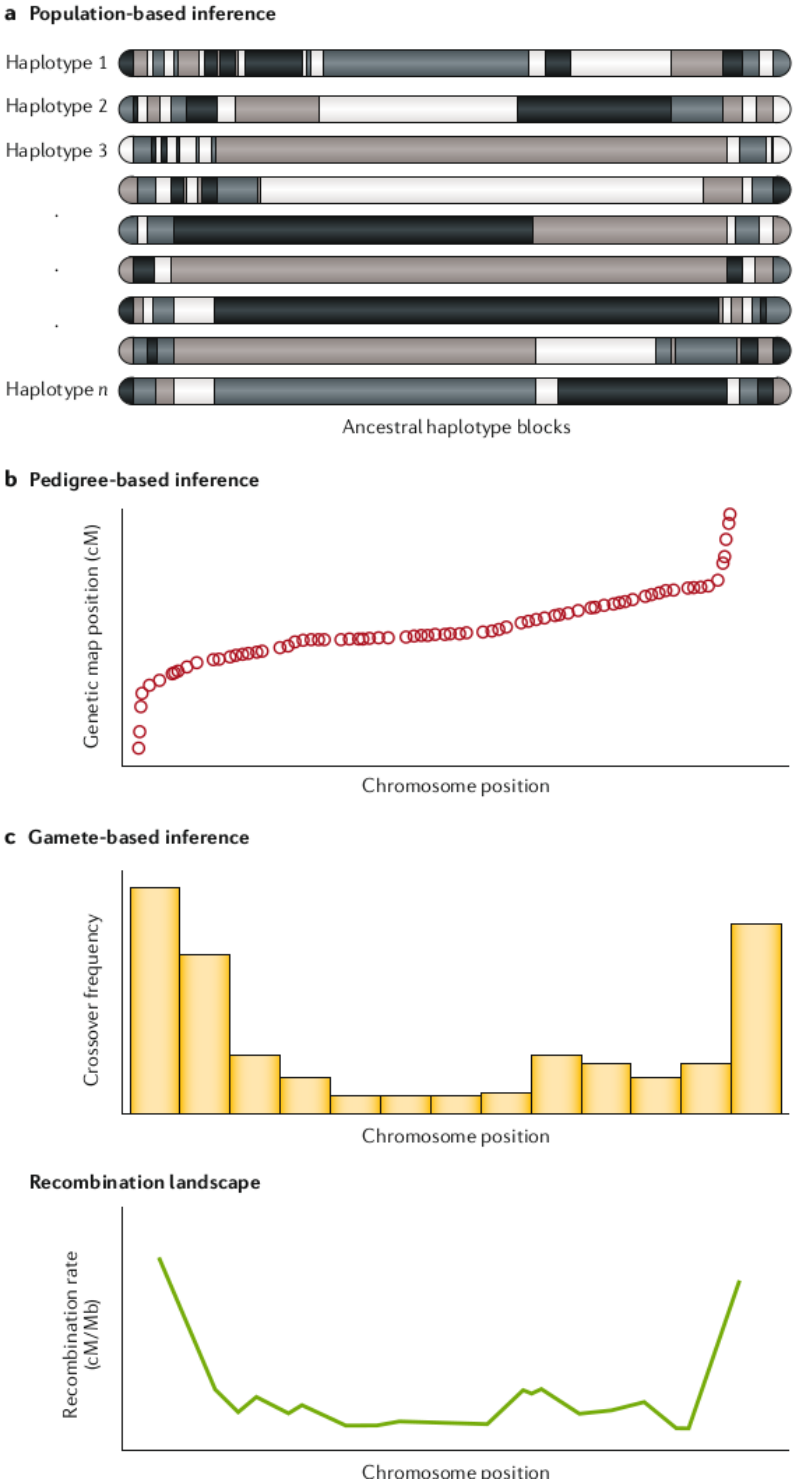


FIGURE 2.2: **From inference to landscape.** The actual result of each inference method and how it translates to the recombination landscape. a | Population-based inference involves direct analysis of haplotype structure along chromosomes. Contemporary haplotypes are composed of ancestral haplotypes (various shades) that arose at different points in the past. The identity and length of ancestral haplotype blocks are a function of the time at which the haplotype arose and recombination. b | Pedigree-based inference involves a comparative representation of genetic distance and physical distance where the local recombination rate is the slope at any given location. c | Gamete-based inference takes the crossover frequency of a given window and translates it into the recombination landscape. cM, centiMorgans. [from (Peñalba and Wolf, 2020)]

2.1.2 An approach to estimate quickly and easily the recombination rate along chromosomes

Thus, we chose to focus on the pedigree-based one, where physical and genetic maps are correlated to infer local recombination rate estimates, based on the Marey Maps (Chakravarti, 1991). This is at the heart of our contribution presented in the next section, since it consists of the one and only input data type we chose to exploit and build our new approach upon (see Figures 2.3, 2.4).

Figure 2.3 illustrates what is a chromosome, genomic data, genetic map (and distance) and physical map (and distance), in addition to the link between them. For more clarity in further details, Figure 2.4 presents the type of data we will be dealing with for the rest of this chapter. It's a simple format: two maps representing the genetic and physical distances, stored as a CSV or TXT data file with a set of markers and their coordinates.

The Marey map approach consists of correlating the physical map with the genetic map representing respectively physical and genetic distances for a set of genetic markers on the same chromosome (Chakravarti, 1991) (see Figure 2.4).

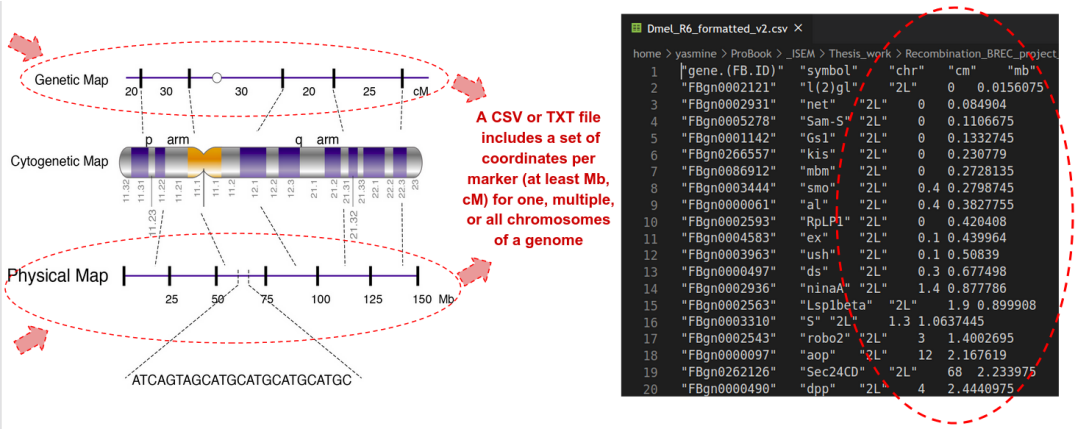


FIGURE 2.3: (Left) Illustration of an ideogram for a chromosome where the cytogenetic map is represented by the colored bands on p arm, the centromere, and the q arm. The corresponding genetic and physical maps pointed out with the red arrows represent the two sets of data of our interest, where the genetic and physical distances are given in centiMorgans (cM) and Mega base-pairs (Mb), respectively. [from https://www.genome.gov/genetics-glossary/Physical-Map. (Right) A sample of the input data file for the chromosome 2 of *D. melanogaster* genome, showing a set of markers (lines) and their coordinates (columns).

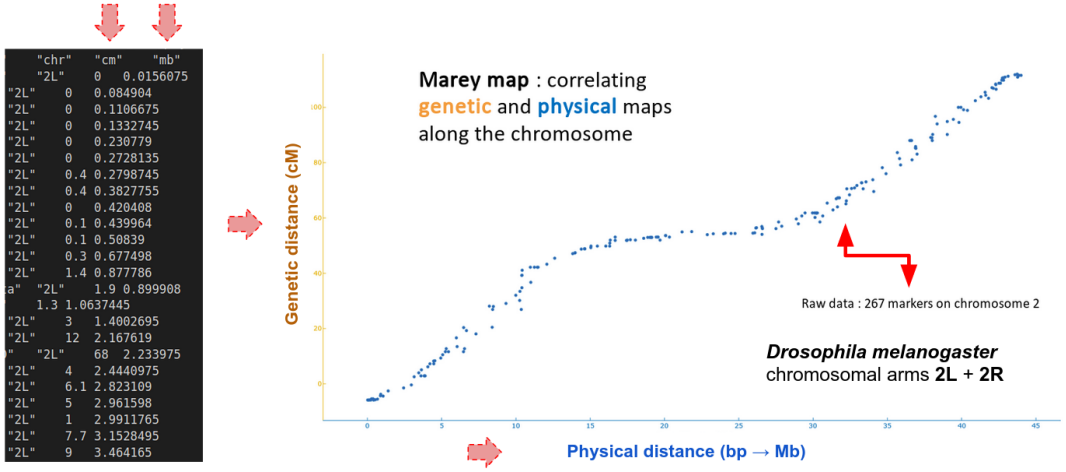


FIGURE 2.4: (Left) A sample of the input data file for the chromosome 2 of *D. melanogaster* genome, serving to build it's Marey Map (Right).

Some Marey map-based tools already exist, two of which are primarily used. The **MareyMap Online** (Rezvoy et al., 2007; Siberchicot et al., 2017) applies to multiple species, which makes it easily exploitable on user-specific data, while provides three regression models for the recombination rate estimates: 3th degree polynomial, Loess, and the cubic spline, as per the user's choice. Since it comes with a Shiny web-based application, it not only easily accessible, but also includes a data cleaning step where the used may select tat data points which appear to be more likely outliers, and proceed to the cleaning step. However, along some specific regions like the chromosome extremities, the recombination rates could not be accurately estimated, as pointed out in Figure 2.5 (e.g. negative values of the recombination rate).

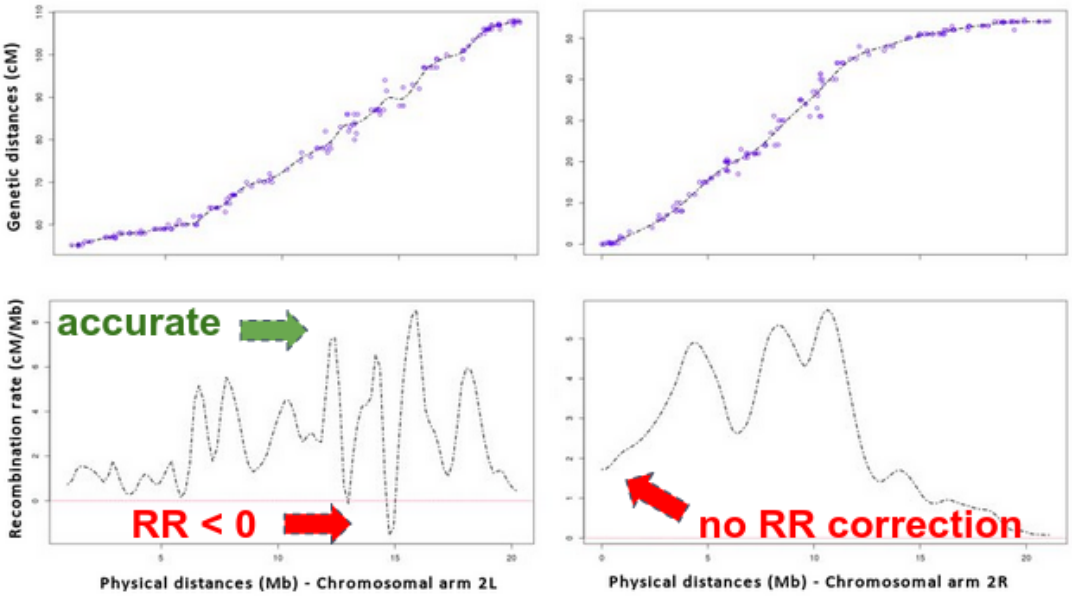


FIGURE 2.5: Screenshot of the MareyMapOnline plots of *D. melanogaster* data: (Right) the arm 2L and (Left) the 2R arm. (top) The Marey maps with the interpolation. (Bottom) the recombination rate estimates (RR). [Adapted from the online version]

Second, the *D. melanogaster* Recombination Rate Calculator (RRC) (Fiston-Lavier et al., 2010) solves the previous issue, by identifying the centromeric and telomeric regions, along which it adjusts the recombination rate estimates, as pointed out on Figure 2.7. However, as indicated by its name, the RRC is *D. melanogaster*-specific, and it applies only the 3rd polynomial regression model for the interpolation, which is broad-scale, and thus less accurate estimates.

With the emerging NGS technologies, accessing whole chromosome sequences has become possible on a wide range of species. Therefore, we may expect an exponential increase in the markers number, requiring more adapted tools to handle such new scopes of data efficiently.

The lack of Fine-scale and/or high density maps like (Comeron, Ratnappan, and Bailin, 2012) It lies mainly in the incorrectly estimated RR on the heterochromatin regions, where the RR is expected to be null or at least very reduces, while this is not the case as shown in Figure 2.6.

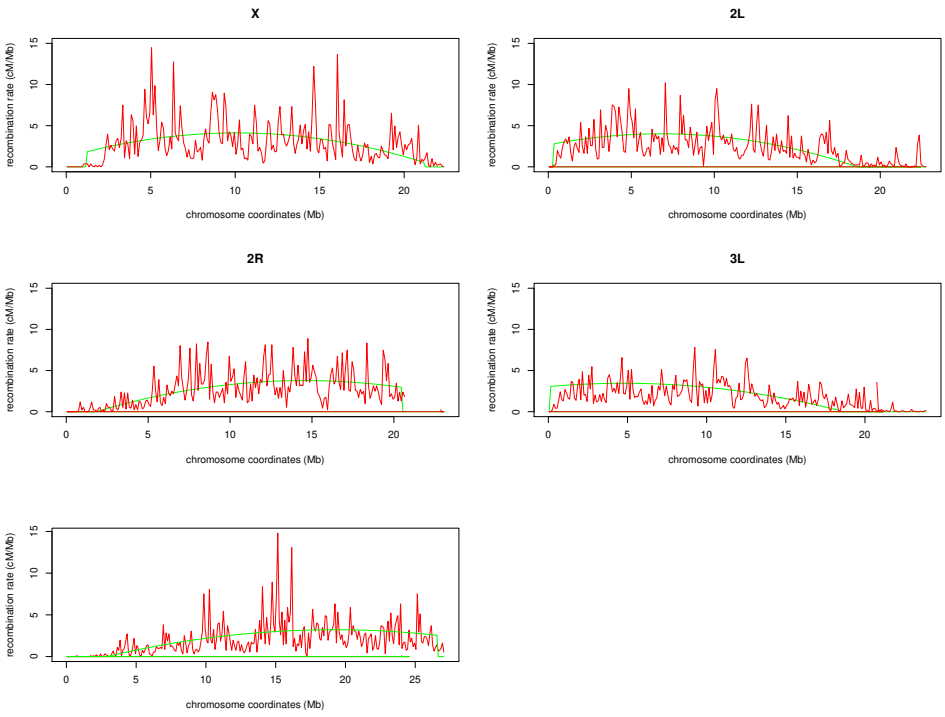


FIGURE 2.6: Comparison of recombination rate estimates between Marey map-based by RRC the (Fiston-Lavier et al., 2010) and population genetics-based by (Comeron, Ratnappan, and Bailin, 2012).

Despite the efficiency of this approach and mostly the availability of physical and genetic maps, generating recombination maps rapidly and for any organism is still challenging. Hence, the increasing need for an automatic, portable, and easy-to-use solution.

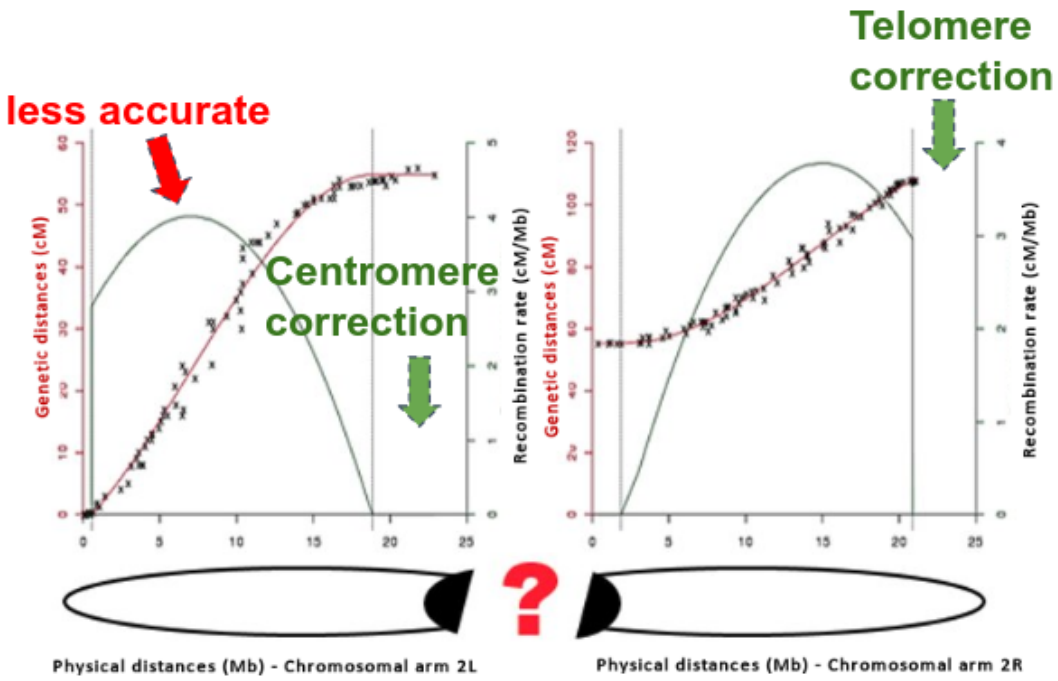


FIGURE 2.7: RRC: Dmel 2L + 2R (the original motivation for BREC development. [Adapted from (Fiston-Lavier et al., 2010)]

Here, we propose a new Marey map-based method as an automated computational solution that aims to, firstly, identify heterochromatin boundaries (HCB) along chromosomes, secondly, estimate local recombination rates, and lastly, adjust recombination rates on chromosome along the chromosomal regions marked by the identified boundaries.

2.2 New Approach: BREC

Different heterochromatin regions exhibit different profiles of recombination rates. Therefore, in order to understand how and why the recombination rate varies, it is vital to break down the chromosome structure into smaller blocks where several genomic features, besides recombination rate, are also known to exhibit different profiles.

Within the context of genome architecture and evolution, introduced in the previous chapter, we will focus on the two first investigated aspects, which are the variation of meiotic recombination rates, and the identification of boundaries between euchromatin and heterochromatin regions on the chromosome scale.

BREC Workflow BREC (Mansour, Chateau, and Fiston-Lavier, 2021) is designed following the workflow represented in Figure 2.8. To ensure that the broadest range of species could be analyzed by our tool, we designed a pipeline that adapts behavior with respect to input data. Each step of the workflow relies mostly on statistical analysis, adaptive algorithms, and decision proposals led by empirical observation.

The workflow starts with a pre-processing module (called "Step 0") aiming to prepare the data prior to the analysis. Then, it follows six main steps: (1) estimate Marey Map-based local recombination rates, (2) identify chromosome type, (3) prepare the HCB identification, (4) identify the centromeric boundaries, (5) identify the telomeric boundaries, and (6) extrapolate the local recombination rate map and generate an interactive plot containing all BREC outputs (see Figure 2.8). Each step is detailed hereafter and summarised in Figure 2.9.

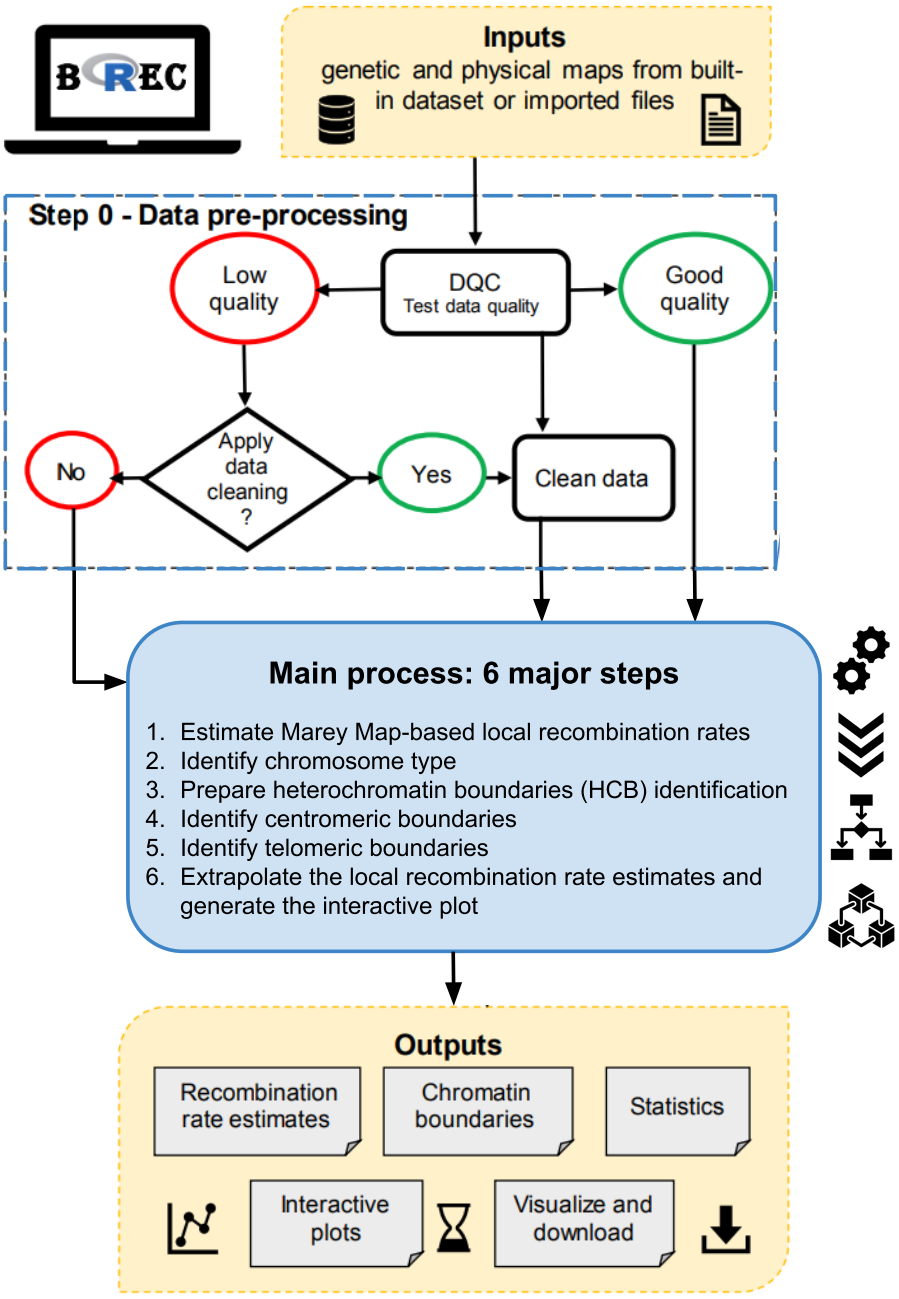


FIGURE 2.8: **BREC workflow.** This figure provides an overview of the tool design explaining how the different modules are linked together and how BREC functionalities are implemented. The top-to-bottom diagram starts with the required input data, how they are pre-processed (Step 0) and exploited (Main process: 6 major steps), then, what outputs are expected to be returned and in which format. A more detailed version is included in the Figure 2.9, where a zoom-in on the main process is clarified for each of the six steps.

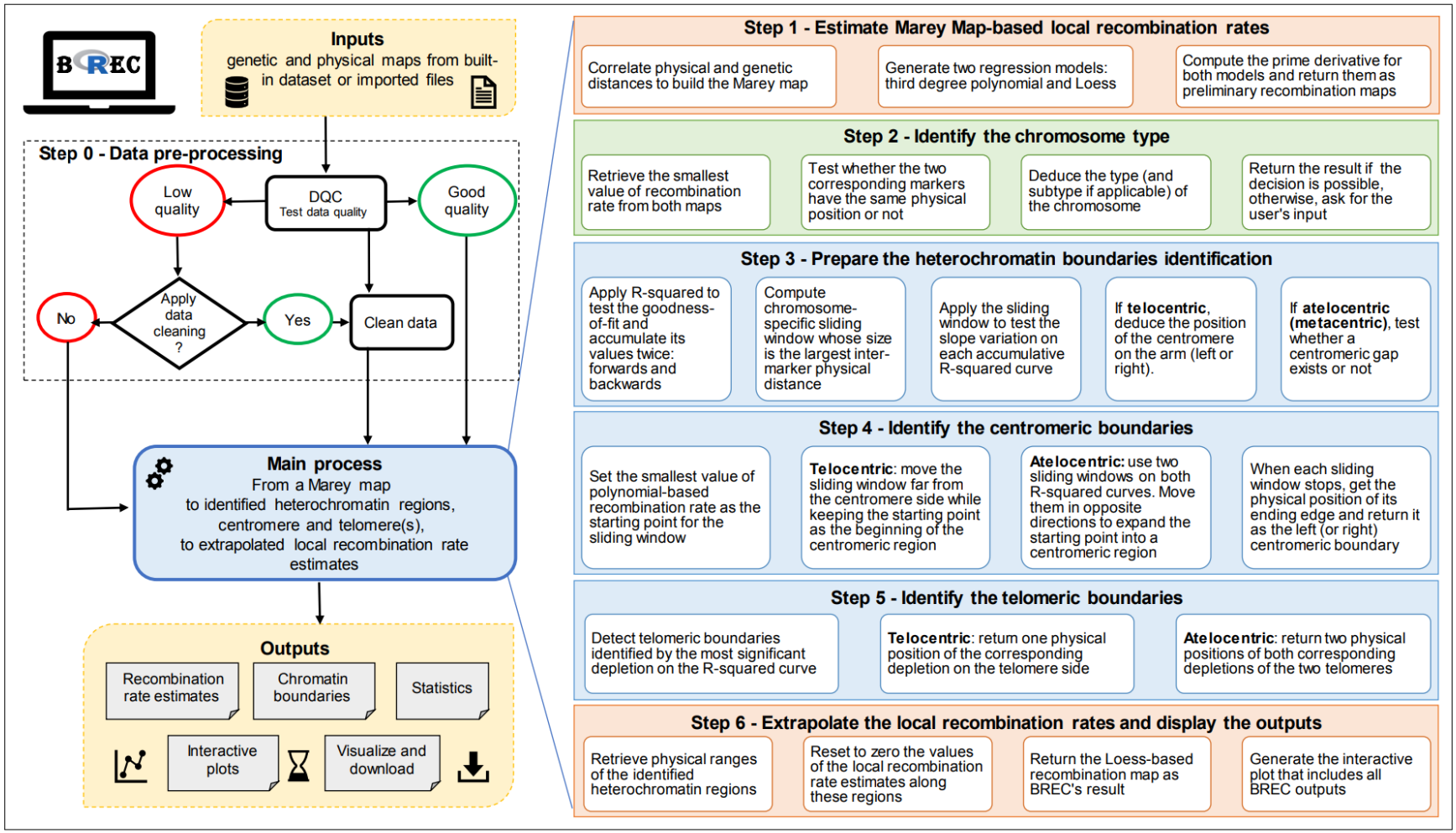


FIGURE 2.9: **BREC workflow.** As a more detailed version of Figure 2.8, this figure provides an overview of the tool design explaining how the different modules are linked together and how BREC functionalities are implemented. The left part represents the top-to-bottom diagram, starting with the required input data, how they are pre-processed (Step 0) and exploited (Main process), then, what outputs are expected to be returned and in which format. The right part of the figure, representing a zoom-in on BREC's main module (estimating recombination rates, identifying chromosome type, identifying HCB, extrapolating the recombination map and generating the interactive plot), clarifies each step following a more detailed scheme.

2.2.1 Step 0 - Apply data pre-processing

Since we have noticed that BREC estimates are sensitive to the quality of input data, we propose a pre-processing step to assess data quality and suggest an optional data cleaning for outliers. As such, we could ensure proper functioning during further steps.

Data quality control

The quality of input data is tested regarding two criteria: (1) the density of markers and (2) the homogeneity of their distribution on the physical map along a given chromosome. First, the mean density, defined as the number of markers per physical map length, is computed. This value is compared with the minimum required threshold of 2 markers/Mb. Based on the displayed results, the user gets to decide if data cleaning is required or not. The threshold of 2 markers/Mb is selected based on a simulation process that allowed to test BREC results while decreasing markers density until the observed HCB estimates seemed to be no longer exploitable (see Section Validation process: Simulated data for quality control testing). Second, the distribution of input data is tested *via* a comparison with a simulated uniform distribution of identical markers density and physical map length. This comparison is applied using Pearson's $Chi - squared$ test (Agresti, 2007), which allows examining how close the observed distribution (input data) is to the expected one (simulated data).

Data cleaning

The cleaning step aims to reduce the disruptive impact of noisy data, such as outliers, in order to provide a more accurate recombination rate and heterochromatin boundary results. If the input data fails to pass the Data Quality Control (DQC) test, the user has the option to apply or not a cleaning process. This process consists of identifying the extreme outliers and eliminating them upon the user's confirmation. Outliers are detected using the distribution statistics of the genetic map (see Figure 2.10). More precisely, inter-marker distances (separating each two consecutive points) are computed along the genetic map. Using a boxplot, distribution statistics (quartiles, mean, median) are applied on these inter-marker distances to identify outliers, which are chosen as the 5% of the data points with a greater genetic distance than the maximum extreme value, and should be discarded. Thus, the cleaning targets markers for which the genetic distance is quite larger than most of the rest. After the first cleaning iteration, DQC is applied again to assess the new density and distribution. The user can also choose to bypass the cleaning step, but BREC's behavior is no longer guaranteed in such cases.

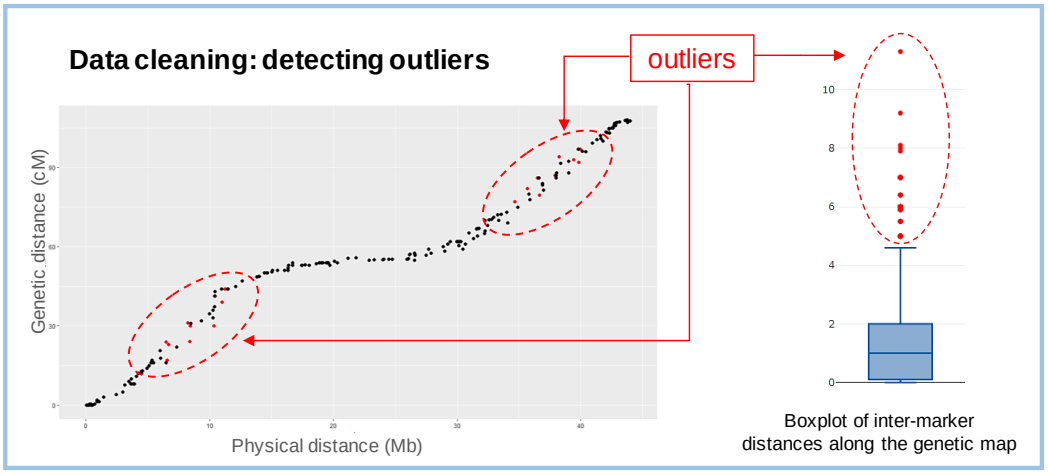


FIGURE 2.10: **The data cleaning process implemented within BREC**. Inter-marker distances (*i.e.* genetic distances between each two consecutive points along the genetic map) are represented using a boxplot in order to identify outliers and give the user the option to remove them. Here is an example showing raw data of a simulated chromosome (left) with the specific markers detected as outliers (red dots circled with red dashed ovals) and the corresponding genetic distances (also in red) on the boxplot (right).

2.2.2 Step 1 - Estimate Marey Map-based local recombination rates

Once the data are cleaned, the recombination rate can be estimated based on the Marey map (Chakravarti, 1991) approach by: (1) correlating genetic and physical maps, (2) generating two regression models -third degree polynomial and Loess- that better fits these data, (3) computing the prime derivative for both models which will represent preliminary recombination maps for the chromosome. The primary purpose of interpolation here is to provide local recombination rate estimates for any given physical position, instead of only the ones corresponding to available markers.

At this point, both recombination maps are used to identify the chromosome type as well as the approximate position of centromeric and telomeric regions. Nevertheless, as a final output, BREC will return only the Loess-based adjusted map for recombination rates since it provides finer local estimates than the polynomial-based map.

2.2.3 Step 2 - Identify chromosome type

BREC provides a function to identify the type of a given chromosome according to the position of its centromere. This function is based on the physical position of the smallest value of recombination rate estimates, which primarily indicates where the centromeric region is more likely to be located. Our experimentation allowed to come up with the following scheme (see Figure 2.11). Two main types are identified: telocentric and atelocentric (Levan, Fredga, and Sandberg, 1964). Atelocentric type could be either metacentric (centromere located approximately in the center with almost two equal arms) or not metacentric (centromere located between the center and

one of the telomeres). The latter includes the two most known subtypes, submetacentric and acrocentric (recently considered types rather than subtypes). It is tricky for BREC to distinguish between submetacentric and acrocentric chromosomes correctly. Their centromeres' position varies slightly, and capturing this variation (based on the smallest value of recombination rate on both maps -polynomial and Loess-) could not be achieved yet. Therefore, we chose to provide this result only if the implemented process allowed to identify the subtype automatically. Otherwise, the user gets the statistics on the chromosome's data and is invited to decide according to further *a priori* knowledge. The two subtypes (metacentric and not metacentric) are distinguished following intuitive reasoning inspired by their definition found in the literature. First, BREC identifies whether the chromosome is an arm (telocentric) or not (atelocentric). Then, it tests if the physical position of the smallest value of the estimated recombination rate is located between 40% to 60% interval. In this case, the subtype is displayed as *metacentric*. Otherwise, it is displayed as *not metacentric*. The recombination rate is estimated using the Loess model ("LOcal regrESSion") (Cleveland and Devlin, 1988; Cleveland and Loader, 1996).

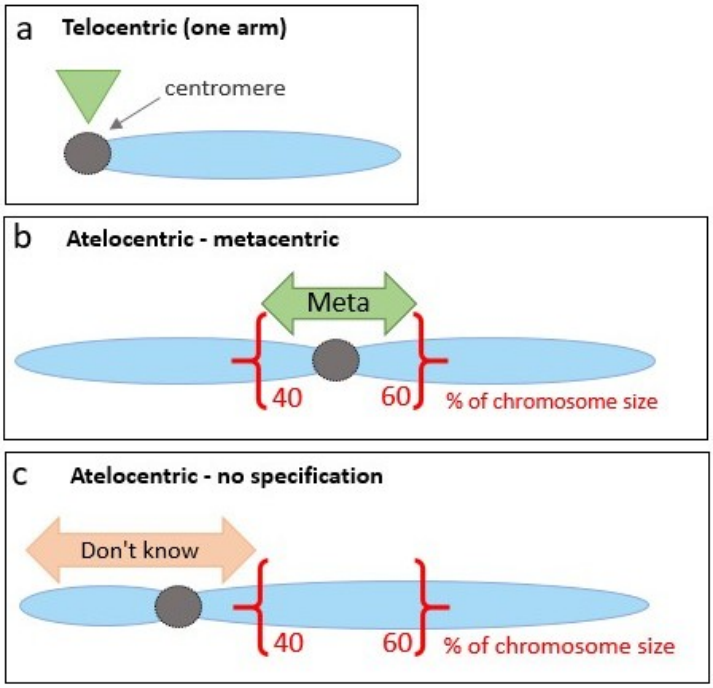


FIGURE 2.11: **A schematic description of the chromosome type identification process implemented within BREC.** (a) Telocentric chromosome type is when the centromere (the grey colored circle) is located on one of the chromosomal arm extremities (indicated with the green upside down triangle). (b) Atelocentric chromosome type -confirmed as metacentric- is when the centromere is located approximately on the middle of the chromosome, here showed within the physical positions 40% and 60% of the chromosome's size (delimited by the red brackets and indicated with the tag "Meta"). (c) Atelocentric chromosome type -with no specification- is when the centromere is located either inside the first arm (between the beginning of the chromosome and 40% of its size), or inside the second arm (between 60% and the end, indicated with the tag "Don't know").

2.2.4 Step 3 - Prepare the HCB identification

The HCB identification is a purely statistical approach relying on the coefficient of determination R^2, which measures how good the generated regression model fits the input data (Zhang, 2017). We chose this approach because the Marey map usually exhibits a lower quality of markers (density and distribution) on the heterochromatin regions. Thus, we aim to capture this transition from high to low quality regions (or *vice versa*) as it reflects the transition from euchromatin to heterochromatin regions (or *vice versa*). The coefficient R^2 is defined as the cumulative sum of squares of differences between the interpolation and observed data. R^2 values are accumulated along the chromosome. In order to eliminate the biased effect of accumulation, R^2 is computed twice: $R^2 - forward$ starts the accumulation from the beginning of the chromosome to provide the left centromeric and left telomeric boundaries. In contrast, $R^2 - backwards$ starts from the end of the chromosome, providing the right centromeric and right telomeric boundaries. These R^2 values were calculated using the `rsq` package in R. To compute R^2 cumulative vectors, `rsq` function is applied on the polynomial regression model. In fact, there is no such function for non-linear regression models like the Loess because, in such models, high R^2 does not always indicate a good fit. A sliding window is defined and applied on the R^2 vectors to precisely analyze their variations (see details in the next step). In the case of a telocentric chromosome, the position of the centromere is then deduced as the left or the right side of the arm, while in the case of an atelocentric chromosome, the existence of a centromeric gap is investigated.

2.2.5 Step 4 - Identify centromeric boundaries

Since the centromeric region is known to present reduced recombination rates, the starting point for detecting its boundaries is the physical position corresponding to the smallest polynomial-based recombination rate value. A sliding window is then applied to expand the starting point into a region based on R^2 variations in two opposite directions. The sliding window's size is automatically computed for each chromosome as the largest value of ranges between each two consecutive positions on the physical map (indicated as i and $i+1$ in Equation 2.1). After making sure the sliding window includes at least two data points, the mean of local growth rates inside the current window is computed and tested compared to zero. If it is positive (resp. negative) on the forward (resp. backward) R^2 curve, the value corresponding to the window's ending edge is returned as the left (resp. right) boundary. Else, the window moves by a step value equal to its size.

$$sliding\_window\_size(chromosome) = max\{|physPos_{i+1} - physPos_i| : 1 \leq i \leq n-1\} \tag{2.1}$$

There are some cases where chromosome data present a centromeric gap. Such a lack of data produces biased centromeric boundaries. To overcome this issue, chromosomes with a centromeric gap are handled with a slightly different approach. After comparing the mean of local growth rates regarding to zero, accumulated slopes of all data points within the sliding window are computed, adding one more point at

a time. If the mean of accumulated slopes keeps the same variation direction as the mean of growth rates, the centromeric boundary is set as the window's ending edge. Else, the window slides by the same step value as before (equal to its size). The difference between the two chromosome types is that only one sliding window is used for the telocentric case, its starting point is the centromeric side, and it moves away from it. As for the atelocentric case, two sliding windows are used (one on each R^2 curve), their starting point is the same, and they move in opposite directions to expand the centromere into a region.

2.2.6 Step 5 - Identify telomeric boundaries

Since telomeres are considered heterochromatin regions as well, they also tend to exhibit low fitness between the regression model and the data points. More specifically, the accumulated R^2 curve tends to present a significant depletion around telomeres. Therefore, a telomeric boundary is defined here as the physical position of the most significant depletion corresponding to the smallest value of the R^2 curve. As such, in the telocentric case, only one R^2 curve is used. It gives one boundary of the telomeric region (the other boundary is defined by the beginning of the left telomere or the end of the right telomere). Whilst in the atelocentric case, where the are two telomeres, the depletion on $R^2 - forward$ detects the end of the left telomeric region, and the depletion on $R^2 - backwards$ detects the beginning of the right telomeric region. The other two boundaries (the beginning of the left telomere and the end of the right telomere) are defined to be, respectively, the same values of the two markers with the smallest and the largest physical position available within the input data of the chromosome of interest.

2.2.7 Step 6 - Extrapolate the local recombination rate estimates and generate interactive plot

The extrapolation of recombination rate estimates at the identified centromeric and telomeric regions automatically performs an adjustment by resetting the initial biased values to zero along these heterochromatin ranges. Finally, all of the above BREC outputs are combined to generate one interactive plot to display for visualization and download (see details in Section BREC results: Easy, fast and accessible tool *via* an R-package and a Shiny app).

It is important to emphasize that throughout the whole main process module, only step 1 "Estimating Marey map-based local recombination rates" comes from previous methods (Chakravarti, 1991; Rezvoy et al., 2007). Otherwise, each of the steps 2-6 are fully developed (designed and implemented) within BREC and represent a new contribution, in addition to step zero "Data pre-processing", as mentioned above.

2.3 Validation process

2.3.1 Validation data

The only input dataset to provide for BREC is genetic and physical maps for one or several chromosomes. A simple CSV file with at least two columns for both maps is valid. If the dataset is for more than one chromosome or the whole genome, a third column, with the chromosome identifier, is required. (see Figure 2.4).

Our results have been validated using Release 5 of the fruit fly *D. melanogaster* (Hoskins et al., 2007; Hoskins et al., 2015) genome as well as the domesticated tomato *Solanum lycopersicum* genome (version SL3.0).

We also tested BREC using other datasets of different species: house mouse (*Mus musculus castaneus*, MGI) chromosome 4 (Cox et al., 2009), roundworm (*Caenorhabditis elegans*, ws170) chromosome 3 (Hillier et al., 2008), zebrafish (*Danio rerio*, Zv6) chromosome 1 (Freeman et al., 2007), respectively (see Figure 2.12), as samples from the multi-genome dataset included within BREC (see further details on the full built-in dataset in Section Validation process: Description of main components of the Shiny app).

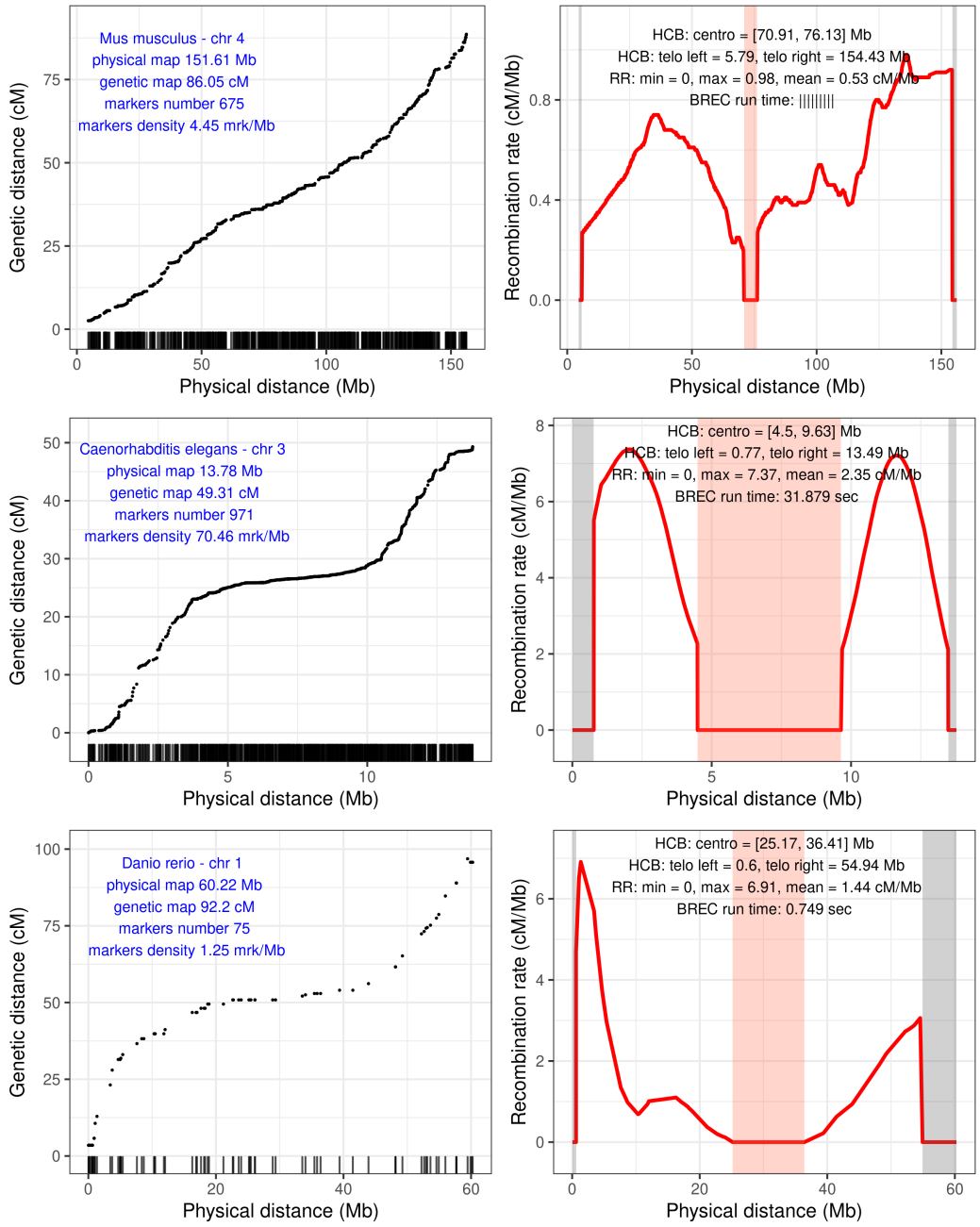


FIGURE 2.12: **BREC results on different species: from top to bottom are *M. musculus* (house mouse) chromosome 4, *C. elegans* (roundworm) chromosome 3, *D. rereo* (zebrafish) chromosome 1, respectively.** For each species, two plots are shown: on the left is the chromosome's genetic markers (black points), their distribution along the physical map (rug on the x-axis), and reported genomic features (label in blue). On the right is BREC results: HCB for centromeric (red highlight) and telomeric (grey highlight) regions, (RR) local recombination rate estimates (red line), and the running time of BREC's algorithms to get these results (loading data and plotting are excluded).

Fruit fly genome *D.melanogaster*

Physical and genetic maps are available for download from the FlyBase website (http://flybase.org/; Release 5) (Thurmond et al., 2019). This genome is represented here with five chromosomal arms: 2L, 2R, 3L, 3R, and X (see Table 2.1), for a total of 618 markers, 114.59Mb of physical map and 249.5cM of genetic map. This dataset is manually curated and is already clean from outliers. Therefore, the cleaning step offered within BREC was skipped.

Chromosomal arms	X	2L	2R	3L	3R	Genome
Markers number	165	110	101	82	160	618
Markers density (marker/Mb)	7.78	4.81	4.78	3.56	5.80	5.39
Physical map length (Mb)	21.22	22.88	21.12	21.81	27.57	114.59
Genetic map length (cM)	65.8	54.8	52.5	45.9	57.5	276.5
BREC run time (sec)	1.278	0.949	0.821	0.916	1.379	5.343

TABLE 2.1: **Genomic features and BREC running time for the *D. melanogaster* Release 5 genome**. The first five columns represent chromosomal arms. Rows represent the genome features as follows: (1) the names of chromosomal arms X, 2L, 2R, 3L, and 3R; (2) the markers number included in the study; (3) the markers density (in markers/Mb); (4) the physical map length (in Mb); (5) the genetic map length (in cM); and (6) the elapsed time when running BREC (in seconds). The last column summarises the same features for the whole genome.

Tomato genome *S. lycopersicum*

Domesticated tomato with 12 chromosomes has a genome size of approximately 900Mb. Based on the latest physical and genetic maps reported by the Tomato Genome Consortium (Sato et al., 2012), we present both maps content (markers number, markers density, physical map length, and genetic map length) for each chromosome in Table 2.2. For a total of 1957 markers, 752.47Mb of physical map and 1434.49cM of genetic map along the whole genome.

Chromosome	1	2	3	4	5	6	7	8	9	10	11	12	Genome
Markers number	232	176	184	160	150	151	145	144	171	148	142	154	1957
Markers density (marker/Mb)	2.58	3.66	2.84	2.55	2.32	3.34	2.22	2.29	2.54	2.32	2.68	2.36	2.64
Physical map length (Mb)	89.85	48.10	64.77	62.79	64.52	45.20	65.18	62.87	67.37	63.66	52.98	65.18	752.47
Genetic map length (cM)	150.72	154.58	134.52	122.64	137.91	106.63	92.48	106.63	108.90	88.92	119.99	110.72	1434.49
BREC run time (sec)	2.164	1.391	1.434	1.295	1.098	1.197	1.102	1.047	1.357	1.095	1.081	1.221	15.479

TABLE 2.2: **Genomic features and BREC running time for the *S. lycopersicum* .** The first twelve columns represent chromosomes. Rows represent the genome features as follows: (1) the identifiers of chromosomes 1 to 12; (2) the markers number included in the study; (3) the markers density (in markers/Mb); (4) the physical map length (in Mb); (5) the genetic map length (in cM); and (6) the elapsed time when running BREC (in seconds). The last column summarises the same features for the whole genome.

2.3.2 Simulated data for quality control testing

We call *data scenarios*, the layout in which the data markers are arranged along the physical map. For experimentally testing the limits of BREC, various *data scenarios* have been specifically designed based on *D. melanogaster* chromosomal arms (see Figure 2.13).

Simulation Data scenarios	Centromeric gap exists (lack of markers)		Markers density on heterochromatic regions
	Yes	No	
Uniform distribution	(a)	(b)	Unvariant
Non uniform distribution	(c)	(d)	High
	(e)	(f)	Low

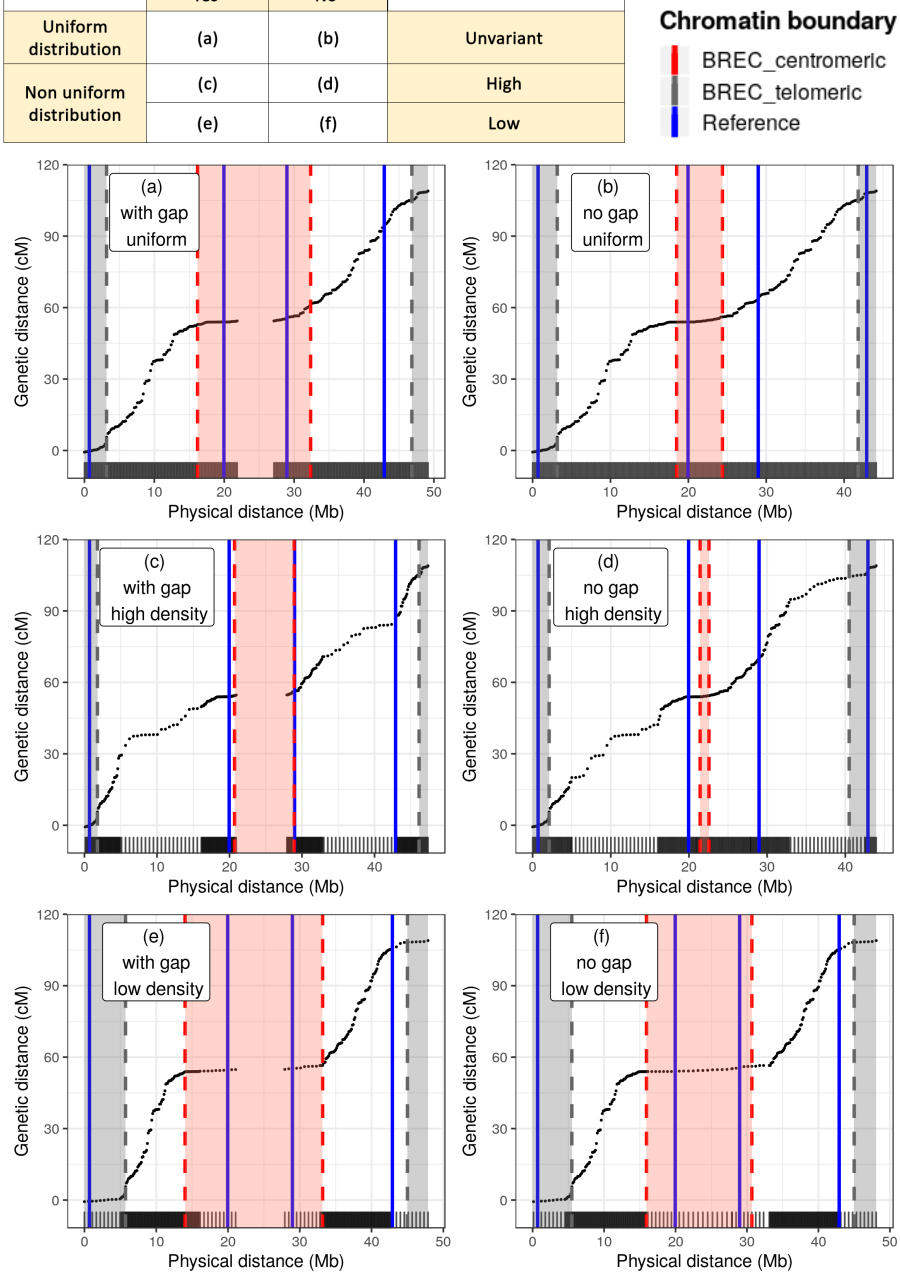


FIGURE 2.13: **Distribution simulations.** BREC results on the simulated chromosomes with different scenarios of markers distribution around heterochromatin regions, as presented in the table (top) . Plots (right after) are presenting the corresponding results for each simulation scenario. On the left, (a, c, e) show the cases with the existence of centromeric gap while the ones on the right (b, d, f) show the cases with no centromeric gap. From top to bottom, cases (a) and (b) show a uniform distributions while (c) to (f) are for non uniform distributions. Cases (c) and (d) show a higher density of markers around heterochromatin regions while cases (e) and (f) show a lower density on the same regions. Black dots represent genetic markers. Vertical lines represent HCB for BREC centromeres (in red dashed line), for BREC telomeres (in grey dashed line) and for the reference (in solid blue line). The heterochromatin regions identified by BREC are highlighted for the centromere (in red) and the telomere (in grey). The rug plot, added on the x axis, shows more clearly the variation in markers density as well as the existence or not of the centromeric gap.

In an attempt to investigate how the markers' density varies within and between the five chromosomal arms of *D. melanogaster* Release 5 genome, the density has been analyzed in two ways: locally (with 1Mb-bins) and globally (on the whole chromosome). Figure 2.14 shows the results of this investigation, where each little box indicates how many markers are present within the corresponding region of size 1Mb on the physical map. The mean value represents the global density. It is also shown in Table 2.1 where the values are slightly different. This is due to computing the marker's density in two different ways with respect to the analysis. Table 2.1, presenting the genomic features of the validation dataset, shows markers density in Column 3, which is simply the result of the division of markers number (in column 2) by the physical map length (in Column 4). For example, in the case of chromosomal arm X, this gives $165/21.22 = 7.78 markers/Mb$. On the other hand, Figure 2.14, aimed for analyzing the variation of local markers density, displays the mean of of all 1Mb-bins densities, which is calculated as the sum of local densities divided by the number of bins, and this gives $165/22 = 7.5 markers/Mb$.

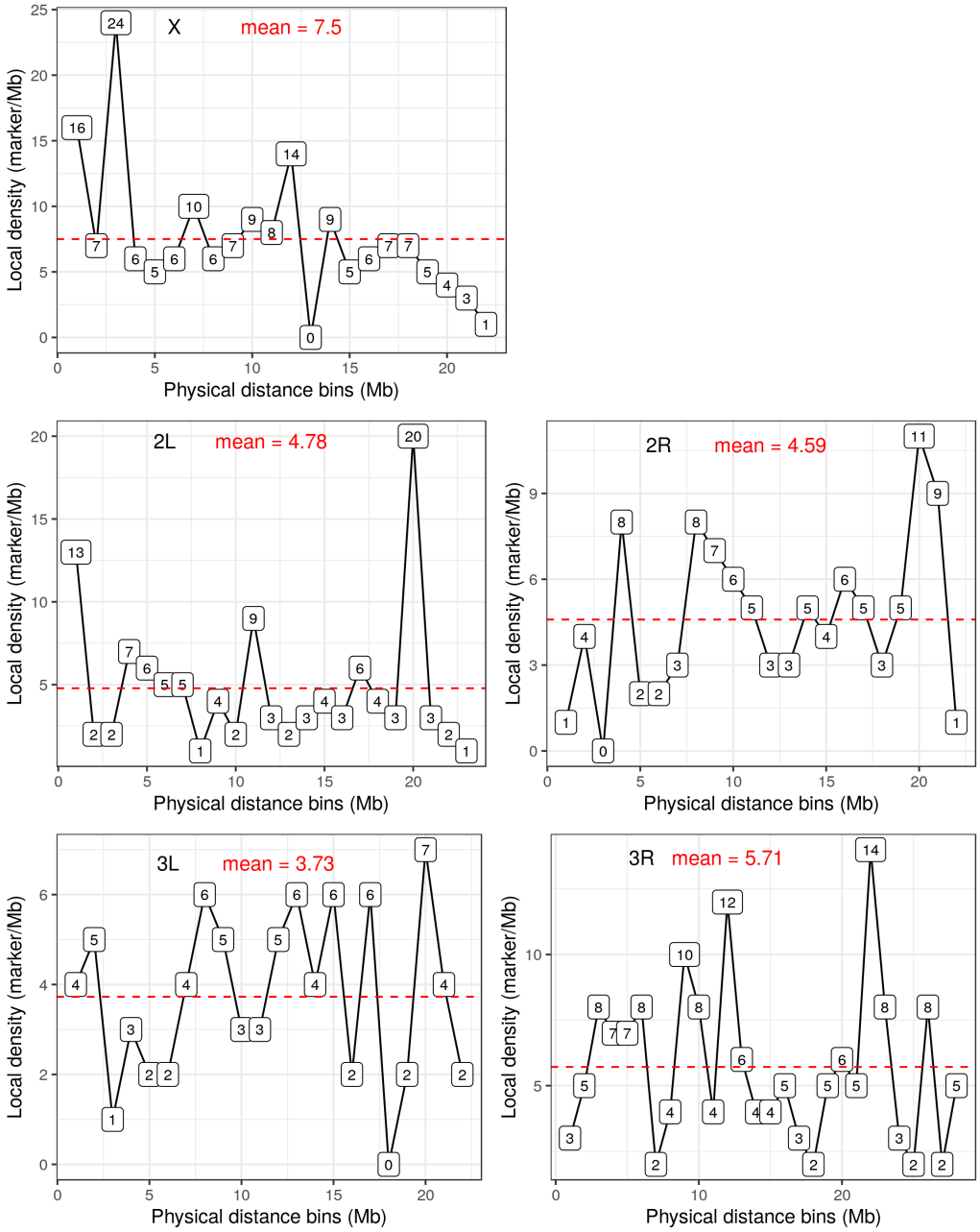


FIGURE 2.14: **Variations of markers local density per 1-Mb bins along *D. melanogaster* Release 5 chromosomal arms.** The red dashed line indicates the mean and represents the global density. Each bin indicates the number of markers it contains. Local density values are represented within the little boxes.

The exact same analysis has been conducted on the tomato genome *S. lycopersicum* where the only difference lies in using 5-Mb instead of 1-Mb bins, due to the larger size of its chromosomes (see Figure 2.15).

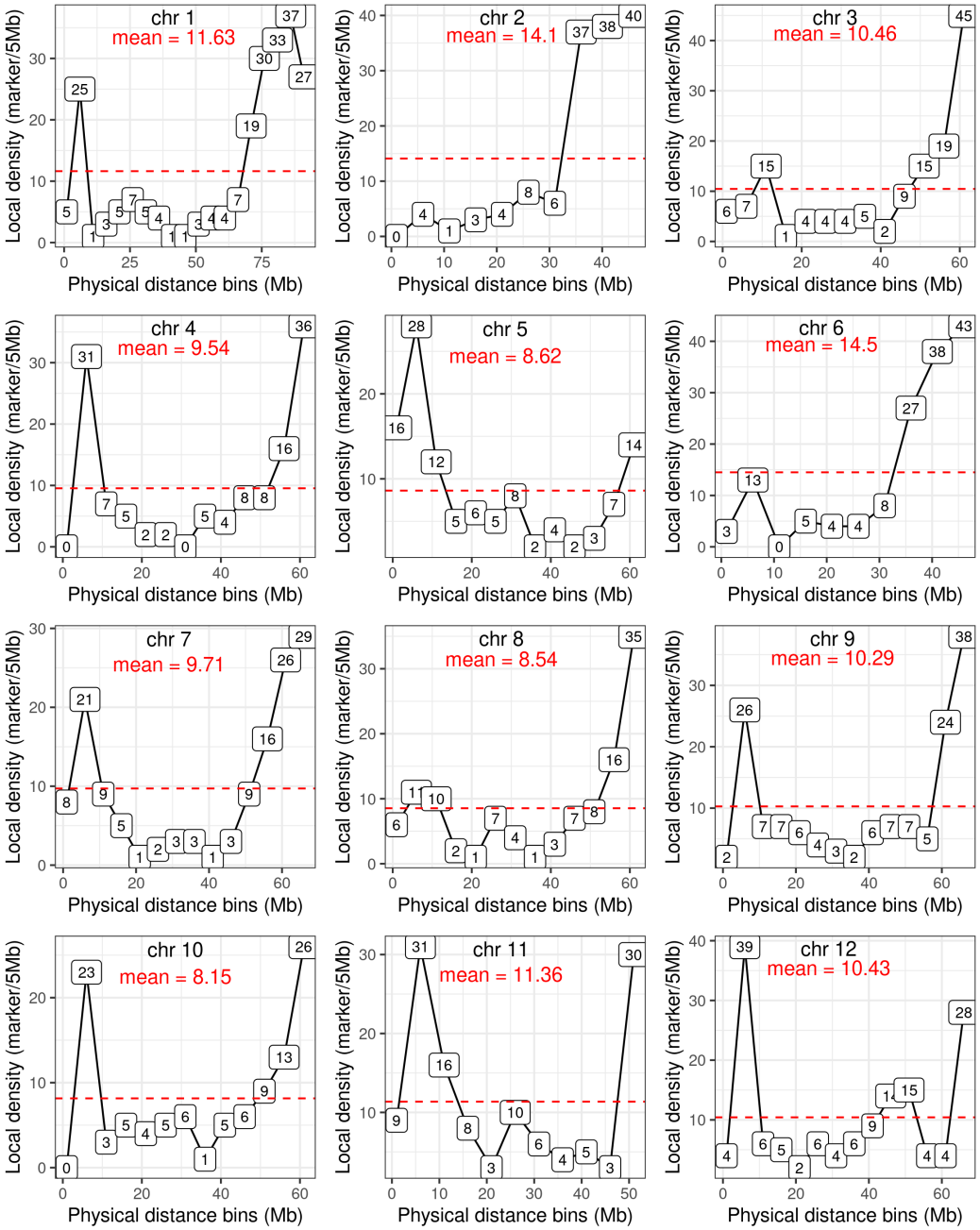


FIGURE 2.15: **Variations of markers local density per 5-Mb bins along the tomato genome *S. lycopersicum* 12 chromosomes.** The red dashed line indicates the mean and represents the global density. Each bin indicates the number of markers it contains. Local density values are represented within the little boxes.

2.3.3 Validation metrics

The measure we used to evaluate the resolution of BREC's HCB is called $shift$ hereafter. It is defined as the difference between the observed heterochromatin boundary ($observed\_HCB$) and the expected one ($expected\_HCB$) in terms of physical distance (in Mb)(see Equation 2.2).

$$shift = |observed\_HCB - expected\_HCB| \tag{2.2}$$

The $shift$ value is computed for each heterochromatin boundary independently. Therefore, we observe only two boundaries on a telocentric chromosome (one centromeric and one telomeric). In comparison, we observe four boundaries in the case of an atelocentric chromosome (two centromeric giving the centromeric region and two telomeric giving each of the two telomeric regions).

The $shift$ measure was introduced not only to validate BREC's results with the reference equivalents but also to empirically calibrate the DQC module, where we are mostly interested in the variation of its value as per variations of the quality of input data.

2.3.4 Implementation and Analysis

The entire BREC project was developed using the R programming language (version 3.6.3 / 2020-02-29) and the RStudio environment (version 1.2.5033) (R Core Team, 2018). The graphical user interface is build using the shiny and shinydashboard packages (RStudio, Inc, 2014). The web-based interactive plots are generated by the plotly package. Data simulations, result analysis, reproducible reports, and data visualizations are implemented using a large set of packages such as tidyverse, dplyr, R markdown, Sweave and knitr among others. The complete list of software resources used is available on the online version of the BREC package accessible at `https://github.com/GenomeStructureOrganization/BREC`.

From inside an R environment, the BREC package can be downloaded and installed using the command in the code chunk in Figure 2.16. In case of installation issues, further documentation is available online on the ReadMe page of the GitHub repository. If all runs correctly, the BREC Shiny application will be launched on your default internet browser.

```
# Install devtools and shiny from CRAN
install.packages("devtools", "shiny")
# Load installed libraries
library(devtools, shiny)


# Download and install the BREC package from the GitHub repository
install_github("ymansour21/BREC")
# Load Brec and shiny
library(Brec)
library(shiny)
# Launch Brec graphical interface in your default internet browser
runApp("shinyApp/Brec_dashboard.R", launch.browser = TRUE)
```

FIGURE 2.16: **Download, install and launch BREC.** Code chunk showing the R commands allowing to download, install and run the BREC Shiny application. The entire R package is available with open access on the indicated GitHub repository.

All BREC experiments have been carried out using a personal computer with the following specs:

- Processor: Intel® Core™ i7-7820HQ CPU @ 2.90GHz x 8
- Memory: 32Mo
- Hard disc: 512Go SSD
- Graphics: NV117 / Mesa Intel® HD Graphics 630 (KBL GT2)
- Operating system: 64-bit Ubuntu 20.04 LTS

2.3.5 Description of main components of the Shiny app

Build-in dataset

Users can either run BREC on a dataset of 44 genomes, mainly imported from (Corbett-Detig, Hartl, and Sackton, 2015), enriched with two mosquito genomes from (Dudchenko et al., 2017) and updated with *D. melanogaster* Release 6 from FlyBase (Thurmond et al., 2019) (see Tables 2.3), already available within the package, or, load new genomes data according to their own interest.

User-specific genomic data should be provided as inputs within at least a 3-column CSV file format, including for each marker: chromosome identifier, genetic distance, and physical distance, respectively. On the other hand, outputs from BREC running results are represented *via* interactive plots.

Species	Common Name	Taxonomy
Aedes aegypti	Yellow fever mosquito	Animal Invertebrate
Anopheles gambiae	African malaria mosquito	
Apis mellifera scutellata	Honeybee	
Bombyx mandarina	Silkworm	
Caenorhabditis briggsae	Roundworm	
Caenorhabditis elegans	Roundworm	
Culex pipiens	Common house mosquito	
Drosophila melanogaster R5	Fruit fly	
Drosophila melanogaster R6	Fruit fly	
Drosophila pseudoobscura	Fruit fly	
Heliconius melpomene melpomene	Postman butterfly	
Bos taurus	Cow	Animal Vertebrate
Canis lupus	Wolf	
Cynoglossus semilaevis	Tongue sole	
Danio rerio	Zebrafish	
Equus ferus przewalskii	Prewalksii's horse	
Ficedula albicollis	Collared flycatcher	
Gallus gallus	Chicken	
Gasterosteus aculeatus	Stickleback	
Homo sapiens	Human	
Lepisosteus oculatus	Spotted gar	
Macaca mulatta	Rhesus macaque	
Meleagris gallopavo	Turkey	
Mus musculus castaneus	House mouse	
Oryzias latipes	Medaka	
Ovis canadensis	Bighorn sheep	
Papio anubis	Olive baboon	
Sus scrofa	Wild boar	
Citrus reticulata	Mandarin Orange	Plant Woody
Gossypium raimondii	New world cotton	
Populus trichocarpa	Black cottonwood	
Prunus davidiana	David's peach	
Arabidopsis thaliana	Thale cress	Plant Herbaceous
Brachypodium distachyon	Purple false brome	
Capsella rubella	Pink Shepherd's Purse	
Citrullus lanatus lanatus	Watermellon	
Cucumis sativus var. hardwickii	Cucumber	
Glycine soja	Wild soybean	
Medicago truncatula	Barrel medic	
Oryza rufipogon	Wild rice	
Setaria italica	Foxtail millet	
Sorghum bicolor subsp. verticilliflorum	Wild Sudan grass	
Solanum lycopersicum	Domesticated tomato	
Zea mays ssp parviglumis	Teosinte	

TABLE 2.3: **BREC's built-in dataset of genomic data.** The available genetic and physical maps for 44 species from (Corbett-Detig, Hartl, and Sackton, 2015), enriched with two recently assembled mosquito genomes: *Cx. pipiens* and *Ae. aegypti* from (Dudchenko et al., 2017), domesticated tomato *S. lycopersicum* from (Sato et al., 2012), and *D. melanogaster* Release 6 (update) from FlyBase (Thurmond et al., 2019).The species in red bold text are the ones used in BREC experiments. Since the data collection process is still ongoing, the current version of this dataset is continuously evolving.

GUI input options

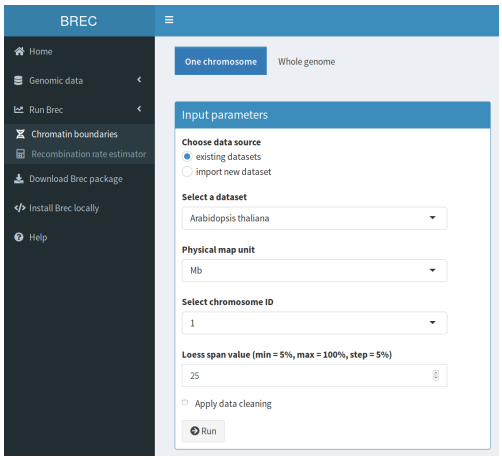
The BREC Shiny interface provides the user with a set of options to select as parameters for a given dataset (see Figure 2.17a). These options are mainly necessary in case the user works on his/her own dataset and this way the appropriate parameters would be available to choose from.

First, a tab to specify the running mode (one chromosome). Then, a radio button group to choose the dataset source (existing within BREC or importing new dataset). For the existing datasets case, there is a drop-down scrolling list to select one of the available genomes (over 40 options), a second one for the corresponding physical map unit (Mb or pb) and a third one for the chromosome ID (available based on the dataset and not the genome biologically speaking). While for the import new dataset case, three more objects are added (see Figure 2.17b); a fileInput to select csv data file, a textInput to enter the genome name (optional), and a drop-down scrolling list to select the data separator (comma , semicolon or tab character -set as the default-).

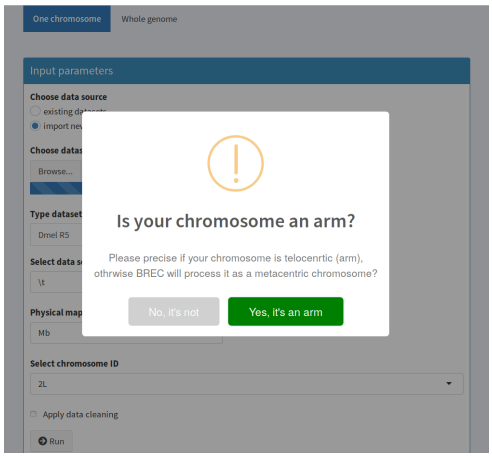
(A) **Inputs - 1** Run BREC for heterochromatin boundaries page, indicated on the left dark panel.

(B) **Inputs - 2** After selecting input parameters and clicking the "Run" button, a popup alert is displayed to ask the user to confirm the chromosome type.

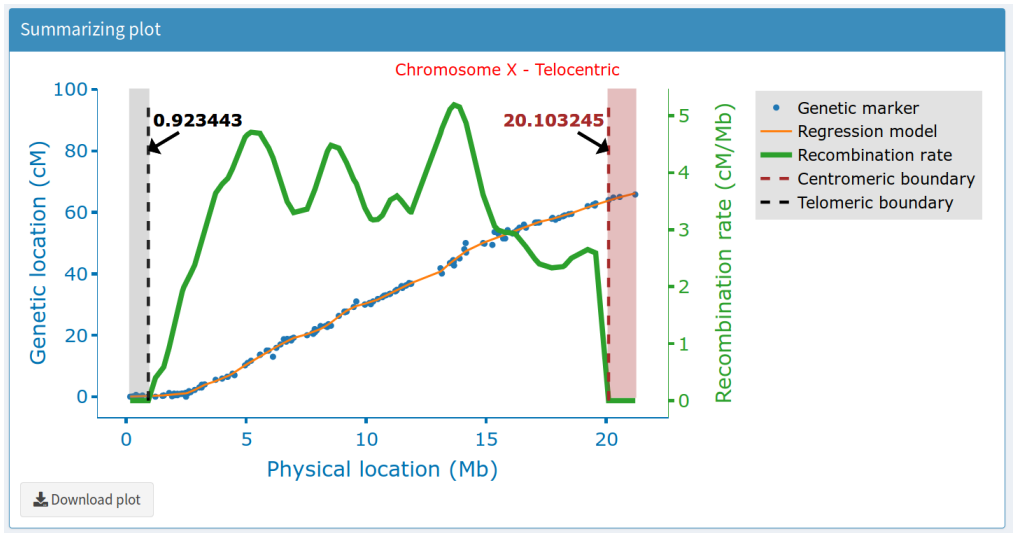
(a)



(b)



(c)



(C) **Outputs -** Here, the interactive summarizing plot of BREC main results is showing the telocentric chromosome X. Respectively with the plot legend order, it includes the input genetic markers (blue dots), the generated regression model (orange line), the local recombination rate estimates (green line), the centromeric boundary (dashed red vertical line on the right) delimiting the centromeric region (highlighted in light red), and the telomeric boundary (dashed black vertical line on the left) delimiting the telomeric region (highlighted in light grey)

FIGURE 2.17: **Screenshots of BREC web application - Run BREC web page** (2.17a) and (2.17b) show the inputs interface. (2.17c) shows the output of running BREC on the specified inputs, represented with an interactive web-based plot as a result.

As for the Loess regression model, the span parameter is required. It represents the percentage of how many markers to include in the local smoothing process. There is

a numericInput object set by default at value 15% with an indication about the range of the span values allowed (min = 5%, max = 100%, step = 5%). The user should keep in mind that the span value actually goes from zero to one, yet, in a matter of simplification, BREC handles the conversion on it's own. Thus, for example, a value of zero basically means that no markers are used for the local smoothing process by Loess, and so, it will induce a running error. Lastly, there is a checkbox to apply data cleaning if checked. Otherwise, the cleaning step will be skipped. This options could save the user some running time if s/he already have *a priori* knowledge that a specific genome's dataset has already been manually curated). The user is then all set to hit the Run button. BREC will start processing the chromosome of interest by identifying its type (telocentric or atelocentric). Since this step is quite difficult to automatically get the correct result, the user might be invited to interfere *via* a popup alert asking for a chromosome type confirmation (see Figure 2.17b).

As shown in Figure 2.18a, all available genomes could be accessed from the left-hand panel (in dark grey) and specifically on the tab "Genomic data" where two pages are available: "Download data files" which provides a data table corresponding to the selected genome on a scrolling list along with download buttons, and "Dataset details" displaying a more global overview of the whole build-in repository (see Figure 2.18b). To give a glance at the GUI outputs, Figure 2.17c shows BREC results displayed within an interactive plot where the user will have the an interesting experience by hovering over the different plot lines and points, visualising markers labels, zooming in and out, saving a snapshot as a PNG image file, and many more available options thanks to the plotly package (Sievert, 2020).

(A) **(Top)**

Download data files page from the Genomic data section, indicated on the left dark panel, is displayed here. After selecting on the top list the *Gallus gallus* genome and clicking the "Download selected" button, a dialog box is open waiting for the user to specify the file path to save the selected data file.

(B) **(Bottom)**

Dataset details page from the Genomic data section is showing a sample of ten available genomes provided within the BREC package. The table is intentionally sorted using the forth column values with descending number of "Total markers".

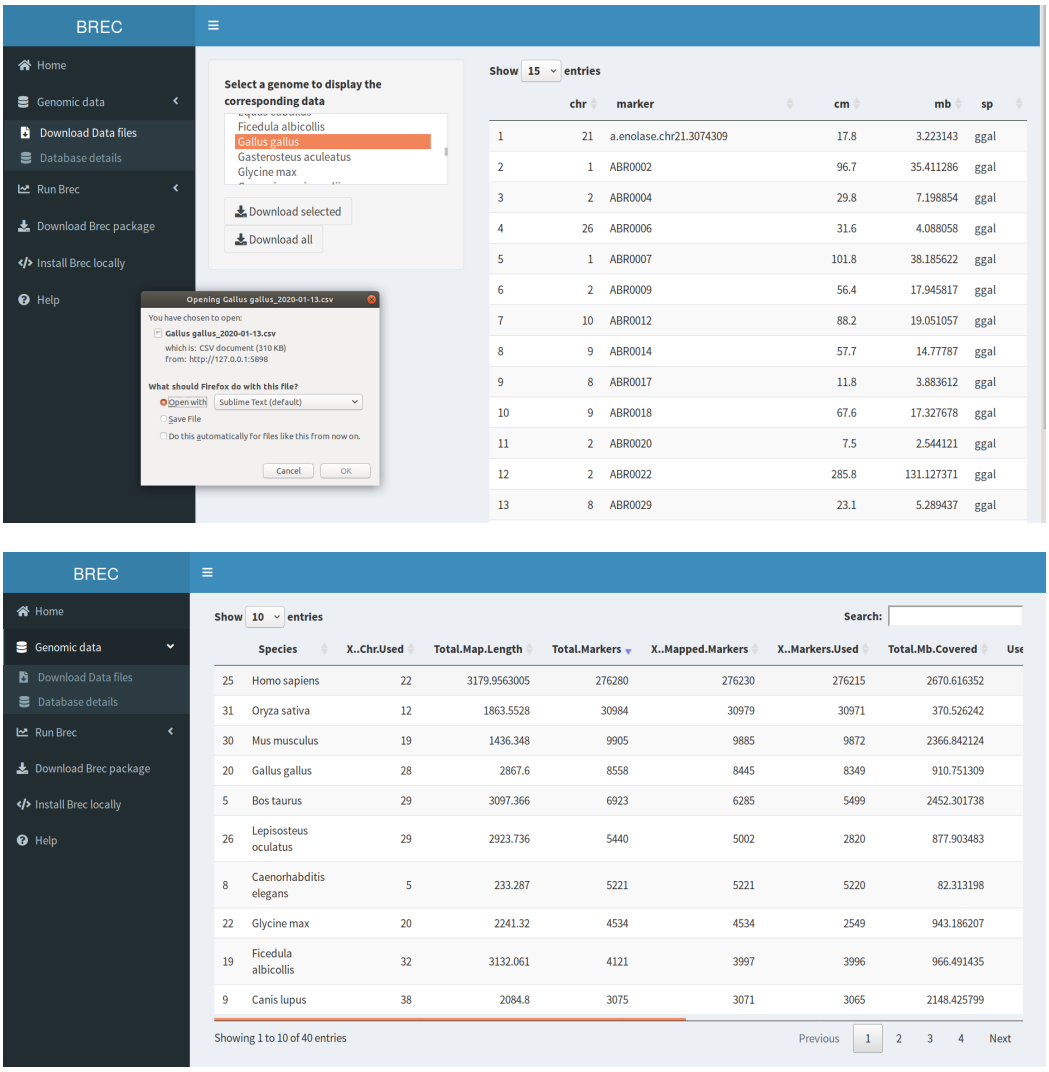


FIGURE 2.18: **Screenshots of BREC web application - Genomic data web pages.**

2.4 BREC results

In this section, we present the results obtained through the following validation process. First, we automatically re-identified HCB with an approximate resolution to the reference equivalents. Second, we tested the robustness of BREC methods according to input data quality, using the well-studied *D. melanogaster* genome

data, for which recombination rate and HCB have already been accurately provided (Fiston-Lavier et al., 2010; Comeron, Ratnappan, and Bailin, 2012; Chan, Jenkins, and Song, 2012; Langley et al., 2012)(Figure 2.19). Besides, we extended the robustness test to a completely different genome, the domesticated tomato *S. lycopersicum* (Sato et al., 2012) to better interpret the study results. Even if the Loess span value does not impact the HCB identification, but only the resulting recombination rate estimates, the span values used in this study are: 15% for *D. melanogaster* (for comparison purpose) and 25% for the rest of the experiments. Our analysis shows that BREC is applicable to data from various organisms, as long as the data quality is good enough. BREC is data-driven, thus, the outputs strongly depend on the markers density, distribution, and chromosome type identified (automatically, or with the user's *a priori* knowledge).

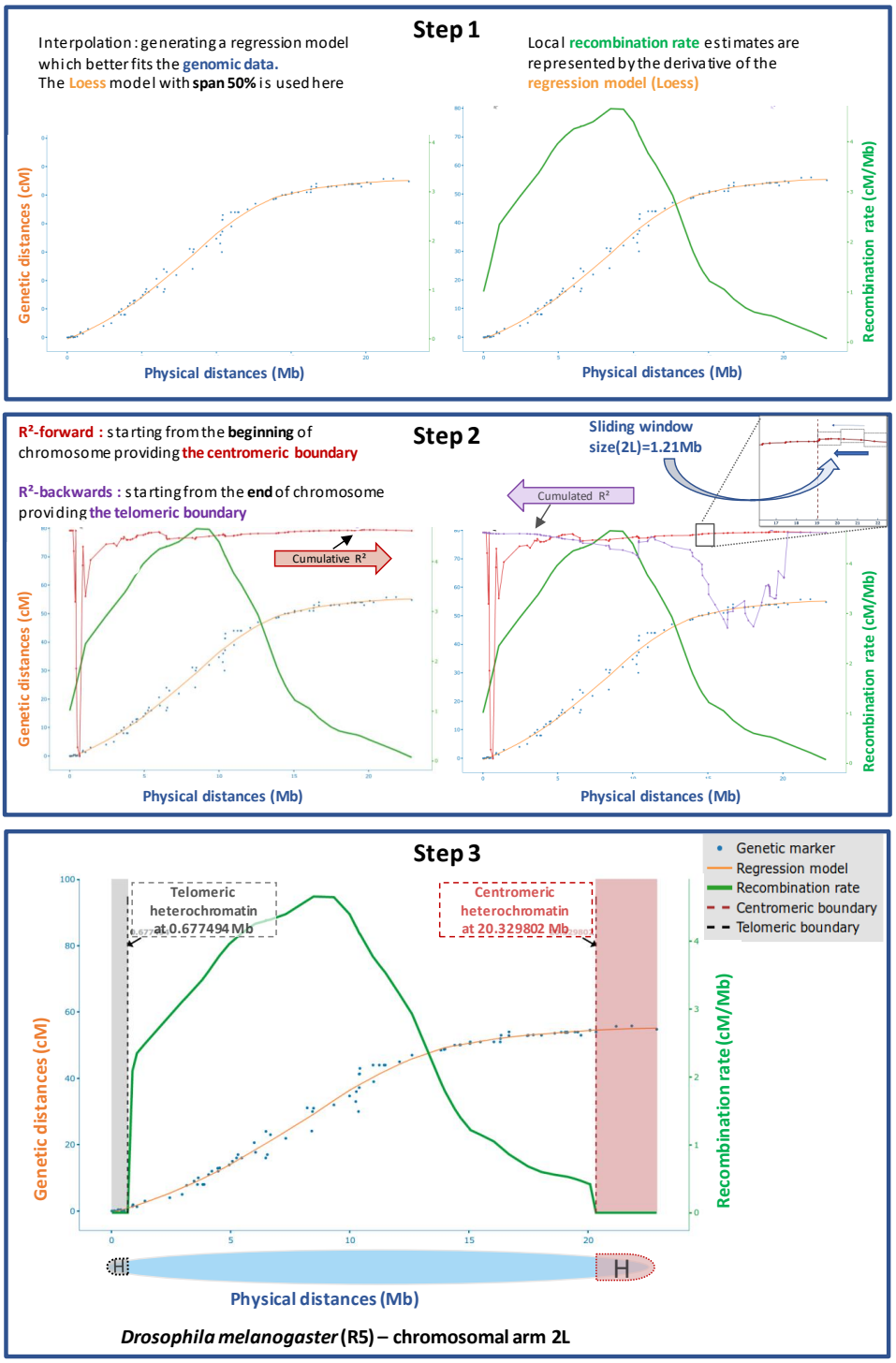


FIGURE 2.19: **BREC workflow steps applied on chromosomal arm 2L of *D. melanogaster* Release 5.** For each one of the five plots, the x and both y axes are the same. The x-axis represents physical distances (Mb). The left y-axis represents genetic distances (cM) shared between markers (blue data points) and the regression model (orange line). The right y-axis represents recombination rates (cM/Mb) for local estimates (green line). For simplification and less redundancy purposes, in steps 1 and 2, both y axes are written only once to be complementary for both plots: the left as well as the right one. R^2 values, varying between zero and one, are following $R^2 - forward$ (red line) and $R^2 - backwards$ (purple line). Left telomere and Right centromere (resp. black and purple dashed lines) indicate HCB for the corresponding identified heterochromatin region.

2.4.1 Approximate, yet congruent HCB

Chromosomal arm	Centromeric (Mb)			Telomeric (Mb)		
	Boundaries		Shift	Boundaries		Shift
	Reference	BREC		Reference	BREC	
X	20.67	20.10	0.56	2.46	0.92	1.54
2L	19.95	20.33	0.38	0.70	0.68	0.02
2R	6.09	5.01	1.08	20.02	20.71	0.69
3L	18.41	20.30	1.90	0.36	2.26	1.91*
3R	8.35	3.77	4.58*	27.25	25.64	1.61
Min. shift	0.38			0.02		
Max. shift	4.58			1.91		
Mean shift	1.70			1.15		
Median shift	1.08			1.54		

TABLE 2.4: **BREC HCB compared to reference boundaries from the reference genome of *D. melanogaster*.** The shift is the absolute value of the distance between the BREC and the reference physical heterochromatin boundary. The first five rows represent all chromosomal arms. Grouped columns present reference, BREC and shift values for the centromeric boundaries (Columns 2-4), and for the telomeric boundaries (Columns 4-6). Here the boundary values correspond to the internal HCB. The external boundaries are represented by the physical positions of the first and the last markers of the chromosomes. All values are expressed in Megabase (Mb). The red asterisk indicates the largest shift value reported on centromeric and telomeric boundaries separately (see corresponding Figure 2.20). The last four rows represent general statistics on the shift value. From top to bottom, they are minimum, maximum, mean, and median respectively. See details on the shift metrics in Section Validation process: Validation metrics (2.3.3).

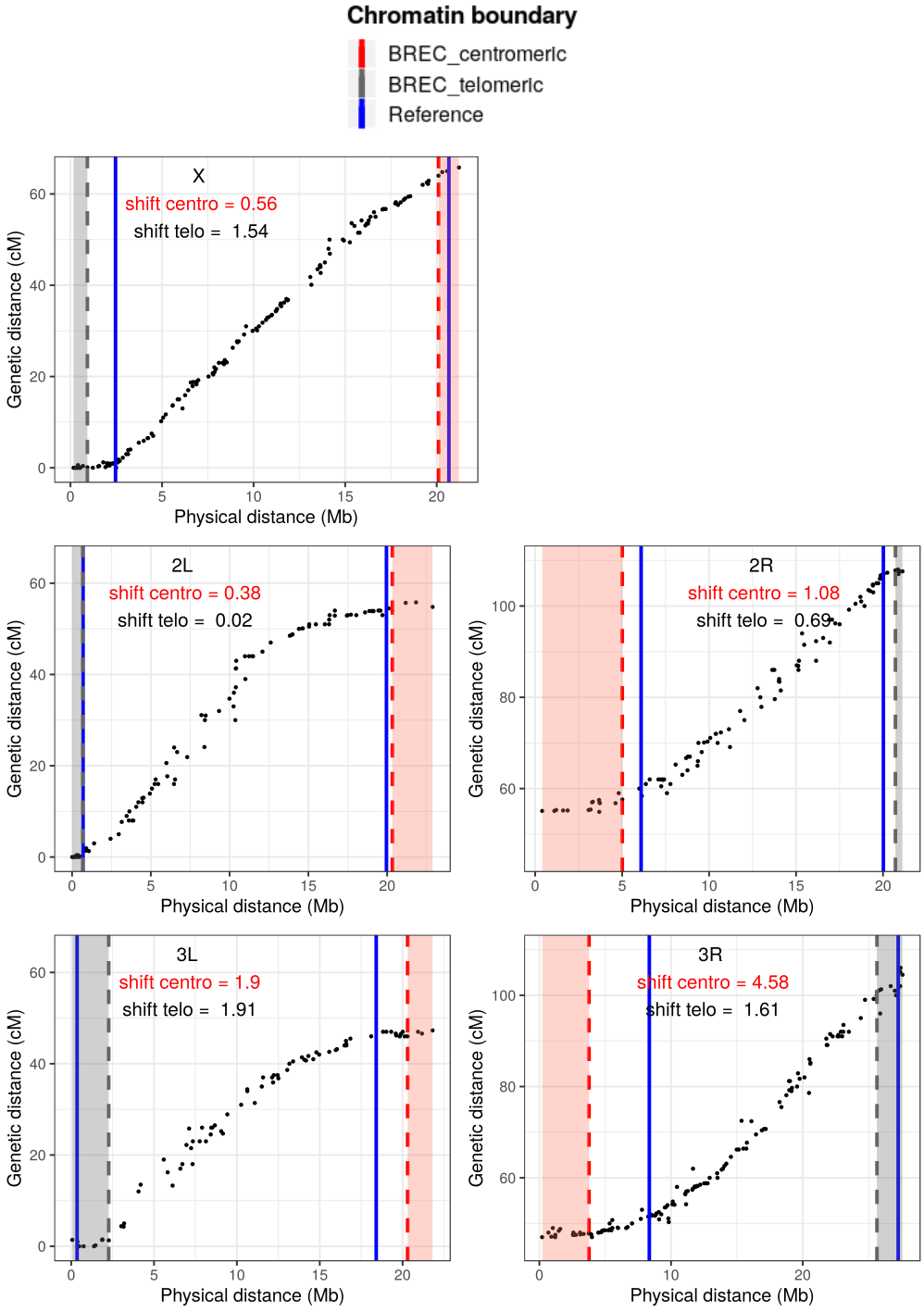


FIGURE 2.20: **Plots representing results of BREC and reference HCB on the *D. melanogaster* genome**. The results are summarized in Table 2.4. From top to bottom are the five chromosomal arms X, 2L, 2R, 3L, 3R, respectively. Black dots represent genetic markers in ascendant order according to their physical position (in Mb). Vertical lines represent HCB for BREC centromeres (in red dashed line), for BREC telomeres (in grey dashed line) and for the reference (in solid blue line). The heterochromatin regions identified by BREC are highlighted for the centromere (in red) and the telomere (in grey). For each chromosomal arm, two shift values of centromeric and telomeric boundaries are shown under the chromosome identifier.

Fruit fly genome *D.melanogaster*

Our approach for identifying HCB has been primarily validated with cytological data experimentally generated on the *D. melanogaster* Release 5 genome (Riddle et al., 2011; Chan, Jenkins, and Song, 2012; Langley et al., 2012; Thurmond et al., 2019). For all five chromosomal arms (X, 2L, 2R, 3L, 3R). This genome presents a mean density of 5.39 markers/Mb and a mean physical map length of 22.92Mb. We obtained congruent HCB with a good overlap and shift, distance between the physical position of the reference and BREC, from 20Kb to 4.58Mb (see Section Validation process: Validation process). We did not observe a difference in terms of mean shift for the telomeric and centromeric BREC identification (χ^2 = 0.10, df = 1, $p - value$ = 0.75)(See Tables 2.1, 2.4). We observe a lower resolution for the chromosomal arms 3L and 3R (see Figure 2.20). This suggests that those two chromosomal arms' data might not present as good quality as the rest of the genome. Interestingly, the local markers density for these two chromosomal arms shows a high variation, unlike the other chromosomal arms. For instance, the 2L for which BREC returns accurate results, shows a lower variation (see Figure 2.14). Without these two arms, the max shift for both centromeric and telomeric BREC boundaries is smaller than 1.54Mb, with a mean shift decreasing from 1.43Mb to 0.71Mb.

This first analysis suggests that BREC methods return accurate results on this genome. However, the boundaries identification process appears very sensitive to the markers' local density and distribution along a chromosome (see Figure 2.20). Therefore, we conducted further experiments on a different dataset, the tomato genome (see Figure 2.15).

Tomato genome *S. lycopersicum*

Results of experimenting BREC behaviour on all 12 chromosomes of *S. lycopersicum* genome (Sato et al., 2012) are shown as values in Table 2.5 and as plots in Figure 2.21. This genome presents a mean density of 2.64 markers/Mb and a mean physical map length of 62.71Mb. We observe a variation in the shift value representing the difference on the physical map between reference HCB and their equivalents returned by BREC. Unlike the *D. melanogaster* genome, which is of a smaller size, with five telocentric chromosomes (chromosomal arms) and a strongly different markers distribution, the tomato genome exhibits a completely different study case. It is a plant genome, with approximately 8-fold bigger genome size. It is organized as twelve atelocentric chromosomes of a mean size of 60Mb, except for chromosomes 2 and 6, which are more likely to be rather considered telocentric based on their markers distribution. Also, we observe a long plateau of markers along the centromeric region with lower density than the rest of the chromosomes. Something which highly differs from *D. melanogaster* data. We believe all these differences between both genomes give a good validation and evaluation for BREC behavior towards various data quality scenarios. Furthermore, since BREC is a data-driven tool, these experiments help analyze data-related limitations that BREC could face while resolving differently. From another point of view, BREC results on the tomato genome highlight the fact that markers distribution along heterochromatin regions, in particular, strongly impacts the identification of eu-heterochromatin boundaries, even when the density is of 2 markers/Mb or more.

Chromosome	Centromeric left (Mb)			Centromeric right (Mb)		
	Boundaries		Shift	Boundaries		Shift
	Reference	BREC		Reference	BREC	
1	5.78	22.88	17.09	67.80	76.48	8.68
2	3.15	1.51	1.64	27.43	21.31	6.12
3	5.75	6.98	1.23	55.34	49.28	6.06
4	5.48	1.21	4.27	54.92	47.21	7.72
5	6.02	15.03	9.01	60.23	51.04	9.19
6	1.50	1.68	0.19	29.62	20.42	9.20
7	5.62	23.05	17.43	52.51	33.52	18.98*
8	5.10	22.87	17.77	51.73	43.96	7.77
9	4.38	32.51	28.12*	61.16	49.16	12.00
10	4.40	24.37	19.97	58.83	49.92	8.91
11	5.56	10.86	5.29	47.57	32.77	14.80
12	7.27	14.34	7.07	60.27	54.33	5.94
Min. shift	0.19			5.94		
Max. shift	28.12			18.98		
Mean shift	10.76			9.61		
Median shift	8.04			8.80		

TABLE 2.5: **Results of BREC and reference HCB on the genome of *S. lycopersicum*.** The shift is the absolute value of the distance between the BREC and the reference physical heterochromatin boundary. The first twelve rows represent all chromosomes. Grouped columns present reference, BREC and shift values for the left centromeric boundaries (Columns 2-4), and for the right centromeric boundaries (Columns 4-6). All values are expressed in Megabase (Mb). The red asterisk indicates the largest shift value reported on centromeric and telomeric boundaries separately (see corresponding Figure 2.21). The last four rows represent some general statistics on the shift value. From top to bottom, they are minimum, maximum, mean, and median respectively. See details on the shift metrics in Section Validation process: Validation metrics. 2.3.3.

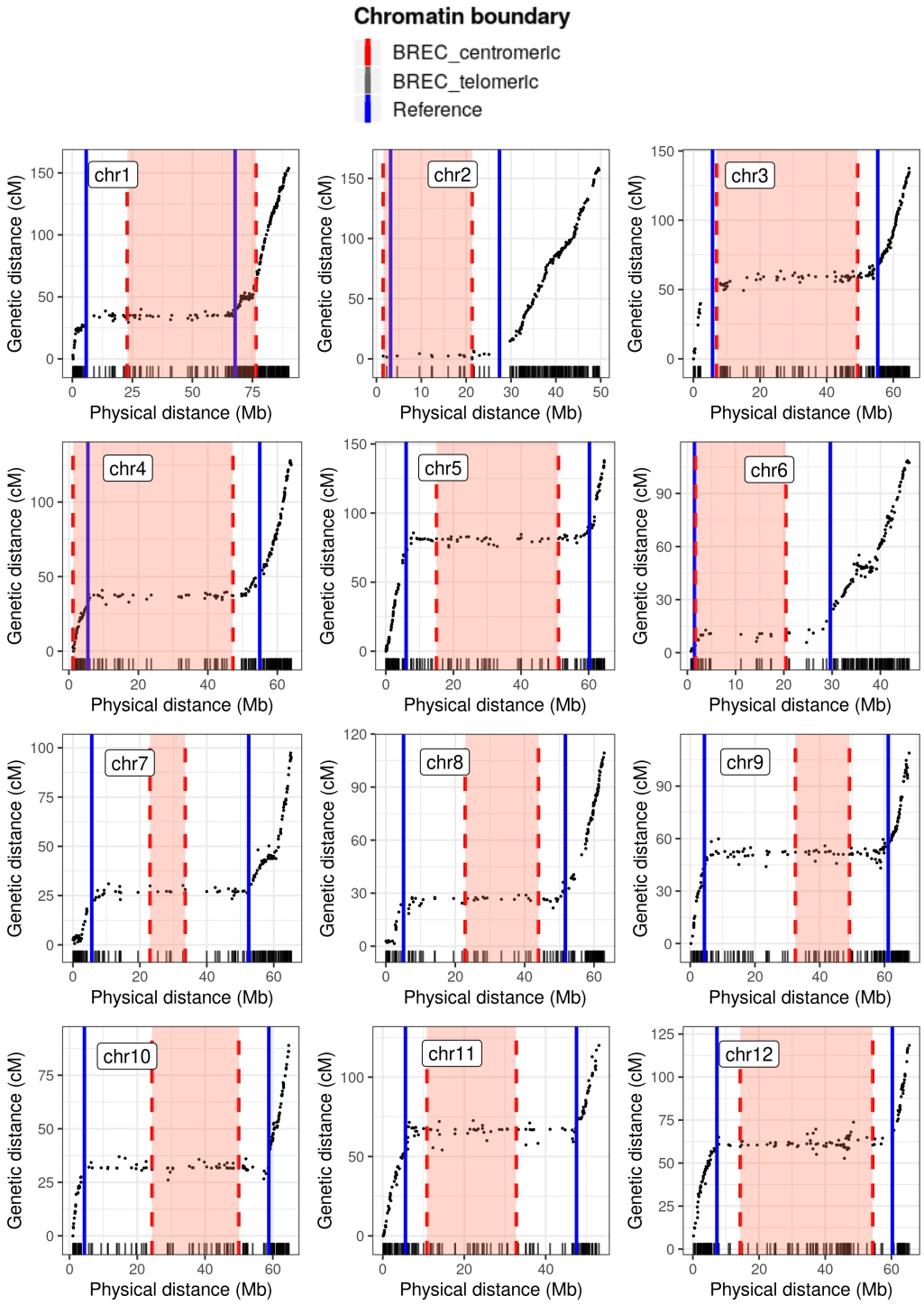


FIGURE 2.21: **Plots representing results of BREC and reference HCB on the *S. lycopersicum* genome**. The results are summarized in Table 2.5. From top to bottom are the twelve chromosomes 1 to 12, respectively. Black dots represent genetic markers in ascendant order according to their physical position (in Mb). Vertical lines represent HCB for BREC centromeres (in red dashed line), and for the reference (in solid blue line). The heterochromatin regions identified by BREC are highlighted for the centromere (in red). Rug plot on the x-axis represents the markers density according to the physical map.

2.4.2 Consistency despite the low data quality

We aim in this part to study to what extent BREC results are depending on the data quality.

BREC handles low markers density

We started by assessing the markers' density on the BREC estimates. We generated simulated datasets with decreasing fractions of markers for each chromosomal arm (from 100% to 30%). For that, we randomly selected a fraction of markers, 30 times, and computed the mean shift between BREC and the reference telomeric and centromeric boundaries. We have noted that BREC's resolution decreases drastically with the fraction and therefore with the marker density (see Figure 2.22). However, BREC results appeared stable until 70% of the data for all the chromosomal arms, more specifically for the telomeric boundary detection. Only for the centromeric boundary of the chromosomal arm 3R, we observed the opposite pattern: BREC returns more accurate telomeric boundary estimates when the markers' number decreases. This supports the low quality of the data around the 3R centromere.

This simulation process allowed to set a minimum density threshold representing the minimum value for data density in order to guarantee accurate results for BREC estimates at 5 markers/Mb (fraction of around 70% of the data) on average in *D. melanogaster*. This analysis also supports the fact that because the markers' density alone can not explain the BREC resolution, BREC may also be sensitive to the marker distribution.

Figure 2.14 clearly shows that markers density varies within and between the five chromosomal arms with a mean of 4 to 8 markers/Mb. The variance is induced by the extreme values of local density, such as 0 or 24 markers/Mb on the chromosomal arm X. Still, the overall density is around 5 markers/Mb for the whole genome.

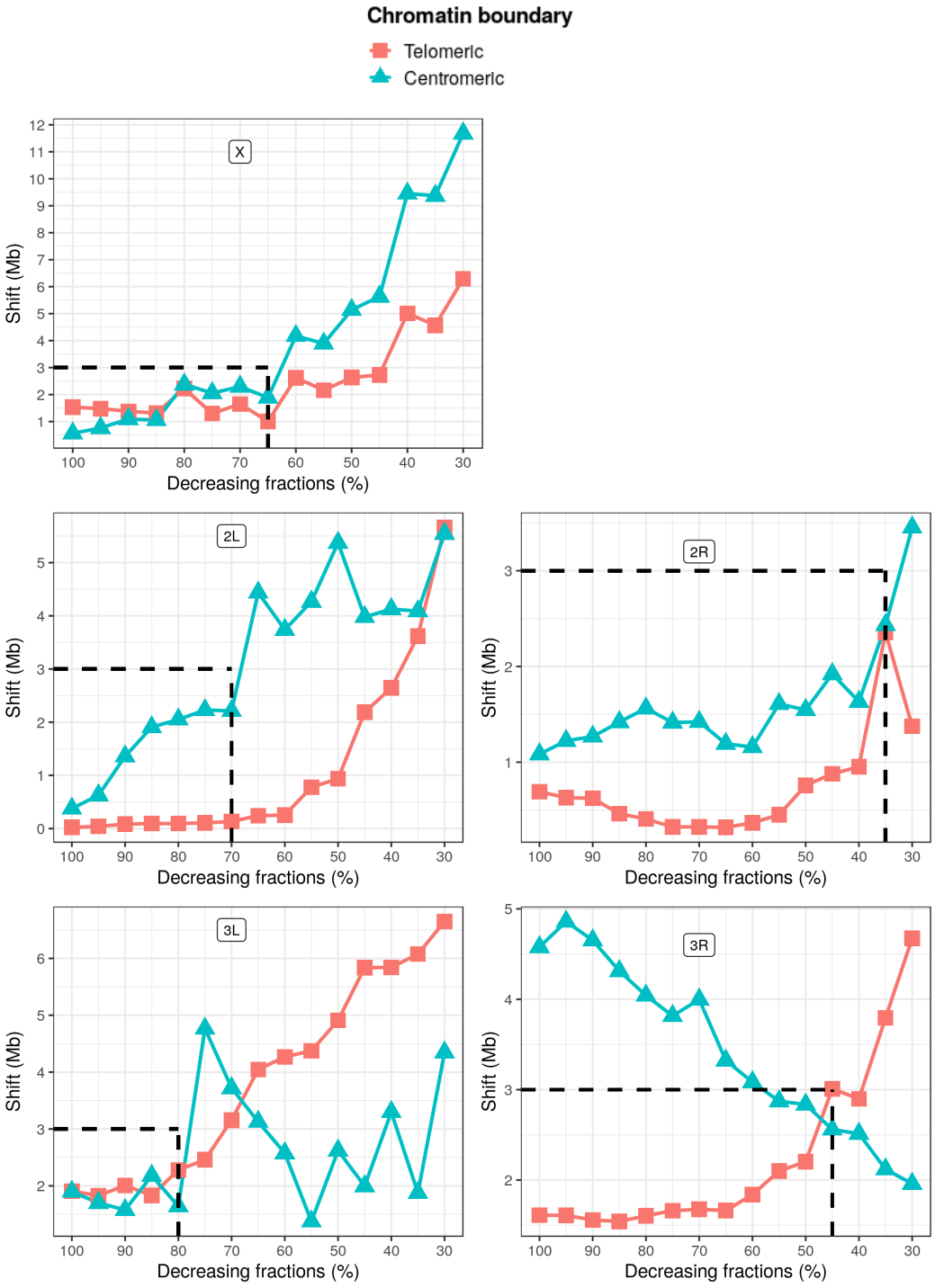


FIGURE 2.22: **The impact of decreasing markers density on the resolution of BREC's HCB expressed by the shift value.** Here is an overview of the variation of shift values (see Equation 2.2) for BREC's HCB compared to reference results for the five *D. melanogaster* chromosomal arms (X, 2L, 2R, 3L, 3R). For each arm, two HCB are shown: squares (in red) for telomeric and triangles (in light blue) for centromeric boundaries. The horizontal dashed line (in black) delimits results smaller than a shift value of 3Mb for all arms while the vertical dashed line (in black) indicates up to which fraction the 3Mb shift is conserved on each chromosomal arm's simulations. Note that the x axis is reversed, so from left to right it goes from 100% to 30% with a step of -5%at each point. The simulation process is further clarified for one fraction on the chromosomal arm 2L and is illustrated in Figure 2.23.

BREC handles heterogeneous distribution

Along chromosomes, genetic markers are not homogeneously distributed. Therefore, to assess the impact of the distribution of markers on BREC results, we designed different *data scenarios* regarding a reference data distribution (see Section Validation process: Simulated data for quality control testing). We choose as reference the chromosomal arms 2L and 2R of *D. melanogaster* as we have obtained the most accurate results with their data. After the concatenation of the two arms, we ended up with a metacentric simulated chromosome as a starting simulation scenario (total physical length of 44Mb). While this length was kept unchanged, markers local density and distribution were modified (see Section Validation process: Simulated data for quality control testing and Figure 2.13).

One particular yet typical case is the centromeric gap. Throughout our analysis, we consider that a chromosome presents a centromeric gap if its data exhibit a lack of genetic markers on a relatively large region on the physical map. Centromeric regions usually are less accessible to sequence due to their highly compact chromatin state. Consequently, these regions are also hard to assemble, and that is why many genomes have chromosomes presenting a centromeric gap. It is essential to know that a centromeric gap is not always precisely located in the middle of a chromosome. Instead, its physical location depends on the chromosome type (see more details in Figure 2.11).

We also assess the veracity of BREC on datasets with variable distributions using simulated data with and without a centromeric gap (see Figure 2.13).

For all six simulation datasets, BREC results overlap the reference boundaries. Thus BREC correctly handles the presence of a centromeric gap (see Figure 2.13: (a)(c)(e)). BREC remains robust to a non-uniform distribution of markers, under the condition that regions flanking the boundaries are greater than 2 markers/Mb (see Figure 2.23). In the case of a non-uniform distribution, BREC resolution is higher when the local density is stronger around heterochromatin regions (see Figure 2.13: (c)(d)(e)(f)). This suggests that low density on euchromatin regions far from the boundaries is not especially a problem either.

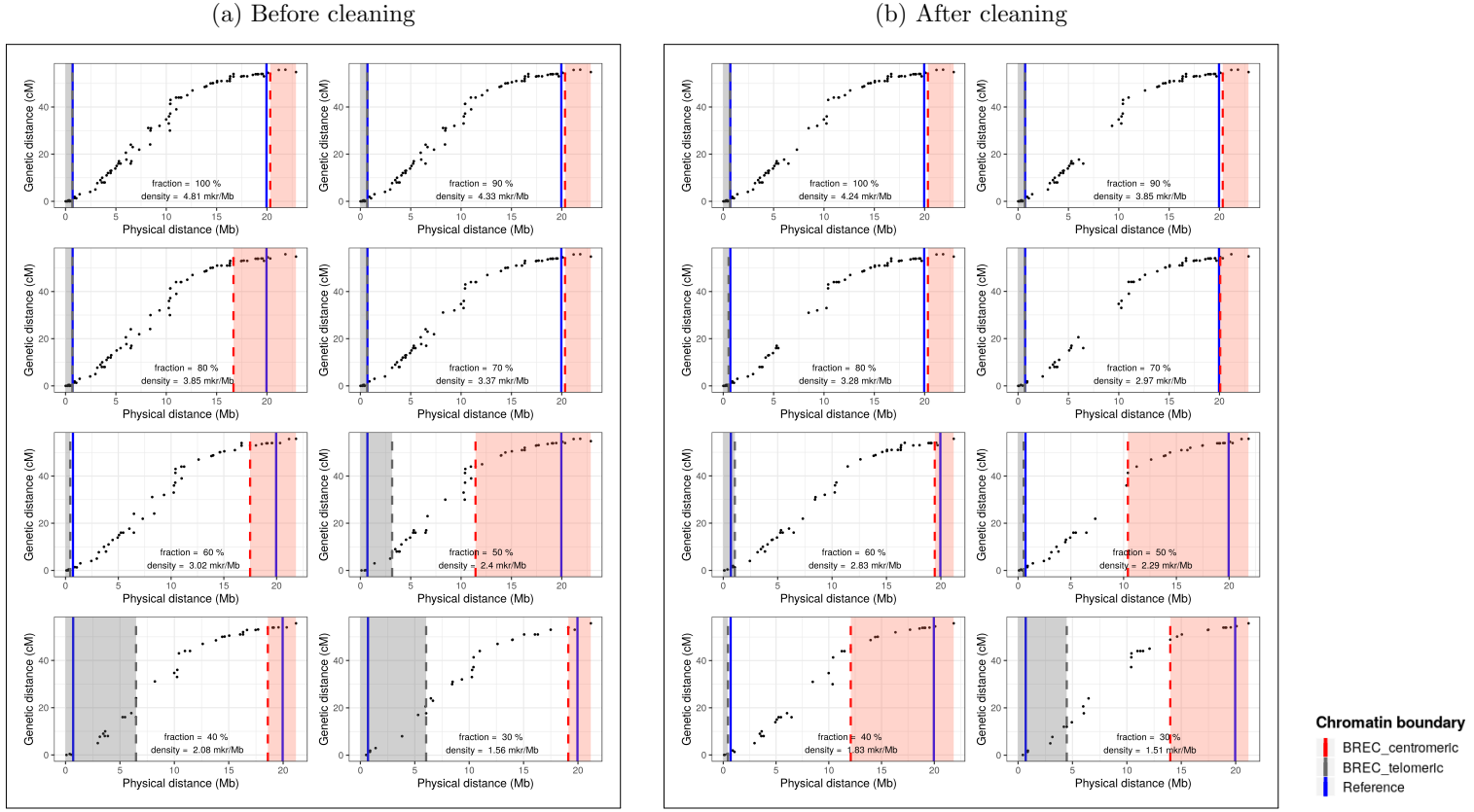


FIGURE 2.23: **Low density simulations** BREC results on the simulated telocentric chromosomes with different density scenarios. Simulating decreasing markers density going from 100% to 30% of the original chromosome 2L (of size 23Mb) of the *D. melanogaster* Release 5 genome. These simulations allow to study the impact of variable markers density on BREC results compared to reference HCB. (a) on the left is before and (b) on the right is after the cleaning step. These simulations have been conducted on each of the five chromosomes (X, 2L, 2R, 3L, 3R) 30 times where the mean shift value is reported in Figure 2.22. Black dots represent genetic markers. Vertical lines represent HCB for BREC centromeres (in red dashed line), for BREC telomeres (in grey dashed line) and for the reference (in solid blue line). The heterochromatin regions identified by BREC are highlighted for the centromere (in red) and the telomere (in grey). The corresponding fraction and markers density is shown on the top left of each simulation plot.

2.4.3 Accurate local recombination rate estimates

Figure 2.24 is a combined recap of the inputs and outputs of BREC when applied on the whole genome of *D. melanogaster R6*.

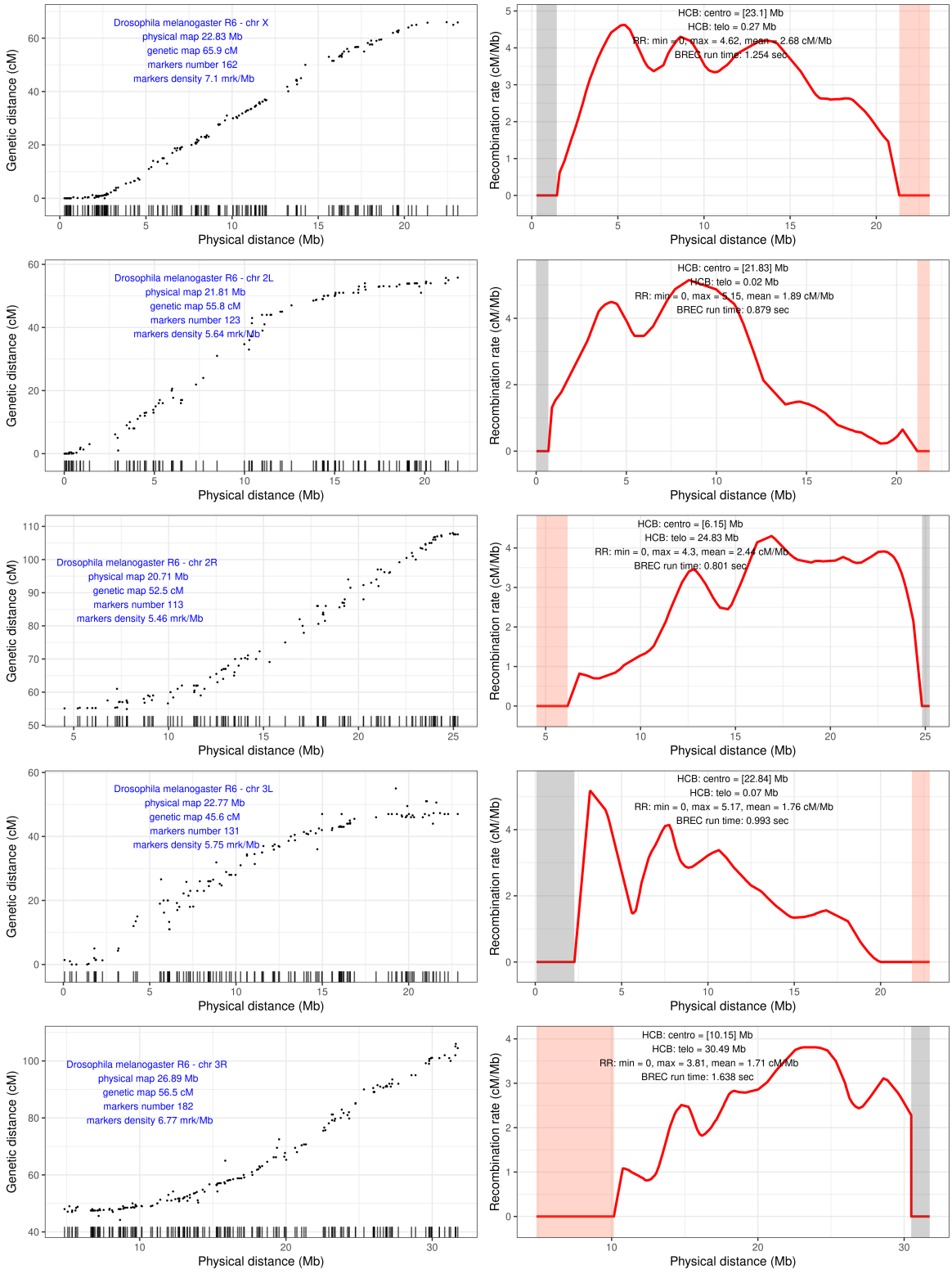


FIGURE 2.24: Genomic features (right) and BREC results (left) for the *D. melanogaster R6* genome.

After identifying the HCB, BREC provides optimized local estimates of recombination rate along the chromosome by taking into account the absence of recombination in heterochromatin regions. Recombination rates are reset to zero across the centromeric and telomeric regions regardless of the regression model. To closely compare the third degree polynomial with Loess, using different span values, we experimented with this aspect on *D. melanogaster* chromosomal arms and reported the results in Figure 2.25.

To assess the veracity of the recombination rates along the whole genome, we compared BREC results with previous recombination rate estimates (see Figure 2.26; (Chan, Jenkins, and Song, 2012; Langley et al., 2012)). BREC recombination rate estimates are significantly strongly correlated with reference data (Spearman's: $P << 0.001$) while the reference estimates fail in telomeric regions.

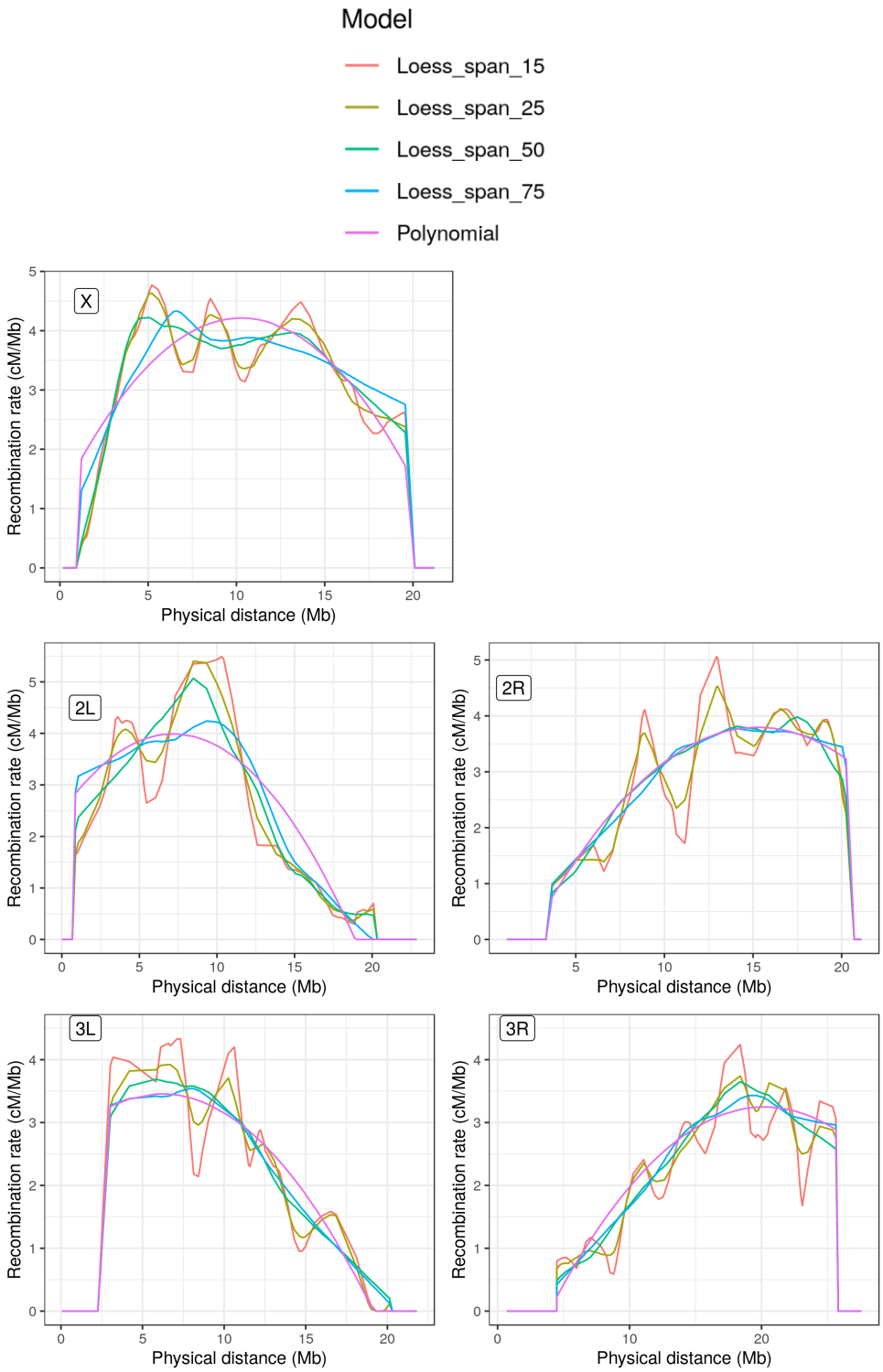


FIGURE 2.25: **Comparison of regression models for recombination rate estimates along the five chromosomes (X, 2L, 2R, 3L, 3R) of *D. melanogaster* Release 5**. Regression models used here are Loess with span values, 15%, 25%, 50%, 75% and third degree polynomial. The HCB defined by BREC remain unchanged and only local recombination rates differ according to the model used to fit the genetic and physical maps. Recombination rate is represented by the derivative of the model. In case of two or more models yielding the same recombination rate estimates on the same physical position, the overlap results in only one curve line. Here, all curves show null recombination rate value on the centromeric and telomeric regions.

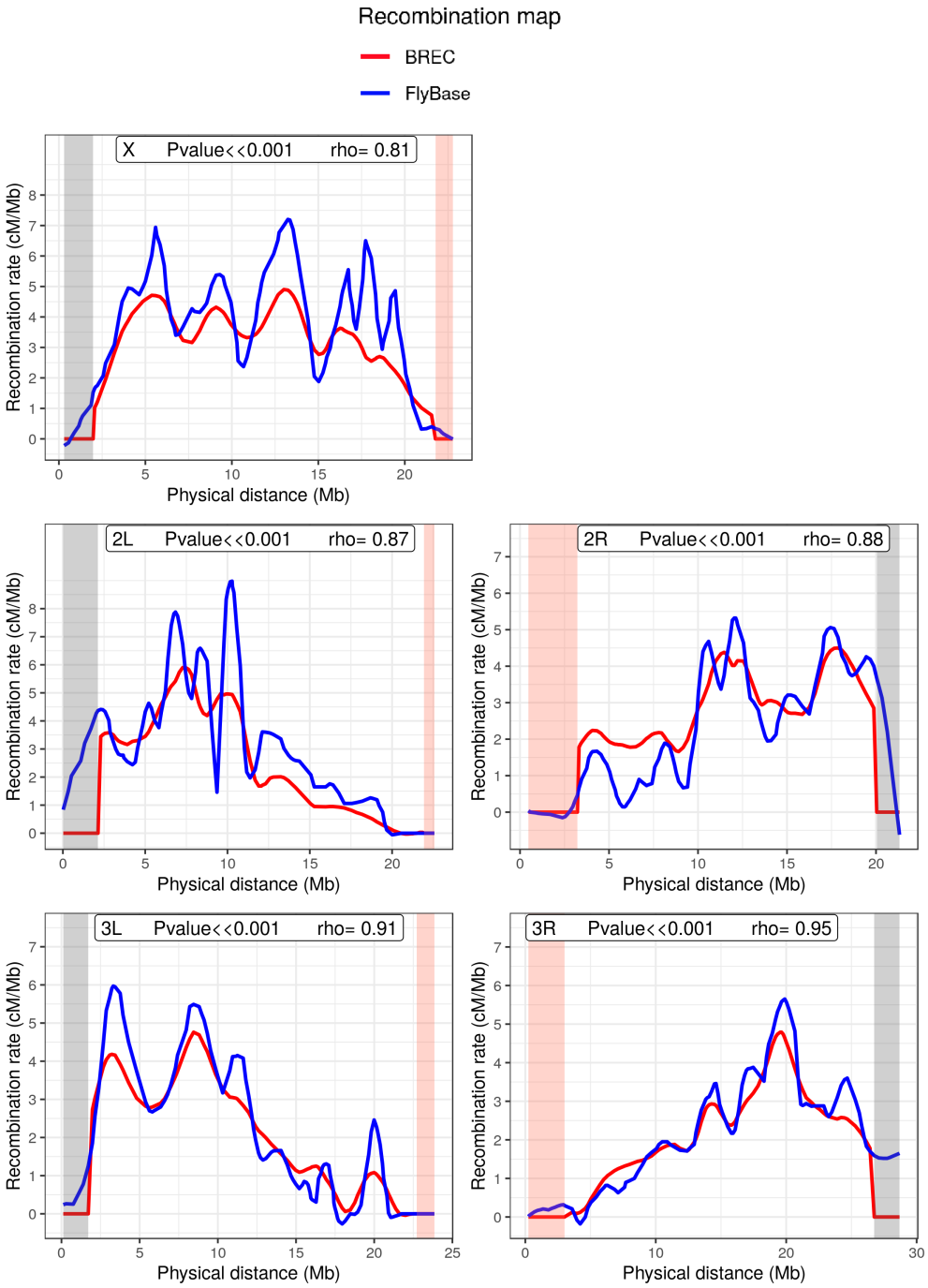


FIGURE 2.26: **Comparison of BREC *vs.* FlyBase recombination rate recombination rates along the five chromosomal arms (X, 2L, 2R, 3L, 3R) of *D. melanogaster* Release 5**. Both recombination maps are obtained using the same regression model: Loess with span 15%. The HCB defined by BREC are represented in red and the reference data are in blue. Heterochromatin regions identified by BREC are highlighted in yellow.

2.4.4 BREC is non-genome-specific

NGS, High Throughput Sequencing (HTS) technologies, and numerous further computational advances are increasingly providing genetic and physical maps with more and more accessible markers along the centromeric regions. Such progress in the availability of data of poorly accessible genomic regions is a huge opportunity to shift our knowledge of heterochromatin DNA sequences and their dynamics, as in the case of Transposable Elements (TEs), for example. Therefore, BREC is not identifying centromeric gaps as centromeric regions as it might seem. Instead, it is targeting centromeric as well as telomeric boundaries identification regardless of the presence or absence of markers neither of their density or distribution variations across such complicated genomic regions (see Figure 2.12). Given that BREC is non-genome-specific, applying HCB identification on various genomes has allowed to widen the experimental design and to test more thoroughly how BREC responds to different *data scenarios*. Despite the several challenges due to data quality issues and following a data-driven approach, BREC is a non-genome-specific tool that aims to help to tackle biological questions.

2.4.5 Easy, fast and accessible tool *via* an R-package and a Shiny app

BREC is an R-package entirely developed with the R programming language (R Core Team, 2018). The current version of the package and documentation are available on the GitHub repository: `https://github.com/GenomeStructureOrganization/BREC`

In addition to the interactive visual results provided by BREC, the package comes with a web-based Graphical User Interface (GUI) build using the shiny and shinydashboard libraries (RStudio, Inc, 2014). The intuitive GUI makes it a lot easier to use BREC without struggling with the command line (see screenshots in Figures 2.17 and 2.18).

As for the speed aspect, BREC is quite fast when executing the main functions. We reported the running time for *D. melanogaster R5* and *S. lycopersicum* in Tables 2.1 and 2.2, respectively (plotting excluded). Nevertheless, when running BREC *via* the Shiny application, and due to the interactive plots displayed, it takes longer because of the plotly rendering. Still, it depends on the size of the genetic and physical maps used, as well as the markers density, as slightly appears in the same tables. The results presented from other species (see Figure 2.12) highlight better this dependence.

2.5 Applying BREC to identify chromatin regions along the mosquito genome: *Ae. aegypti*

In this section, we present some preliminary results obtained using BREC (see Chapter 2) on the mosquito genomes presented in Chapter 1. We then carry out the study on TE distribution in the mosquito genomes and assess the association between some TE families known to be actives and the recombination pattern along the chromosomes. In March 2017, a research group (Dudchenko et al., 2017) ended up with the first chromosome-length scaffolds for *Cx. pipiens quinquefaciatus* and *Ae. aegypti* genomes. According to the TE evolutionary models, we may expect to observe an enrichment of TEs in regions poor in genes and regions of reduced recombination

(Petrov et al., 2011). To test this hypothesis, we estimated the local recombination rates using BREC (see Chapter 2) on both mosquito genomes *Ae. aegypti*and *Cx. pipiens quinquefaciatus*. Two mosquito genomes were recently released at the beginning of this study : *Ae. aegypti* and *Cx. pipiens* for which the linkage maps were updated. For both genomes, the quality have been improved adding Hi-C information to contigs from the previous versions (Dudchenko et al., 2017). The authors re-sequenced both genomes and checked the order of genetic markers to assess the quality on these new assemblies (see Figure 2.27). In the following, we indifferently use *Cx. pipiens* or *Cx. pipiens quinquefasciatus* to refer to the latter.

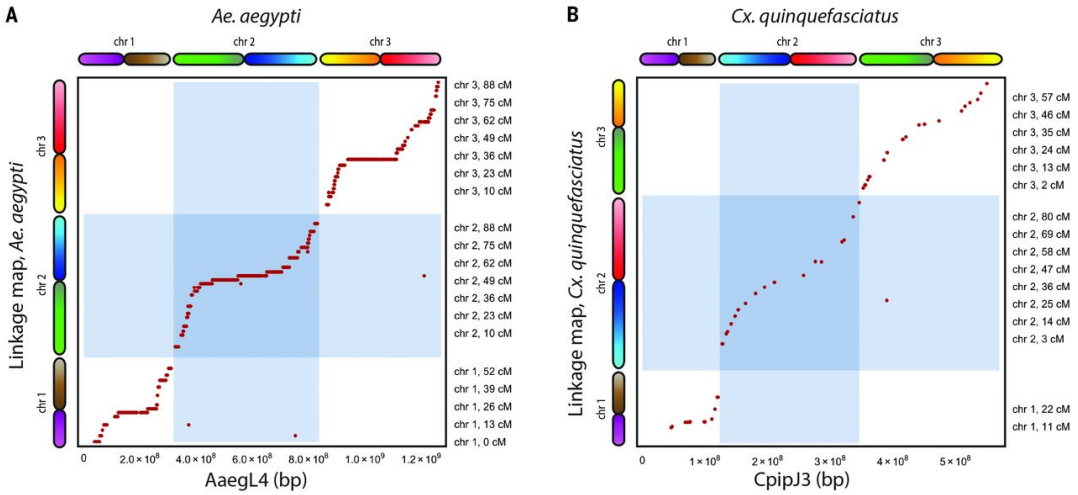


FIGURE 2.27: **Comparison of *AaegL4* and *CpipJ3* with genetic maps.** (A) They compared *AaegL4* with a genetic map of *Ae. aegypti*. Their assembly agreed with the genetic map on 1822 out of 1826 markers. The exceptions are due to misjoins in *AaegL2* that were not corrected in *AaegL4*. (B) Similarly, *CpipJ3* is in agreement with a genetic map of *Cx. pipiens quinquefasciatus*. [Figure by (Dudchenko et al., 2017)]

As BREC is a data-driven tool, we started by investigating the quality of the data. The *Ae. aegypti* genome release used here is *AaegL4* where the genetic and physical maps were downloaded from (Dudchenko et al., 2017). The whole genome dataset provides three huge chromosome-length scaffolds of 307, 472, and 404Mb, with 317, 923, and 586 markers, respectively, providing an average density of 1.49 markers/Mb for a total genetic map length of 235 cM. We then launched BREC. BREC identifies 3 metacentric chromosomes with a large decrease of recombination in the middle and also at the extremities of the chromosomes. The centromeric regions including the pericentromeric regions range from 0.42Mb (chr3) to 52.99Mb (chr1) (see Figure 2.28).

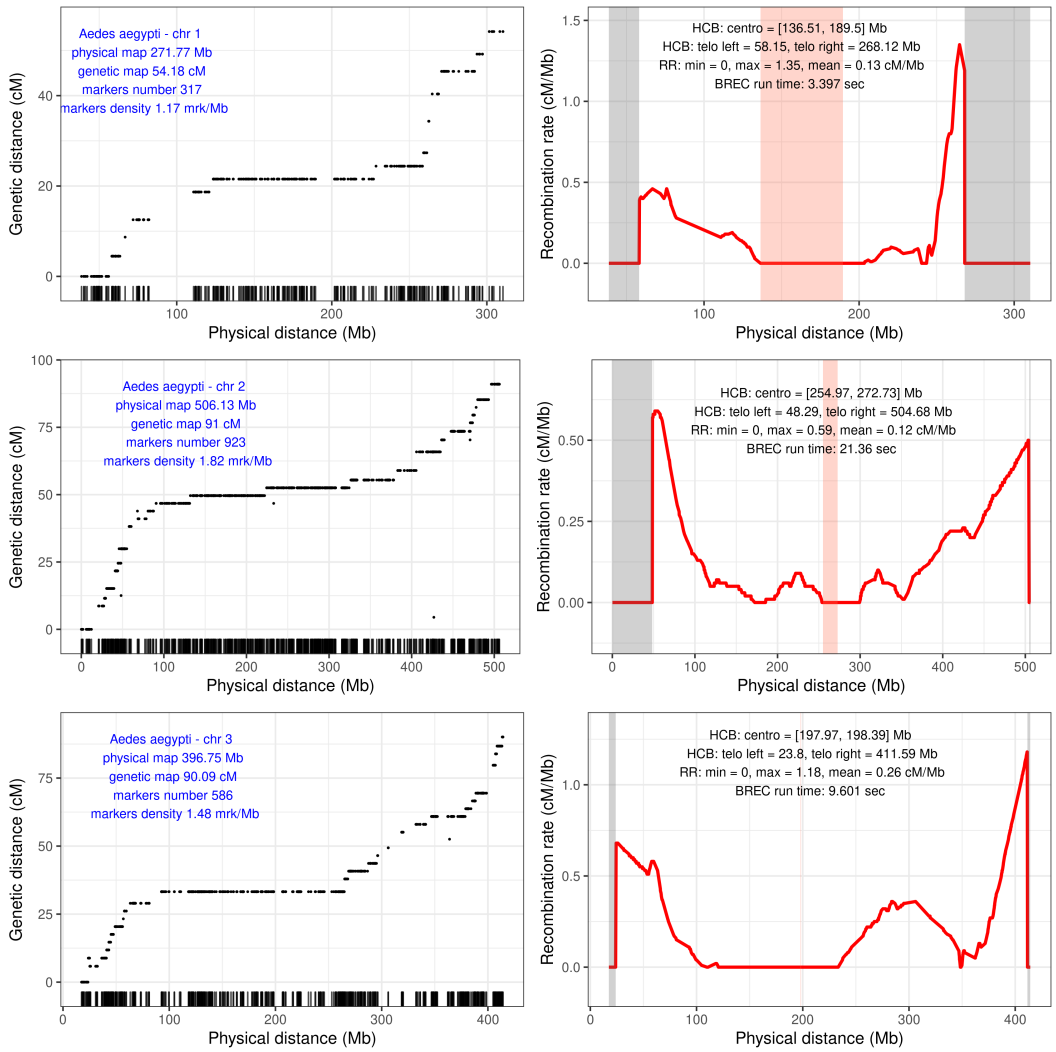


FIGURE 2.28: Genomic features (right) and BREC results (left) for the *Ae. aegypti AaegL4* genome. A specific pattern is observed on the three chromosomes where a large plateau region around the centromere is highlighting almost no variation on the genetic map, and expected to yield large heterochromatin regions with reduces/suppressed recombination rates.

According to (Matthews et al., 2018), it is not relevant comparing AaegL5 to AaegL4 as these genomes derive from different strains. In addition, there is a high degree of natural diversity between *Ae. aegypti* strains. The authors estimated that only 70% of the older AaegL4 reference aligns to the new AaegL5 assembly with >95% identity (Matthews et al., 2018). The comparison of the two assemblies with an old assembly version (AaegL3) revealed only very few shifts of coordinates. In addition, the analysis of the nucleotide diversity between several *Ae. aegypti* strains highlight putative centrometic regions. The approximate physical position of the pericentromere/centromere reported on the latest *AaegL5* version of this genome are: chr1: 145-177Mb/166Mb, chr2: 219-258Mb/243Mb, chr3: 184-219Mb/206Mb (Matthews et al., 2018) (Supp Data 12). BREC chromatin boundaries showed a clear overlap for the centromeric regions with 4 to 12Mb distance from the centromeres to the closest BREC boundaries (chr1: 137-170Mb, chr2: 255-273Mb, chr3: 197-198Mb) 2.29. BREC supports the high variation of the pericentromeric regions between the

three chromosomes. Also, for the first time, using BREC we were able to define the physical location of the telomeric regions of this genome.

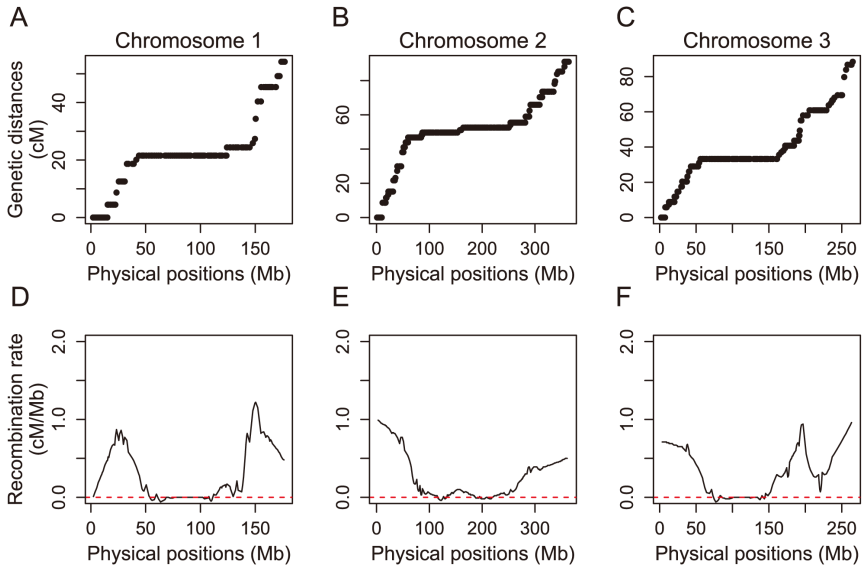


FIGURE 2.29: **The relationship between genetic (cM) and physical map (Mb) positions and estimated local recombination rates across the three chromosomes of a previous version of the *Ae. aegypti* genome.** The physical length was measured as the number of base pairs mapped to a particular genetic position for chromosomes A) 1, B) 2, and C) 3. Local recombination rates for chromosomes D) 1, E) 2, and F) 3, estimated using the Loess function with the MareyMap R package (Rezvoy et al., 2007), show depressed recombination in the centromeric regions of each chromosome. [from (Juneja et al., 2014)]

Taking together, These preliminary results are encouraging as they suggest accurate chromatin boundaries definitions using BREC. To test if we can obtain the same quality of results in other species, we launched BREC on *Cx. pipiens* genome (see Appendix C).

2.6 Discussion and Conclusion

The main two results of BREC are the eu-heterochromatin boundaries and the local recombination rate estimates (see Figures 2.26 and 2.20).
The HCB algorithm, which identifies the location of centromeric and telomeric regions on the physical map, relies on the regression model obtained from the correlation between the physical distance and the genetic distance of each marker. Then, the goodness-of-fit measure, the R-squared, is used to obtain a curve upon which the transition between euchromatin and heterochromatin is detectable.
On the other hand, the recombination rate algorithm, which estimates local recombination rates, returns the first derivative of the previous regression model as the recombination rates, then resets the derivative values to zero along the heterochromatin regions identified (see Figure 2.9).

We validated BREC methods with a reference dataset known to be of high quality: *D. melanogaster*. While two distinct approaches were respectively implemented for

the detection of telomeric and centromeric regions, our results show a similar high resolution (see Table 2.4 and Figure 2.20). Then we analysed BREC's robustness using simulations of a progressive data degradation (see Figures 2.22 and 2.23). Even if BREC is sensitive to the markers' distribution and thus to the local markers' density, it can correctly handle a low global markers' density. For the *D. melanogaster* genome, a density of 5 markers/Mb seems to be sufficient to detect the HCB accurately.

We also validated BREC using the domesticated tomato *S. lycopersicum* dataset (see Table 2.5 and Figure 2.21). At first glance, one might ask: why validating with this species when the results do not seem really congruent? In fact, we have decided to investigate this genome as it provides a more insightful understanding of the data-driven aspect of BREC and how data quality strongly impacts the heterochromatin identification algorithm. Variations in the local density of markers in this genome are particularly associated with the relatively large plateaued centromeric region representing more than 50% of the chromosome's length. Such *data scenario* is quite different from what we previously reported on the *D. melanogaster* chromosomal arms. This is partially the reason for which we chose this genome for testing BREC limits.

While analyzing the experiments more closely, we found that BREC processes some of the chromosomes as presenting a centromeric gap, while that is not actually the case. Thus, we forced the HCB algorithm to automatically apply the *with-no-centromeric-gap-algorithm*, then, we were inspired to implement this option into the GUI in order to give the users the ability to take advantage of their *a priori* knowledge and by consequence to use BREC more efficiently. Meanwhile, we are considering how to make BREC completely automated regarding this point for an updated version later on. Besides, the reference heterochromatin results we used for the BREC validation are rather an approximate than an exact indicator. The physical positions used as reference correspond to the first and last markers tagged as "heterochromatin" on the spreadsheet file published by the Tomato Genome Consortium authors in (Sato et al., 2012). However, we hesitated before validating BREC results with these approximate reference values due to the redundant existence of markers tagged as "euchromatin" directly before or after these reference positions. Unfortunately, we were unable to validate telomeric regions since the reference values were not available. As a result, we are convinced that BREC is approximating well enough in the face of all the disrupting factors mentioned above.

On the other hand, this method's ambition is to escape species-dependence, which means it is conceived to apply to a various range of genomes. To test that, we also launched BREC on genomic data from different species (the house mouse's chromosome 4, roundworm's chromosome 3, and the chromosome 1 of zebrafish). Experiments on these whole genomes showed that BREC works as expected and identifies chromosome types in 95% of cases (see Figure 2.12).

One can assume, with the exponential increase of genomic resources associated with the revolution of the sequencing technologies, that more fine-scale genetic maps will be available. Therefore, BREC has quite the potential to widen the horizon of deployment of data science in the service of genome biology and evolution. It will be crucial to develop a dedicated database to store all this data.

BREC package and design offer numerous advantageous functionalities (see Table 2.6) compared to similar existing tools (Siberchicot et al., 2017; Fiston-Lavier et al., 2010). Thus, we believe our new computational solution will allow a large set of

scientific questions, such as the ones raised by the authors of (Lenormand et al., 2016; Stapley et al., 2017), to be addressed more confidently, considering model as well as non-model organisms, and with various perspectives.

TABLE 2.6: **Comparing BREC with similar widely used tools.** BREC's provided features and functionalities are compared along with the Recombination Rate Calculator (RRC) (Fiston-Lavier et al., 2010) and the MareyMapOnline (Siberchicot et al., 2017), following a chronological order (the oldest first).

Features / Tool		**RRC**	**MMO**	**BREC**
Publication year		2010	2017	2020
Genome-specific		D. melanogaster	all	all
Interpolation method	**Polynomial**	yes	yes	no
	Loess	no	yes	yes
	Cubic spline	no	yes	no
Data cleaning		no	manually	automatically
Data Quality Control		no	no	yes
Chromatin boundaries identification		no	no	yes
Software	R package	no	yes	yes
	Web-based GUI	Perl CGI	Shiny	Shiny

Chapter 3

Improving the quality of genome assembly using repeats

Contents

3.1 Context and motivation 81
3.1.1 Genome assembly overview 82
3.1.2 *de novo* assembly approaches 87
3.1.3 Genome scaffolding 94
3.1.4 Correcting short read assembly errors 97
3.2 New approach: From classic to enhanced scaffolding 98
3.2.1 Method description 98
3.2.2 Data simulated 101
3.3 Results 101
3.3.1 Impact on the assembly quality 101
3.3.2 RR within the misassemblies 103
3.4 Discussion and conclusion 105

Conducting studies to further the understanding of genome dynamics requires the availability, accessibility, and mostly the completeness of the genome sequence of interest, thus, the high quality of its assembly.

Although a new genome sequence of the *Cx. pipiens* (*CpipJ3*) has been recently released (Dudchenko et al., 2017), our preliminary analysis highlighted a serious amount of assembly errors (chimeric, repeat collapse, ...). The improvement of the *Cx. pipiens* genome is currently in progress within our team, through the re-sequencing of this species, yet, we have been interested in investigating the possibility of making such improvement of genome quality achievable by means of optimizing the scaffolding process, instead of re-sequencing.

3.1 Context and motivation

Looking back at when it all begun, the first milestone for genome assembly goes back to the 1960s when the small genomes of yeast and *E.coli* have just been sequenced. Since then, enormous progress has been achieved, and the amount of genomic sequences produced has been regularly increasing tenfold. Figure 3.1 shows the evolution of both sequencing technologies and volume of genomic data. We can notice that most of reference genomes including mosquito genomes like *D. melanogaster*

and *An. gambiae* have been produced during the Sanger sequencing technology era. After a decade of Next-Generation Sequencing (NGS) and expansion of the range of sequenced organisms, the field is currently undergoing the era of Third Generation Sequencing (TGS) and population scale sequencing.

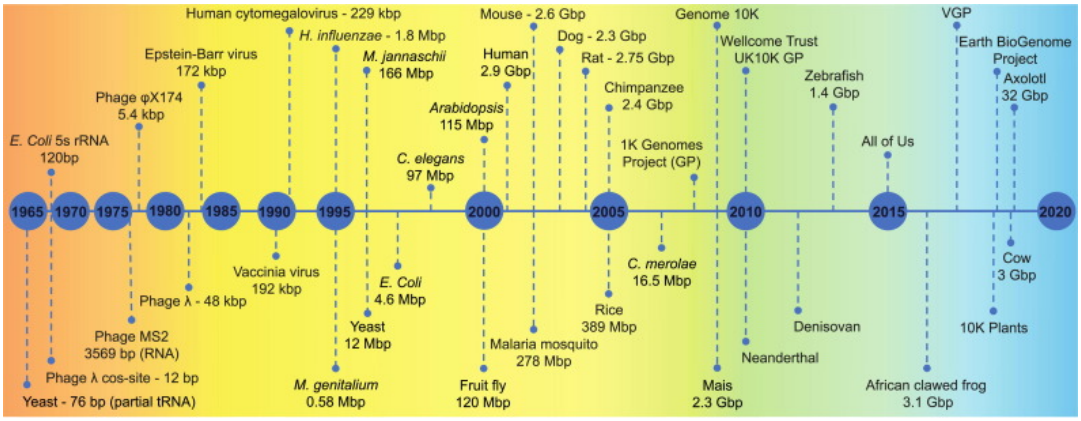


FIGURE 3.1: **Milestones in genome assembly.** Timeline illustrating many of the major genome assembly achievements ranging from the beginning of the sequencing era to the large-scale genome projects currently ongoing. Each genome or genome project (GP) is placed under a color-coded background according to the sequencing approach adopted. Light red: early sequencing methods, Yellow: Sanger-based shotgun sequencing, Green: NGS, Light blue: TGS (Third Generation Sequencing). [from (Giani et al., 2020)]

Due to the characteristics of these sequencing technologies, it is extremely rare to get the correct sequence of a whole genome directly from the sequencing data. When using NGS in particular, reads produced are short in length, only a few hundreds of bp. On the other hand, long reads produced by TGS reach a more interesting scale for inferring global information, e.g. thousands of bp. Yet, this is still not sufficient to get one unique complete genome sequence per chromosome. Therefore, sequencing data need to be computationally *assembled* into larger DNA fragments.

In this chapter, we focus on the way of producing *de novo* genomes, and what could be done to achieve chromosome scale sequences with better quality, using already available datasets. Thus, we (1) describe the assembly process yielding contigs and its potential weaknesses, (2) focus on the scaffolding step which is post-processing of the assembled sequences, and (3) highlight the role of repeated elements throughout these steps. Then, we describe a new approach to take into account the existing repeats between the contigs and the scaffolds, in order to improve the assembled genome.

3.1.1 Genome assembly overview

Genome assembly: Reconstructing the reference genome

Genome assembly is the computational process of reconstructing a genome, as complete as possible, based on the fragmented DNA sequences produced by the sequencers. Sequencers are the machines used for reading the genetic material of an organism, and converting it into a data file ready to be analysed with a computer

for multiple purposes. Figure 3.2 shows an overview of the genome assembly process which aims to produce the reference genome based on a set of sequencing data (Ghurye and Pop, 2019).

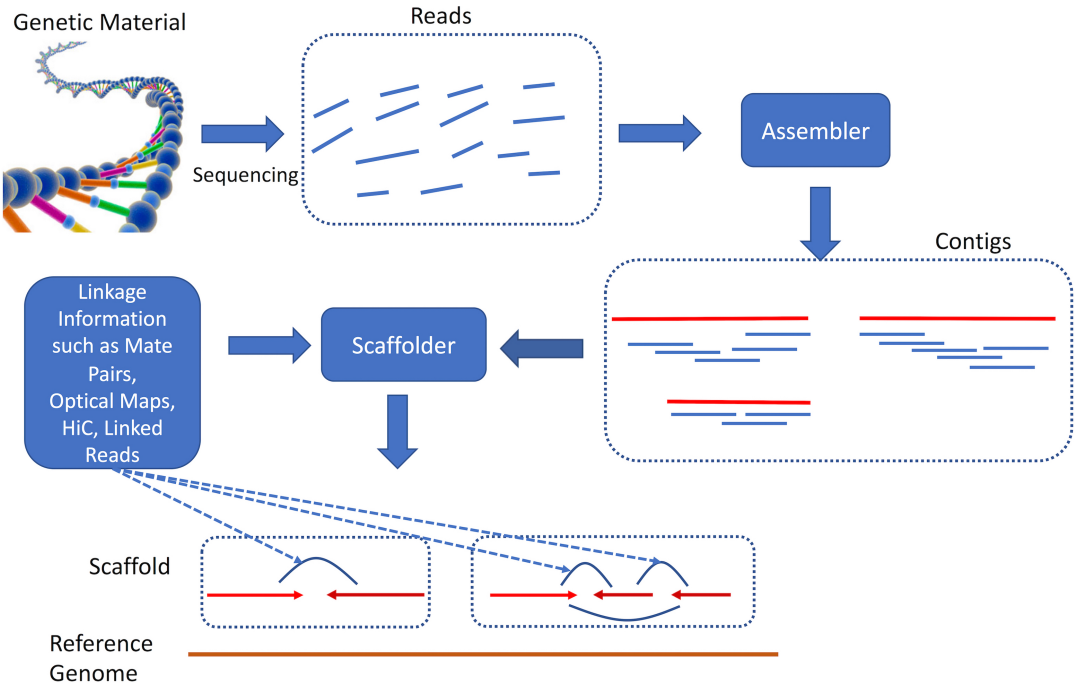


FIGURE 3.2: **Overview of the genome assembly process.** First, genetic material is sequenced, generating a collection of sequenced fragments (reads). These reads are processed by a computer program called an *assembler*, which merges the reads based on their overlap to construct larger contigs. Contigs are then oriented and ordered with respect to each other with a computer program called a *scaffolder*, relying on a variety of sources of linkage information. The scaffolds provide information about the long-range structure of the genome without specifying the actual DNA sequence within the gaps between contigs. The size of the gaps can also only be approximately estimated. (contig, contiguous genomic segment). [from (Ghurye and Pop, 2019)]

To have a closer look at the assembly process, Figure 3.3 from (Sohn and Nam, 2018) details the general steps of an assembly workflow that may be applied to most genomes. The quality of the assembly is strongly impacted by not only the assembler's algorithms, but mainly by the starting quality of the reads. Therefore, making the right choice about the sequencing technology to rely on is of great importance.

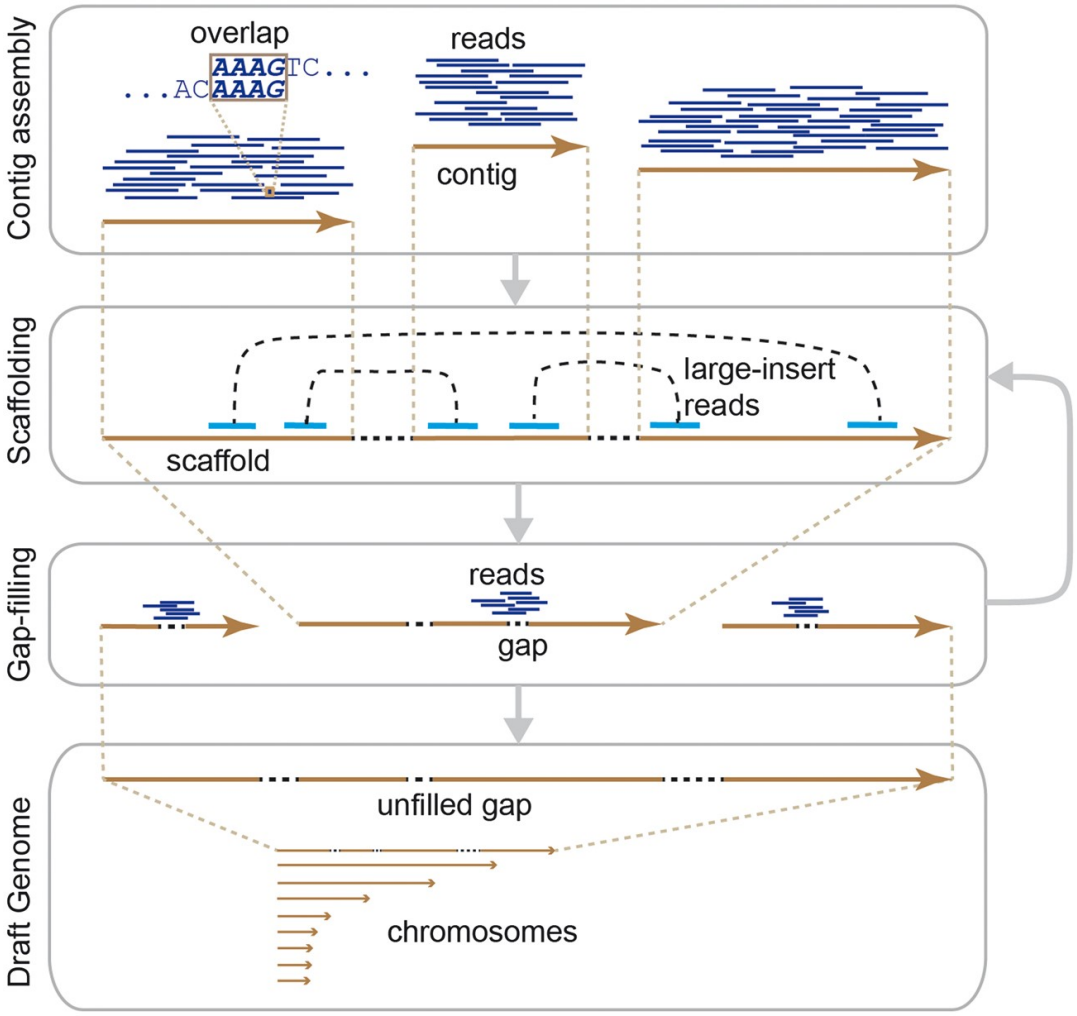


FIGURE 3.3: **General workflow of the de novo assembly of a whole genome** By overlapping reads, contigs are assembled from short reads before scaffolding by large-insert reads, and the remaining gaps are filled. The scaffolding and gap-filling steps can be iteratively performed until no contigs are scaffolded or no additional gaps are resolved before completion. Through this procedure, a draft genome consisting of chromosomes is built. Some unfilled gaps may remain in the draft genome. [from (Sohn and Nam, 2018)]

Sequencing data: the inputs for genome assembly approaches

In order to comprehend the scales variation between the different sequencing data currently available, Figure 3.4 by (Peona, Weissensteiner, and Suh, 2018) presents real-life inspired examples. This figure highlights the fact that dealing with short reads and long reads may be totally different.

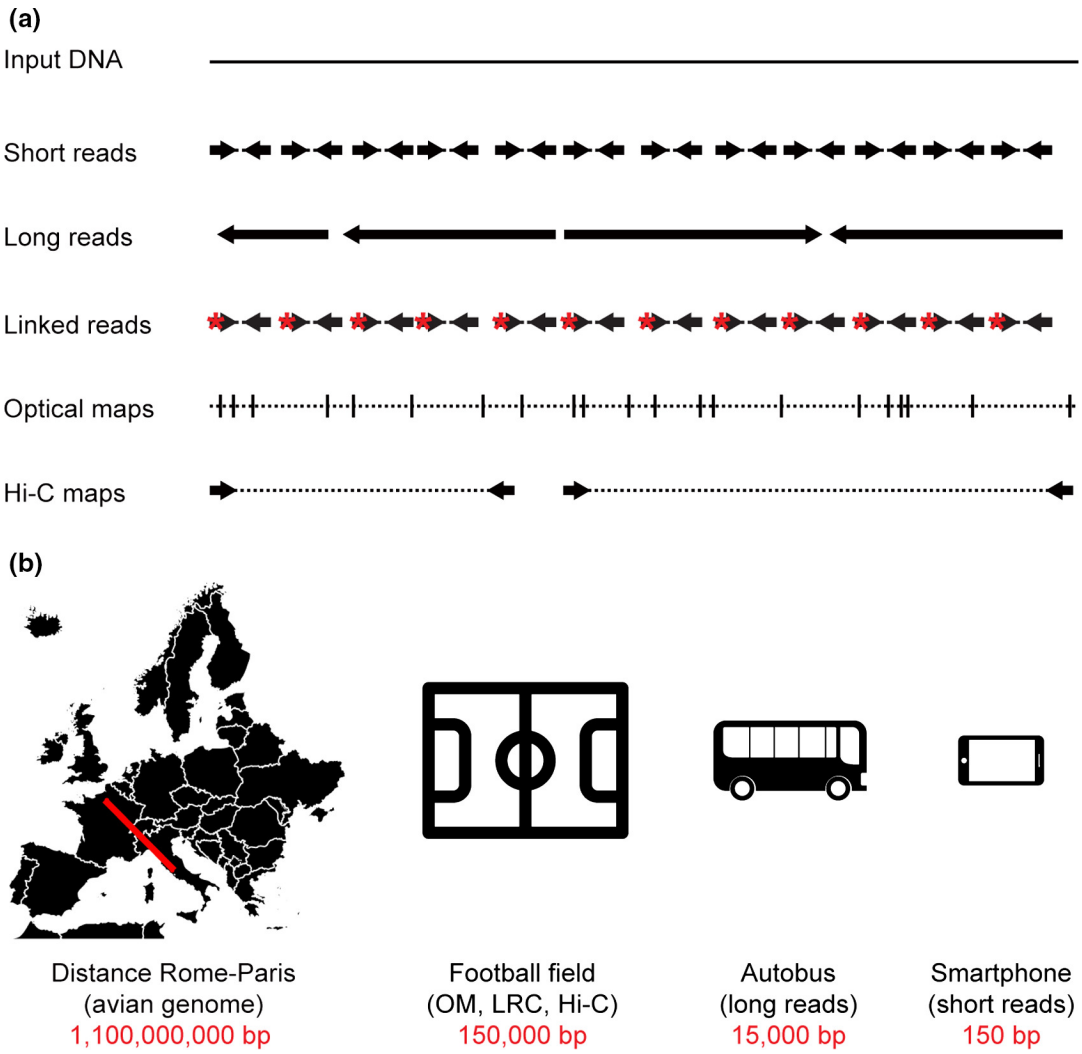


FIGURE 3.4: *Currently available genomics technologies.* **(a)** Schematic illustration of the data structure of these technologies produced from a hypothetical input DNA molecule. Short reads come in read pairs, long reads as single reads, linked-read clouds (LRC) as short-read pairs with a unique barcode (red asterisk) for each input molecule. Optical maps (OM) contain physical distances between short sequence motifs, and Hi-C maps are short-read pairs of 3D genome interactions obtained through chromatin conformation capture. **(b)** Schematic size relations of the data structure from panel (a). Examples are scaled by illustrating 1 base pair as 1 mm. (Icons made by Freepik from www.flaticon.com). [from (Peona, Weissensteiner, and Suh, 2018)]

For simplification purposes, we chose to only introduce the short reads and long reads from a comparative point of view (Murigneux et al., 2020), especially because we will be focusing on the use of short reads in further sections of this chapter. Here, we focus on the main three used technologies:

- **Short reads**: Illumina is the leader of short-read sequencing technology, and Illumina data constitute the vast majority of genomic data stored in public databases. It represents about 80% of the sequencing market share[1].

[1] https://frontlinegenomics.com/how-did-illumina-monopolize-the-sequencing-market/

- **Long reads**: Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT) have been increasingly exploited, and more frequently combined with short-read data.

The read length is obviously not the only feature distinguishing the sequencing products mentioned above. There is also the quality of reads which depends on the error rate, as well as the cost covering the whole sequencing process. Table 3.1 presents an example of a brief comparison of such features between the Illumina, PacBio SMRT, and Oxford Nanopore MinION technologies (Bansal and Boucher, 2019).

Technology	Read Length	Error Rate (%)	Estimated cost per Gb (US$)
Illumina	100–300 bp	0.1	40-60
PacBio SMRT	10–100 kb	5–15	300–900
ONT MinION	Variable (up to 1,000 kb)	5–20	50-500

TABLE 3.1: Comparison of the read lengths, error rates, and costs of various DNA Sequencing Technologies. [table from (Bansal and Boucher, 2019), costs from (Logsdon, Vollger, and Eichler, 2020)]

To what extent the current reference genomes are complete?

Model organisms across eukaryotic genomes, such as *Arabidopsis thaliana* (plant), *D. melanogaster*, and *H. sapiens* (human), have always been at the heart of most studies either in the genetics or genomics fields. Thus, the community is continuously in need of the whole genome sequence for these species in addition to others, in order to increase the accessibility to the hidden messages of their DNA.

Surprisingly, despite the enormous advances in terms of performance achieved by the sequencers as well as the assemblers, reaching the optimum goal of 100% fully assembled genome is still a dream. Table 3.2 by (Peona, Weissensteiner, and Suh, 2018) highlights this point by presenting the current state of the three genomes mentioned above.

Name	Chrs (n)[a]	Scaffolds	Expected size (Gb)[a]	Assembly size (Gb)	Missing (Mb)[b]	"N" gaps (Mb)[c]	% missing DNA[d]
***A. thaliana* [TAIR10]**	5	7	0.125	0.12	5.33	0.20	4.4
***D. melanogaster* [dm6]**	4	1,870	0.17	0.14	30.00	1.10	18.0
***H. sapiens*[hg38]**	23	594	3.42	3.25	162.00	161.00	10.3

TABLE 3.2: **Quantification of missing DNA in the reference genomes of three model organisms.** [from (Peona, Weissensteiner, and Suh, 2018)] ***Notes*** | **n:** Haploid chromosome number. | Weblinks to sampled genome assemblies are listed in Supporting Information Data S1 of the original reference. | [a] Chromosome number and genome size estimates. Genome size estimates were converted from C-values into billion basepairs (Gb) assuming 1 pg = 0.978 Gb (see original reference for data sources). | [b] Assembly size subtracted from expected genome size. | [c] Sum of all "N" nucleotides present in the genome assembly. | [d] Percentage of the expected genome size either missing in the assembly or assembled as "N" nucleotides.

In addition to the fragmentation of the genome and the missing DNA, existing genomes qualified as complete may also present some errors and imprecise parts, like sequences composed of the generic nucleotide "N" meaning "any nucleotide". Also, some errors related to structural mistakes, occurred during the assembly, lead to so-called "misassemblies" which will be discussed hereafter. Some of them are unexpected insertion, deletion or genome rearrangements identified when compared to the reference genome.

3.1.2 *de novo* assembly approaches

The name "*de novo*" means that the assembly process is going to reconstruct the genome sequence from scratch, and not based on a reference genome, which is called "reference assembly" (also known as "mapping assembly").

de novo assembly allows to produce the genome of a newly sequenced organism, or to preserve the genetic diversity of already existing reference genomes instead of losing specific characterising motifs that may exist in a new version of a genome but not in its assembled reference (Mukherjee et al., 2019).

In the literature, *de novo* assembly approaches are mainly based on 3 paradigms: Greedy algorithms, Overlap-Layer-Consensus (OLC), and De Bruijn graphs (DBG) (Nagarajan and Pop, 2013). The generic problem of assembly is often stated as the *Shortest Superstring Problem* (SSP): "from a set of strings, find the shortest string that contains them as factors". The modeling of the assembly problem does not take into account the fact that some repeats may appear in the result. Therefore, it is not surprising that such repeats are not well handled by the existing methods addressing such problem.

Greedy algorithms

The idea underlying the greedy algorithm is to greedily merge reads that "best" overlap, where the optimality criteria is the length of the overlap. The methods based on such algorithm are simple and quite easy to apply, but the memory consumption is absolutely crippling when it comes to assemble genomes of average size. Thus, this idea was mainly exploited at the beginning of the sequencing era, and quite abandoned after that for application on NGS data.

Overlap-Layer-Consensus

Figure 3.5 from (Bleidorn, 2017) illustrates the Overlap-Layout-Consensus assembly algorithm which consists of 3 phases as follows:

1. **Overlap:** compute overlaps between reads and infer a (directed) *overlap graph* defined as follows: vertices are sequences and edges represent their overlap, labeled by the length of this overlap.
2. **Layout:** Find an optimal path in the overlap graph, through a Hamiltonian-like process. We remind that the Hamiltonian Path Problem is *NP-hard*, thus this step is very time-consuming, except when using heuristics.

3. **Consensus:** From the previous path and together with the read sequence information, infer a consensus sequence as the final result.

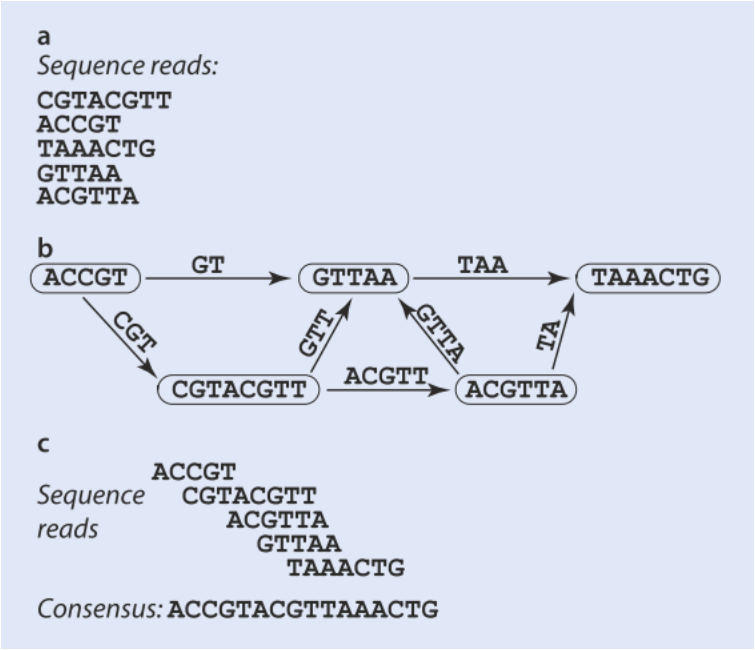


FIGURE 3.5: **Overlap–layout–consensus genome assembly algorithm: Reads are provided to the algorithm. (a)** Sequence reads for assembly. **(b)** Overlap graph. **(c)** Alignment of reads after layout step, in which a Hamiltonian path was searched for in the overlap graph. The consensus sequence is the resulting contig. [from (Bleidorn, 2017)]

OLC strategy is applicable on relatively modest datasets, thus it is not used for large organisms sequenced with short reads. However, they regained popularity with the TGS era, since long reads datasets are smaller.

De Bruijn graphs

A *De Bruijn Graph* of strings of size k on a given alphabet, is a graph defined by: (1) the set of vertices is the set of all existing strings of size k on this alphabet, (2) the set of edges so that there is an edge between u and v when u and v present an overlap of size $k-1$.

DBG-based methods became popular when the amount of data overwhelmed the ability of other methods, especially when the sequencing depth (i.e. the average number of times a genomic position is read by the sequencer) increased. Users must find a way to store the overlapping information without memory and time-consuming redundancy handling, and a way to do so is to consider reads as sets of k-mers, which are its factors of length k. Using then a k-mer graph instead of an overlap graph reduces the problem and allows to deploy several efficient storing strategies.

Figure 3.6 by (Compeau, Pevzner, and Tesler, 2011) shows an example of the *k*-mer graphs that may be built from a sequencing dataset and the way to solve the assembly problem in such graphs. Though moving from an *NP-hard* problem, the Hamiltonian Cycle Problem, to a polynomial one, the Eulerian Cycle Problem, this kind of methods is not magically solving all the issues. First, the choice of the *k* value is crucial for both: (1) the size of the data-structure to store *k*-mers issued from reads, since larger are the *k*-mers, the more they can be, (2) and the handling of repeats, since the smaller are the *k*-mers, the less precision we get.

Mostly used short-read assembly tools use one or more *k*-mer graphs and the traversals of these graphs, which we will be focusing on in the following. Though the *k*-mer graph does not contain *every* possible *k*-mer but only those which are present in the reads, we will undifferentially use the terms *k*-mer graph and De Bruijn graph hereafter, since this is the common usage by the community.

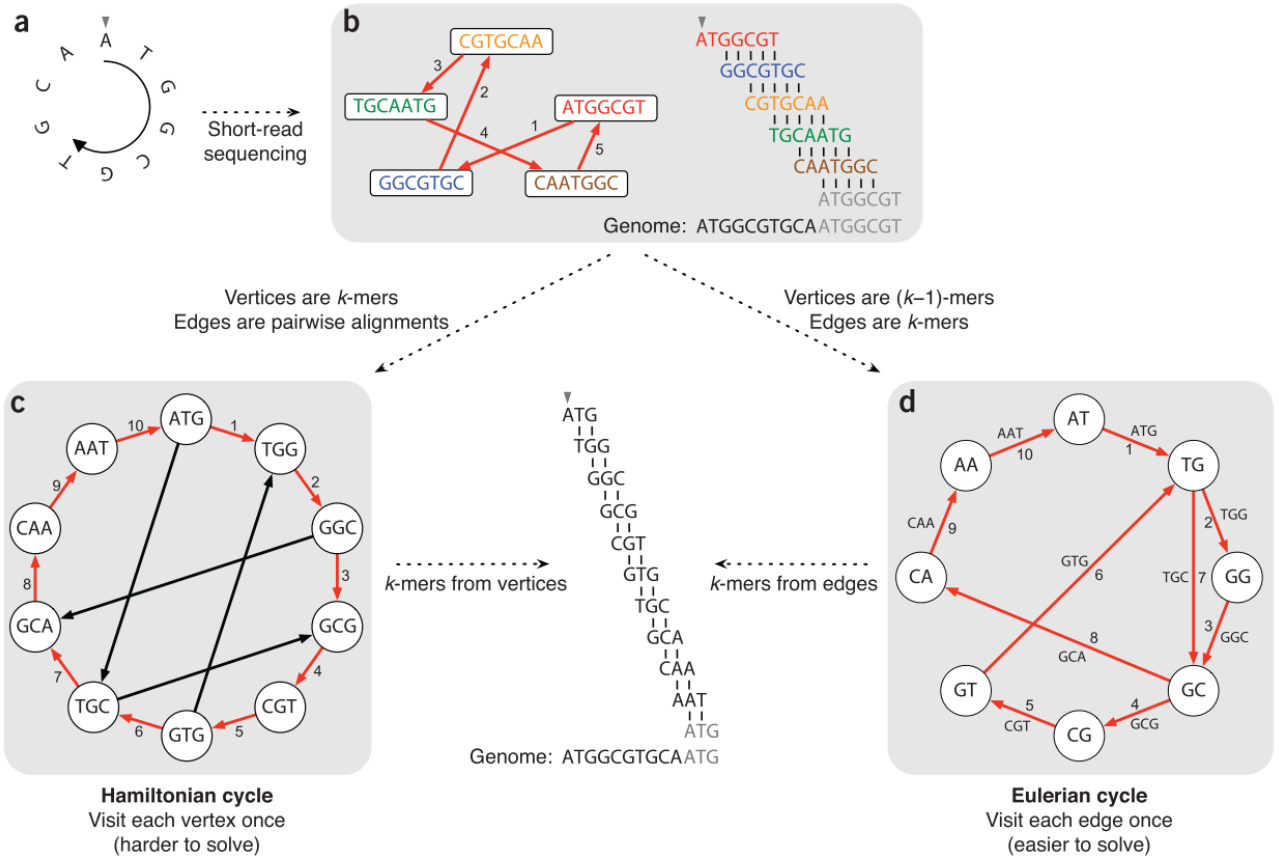


FIGURE 3.6: **Two strategies for genome assembly: from Hamiltonian cycles to Eulerian cycles.** **(a)** An example small circular genome. **(b)** In traditional Sanger sequencing algorithms, reads were represented as nodes in a graph, and edges represented alignments between reads. Walking along a Hamiltonian cycle by following the edges in numerical order allows one to reconstruct the circular genome by combining alignments between successive reads. At the end of the cycle, the sequence wraps around to the start of the genome. The repeated part of the sequence is grayed out in the alignment diagram. **(c)** An alternative assembly technique first splits reads into all possible k-mers: with k = 3, ATGGCGT comprises ATG, TGG, GGC, GCG and CGT. Following a Hamiltonian cycle (indicated by red edges) allows one to reconstruct the genome by forming an alignment in which each successive k-mer (from successive nodes) is shifted by one position. This procedure recovers the genome but does not scale well to large graphs. **(d)** modern short-read assembly algorithms construct a de Bruijn graph by representing all k-mer prefixes and suffixes as nodes and then drawing edges that represent k-mers having a particular prefix and suffix. For example, the k-mer edge ATG has prefix AT and suffix TG. Finding an Eulerian cycle allows one to reconstruct the genome by forming an alignment in which each successive k-mer (from successive edges) is shifted by one position. This generates the same cyclic genome sequence without performing the computationally expensive task of finding a Hamiltonian cycle. [from (Compeau, Pevzner, and Tesler, 2011)]

Repeats: a big challenge facing the assembly process

Due to their repetitive nature, repeats present in the genome yield strong disruption in the assembly process. This occurs quasi-systematically when reads are too small to entirely represent the repeated sequence and its copies. Figure 3.7 from (Bleidorn, 2017) presents an example of a wrong assembly result, due to a repeated sequence. Thus, repeats can lead to **erroneous overlaps**.

FIGURE 3.7: **Example of a wrong assembly of a repetitive region.** The repeat motive is given in red, a stretch of the true sequence which is missing in the resulting assembly is given in blue [from (Bleidorn, 2017)]

To more precisely analyse the impact of repeats on *k*-mer graphs, Figure 3.8 from (Bleidorn, 2017) shows an example of an anomaly causing ambiguous choices in the graph. Repetitive sequences can produce loops in the DBG, which make it difficult for the path search to be resolved.

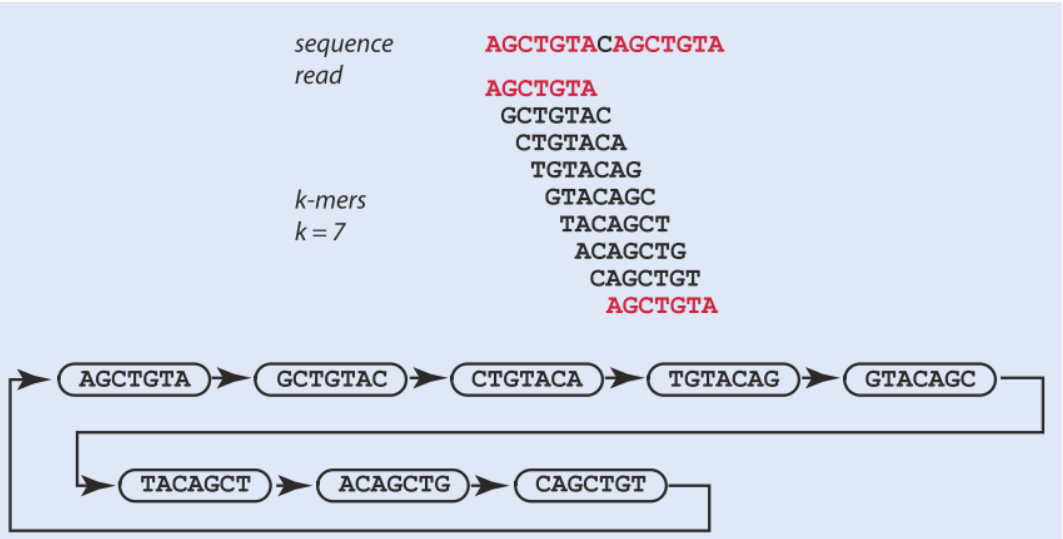


FIGURE 3.8: **Repetitive sequences can lead to loops in a de Bruijn graph.** The repetitive motive is indicated in red. [from (Bleidorn, 2017)]

Figure 3.9 from (Bleidorn, 2017) shows other possible complicated structures that may occur in the graph when repeats are around.

To handle such messy subgraphs involving repeats, decisions must be made during the traversal of the graph. Figure 3.10 from (Wajid and Serpedin, 2016) gives examples of such decisions, which are often subject to arbitrary parameters, and lead to a fragmented set of contigs.

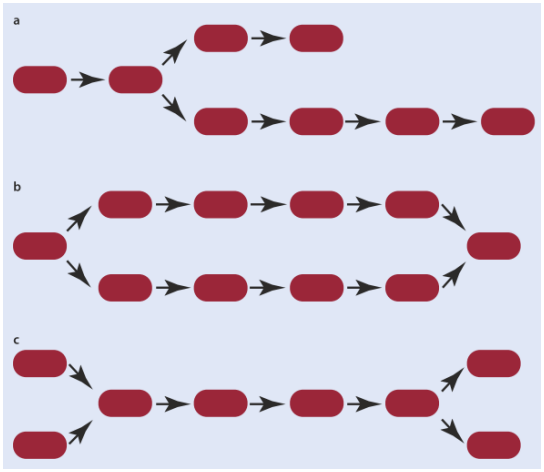


FIGURE 3.9: **Sequence errors and repeats lead to more complex k-mer graphs.** Nodes representing k-mers are indicated by red boxes. **(a)** Errors at the end of sequence introduce dead ends into the graph. **(b)** Errors in the middle of sequences introduce bubbles into the graph. **(c)** Repeat sequences lead to a pattern of convergent and divergent paths [from (Bleidorn, 2017)]

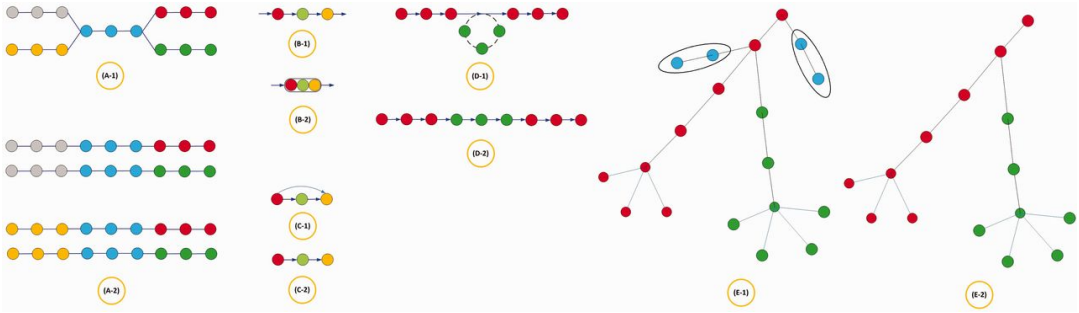


FIGURE 3.10: **Graph simplification techniques: (A-1)** Ambiguous paths; **(A-2)** Pulling apart operation: the resultant graph is divided into four possible paths. **(B-1)** Simplistic path; **(B-2)** Removing intermediate nodes: nodes that have an indegree = outdegree = 1 are collapsed to form one giant node, also referred to as a 'unitig'. **(C-1)** Unnecessary edges; **(C-2)** Removing edges: an edge between two nodes is removed if there is an intermediate node between them that connects them simplistically. **(D-1**) Loop; **(D-2)** Disambiguation: the loop edge is unrolled and integrated in the continuous edge from left to right. **(E-1)** Shorter paths are shown encircled; **(E-2)** Removing tips: a tip is defined as a chain of nodes that is disconnected at one end. Tips are removed if they are shorter than t, where t is a user-defined parameter. Furthermore, if there is a longer/common path, it will also trigger a tip's removal. [from (Wajid and Serpedin, 2016)]

In this paragraph, we focused on the DBG approaches, but it is important to bring the reader's attention to the fact that other methods are not better-armed to face this problem. By nature, the greedy approach will expurgate repeated regions from the solution. In OLC graphs, repeats may be represented with more precision, but there are still ambiguities that have to be solved (see Figure 3.11 from (Li et al., 2012)).

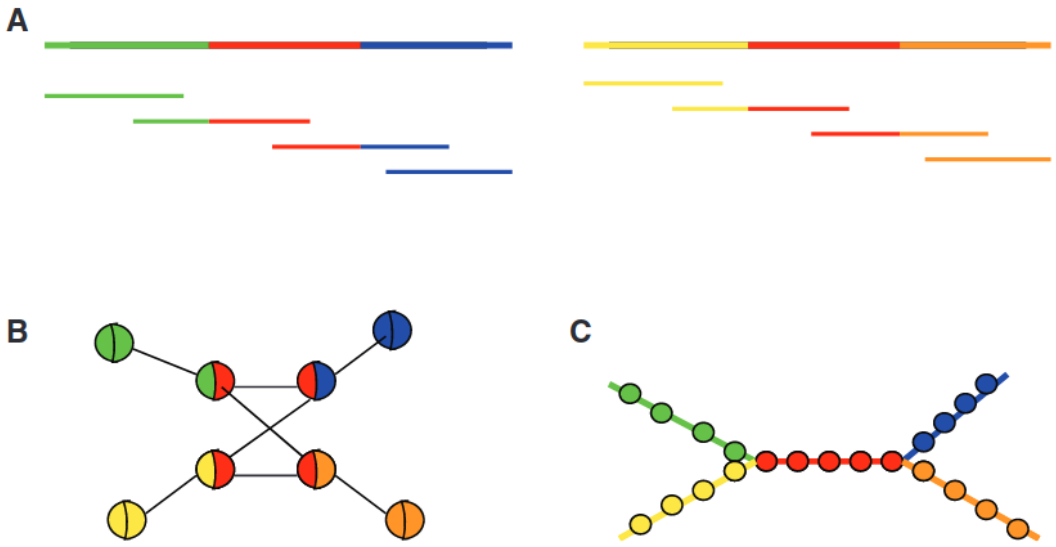


FIGURE 3.11: **The difference to represent repeats in OLC and DBG graphs. (A)** Two separate genomic regions share a repeat fragment (in the middle) and the flanking regions are unique sequences. Top is the genomic sequence and bottom are the sequenced reads. **(B)** The OLC reads graph. The nodes represents reads and the links show overlap relations. All the repeat reads are placed on the graph as nodes. **(C)** The *k*-mer graph. The reads are chopped into shorter *k*-mers. The *k*-mers from repeat regions are collapsed together. [from (Li et al., 2012)]

Assemblers

Based on the assembly approaches previously mentioned, the literature counts a large variety of software tools that implement various algorithms to handle the genome assembly process.

Overall, there are 3 categories of computational resources for genome assembly, also called assemblers, as per the type of the sequencing data used:

- **Short read assemblers**: such as Velvet (Zerbino and Birney, 2008), Abyss (Jackman et al., 2017), AllPath (Butler et al., 2008), SPAdes (Bankevich et al., 2012), and Minia (Chikhi and Rizk, 2013).
- **Long read assemblers**: such as Canu (Nurk et al., 2020) and Flye (Freire, Ladra, and Parama, 2021). A review of assembly tools for long reads is published by (Wee et al., 2019).
- **Hybrid assemblers**: which combine the short reads and long reads in order to optimise the assembly output, such as HybridSPAdes (Antipov et al., 2016), and MaSuRCA (Zimin et al., 2013).

It is quite interesting to consider diving into this direction in order to better choose which tool for which project upon the research questions addressed (Jung et al., 2020). Nevertheless, such tedious comparison is not within the scope of our work and we believe including this aspect is not of value here.

3.1.3 Genome scaffolding

As mentioned before, repeats are responsible for misassemblies, particularly by fragmenting the assembled sequence into contigs that represent correct parts of the genomes, yet, are quite short compared to the expected sequences. Fortunately, once the contigs are produced, it is still possible to go further towards a chromosome-scale sequence by means of the scaffolding step.

Assembling contigs into scaffolds

The scaffolding problem considers a set of contigs and outputs an orientation and an order on the oriented contigs, which should correspond to the orientation and order on the original genome. The oriented and ordered contigs form *scaffolds*. This problem is *NP-hard*. Thus, is it computationally difficult to handle large instances without using heuristics.

Interesting surveys on recent scaffolding methods are available in (Mandric et al., 2016; Rice and Green, 2019; Luo et al., 2021). The underlying idea is to take advantage of additional information which is not considered during the contigs production process. For instance, some methods are based on the use of pairing of reads. Paired-end reads are produced by NGS technologies and correspond to external sequences of one same fragment. Considered individually during the assembly, paired-end reads may provide precious information on the proximity of contigs (see Figure 3.12 from (Ghurye and Pop, 2019)).

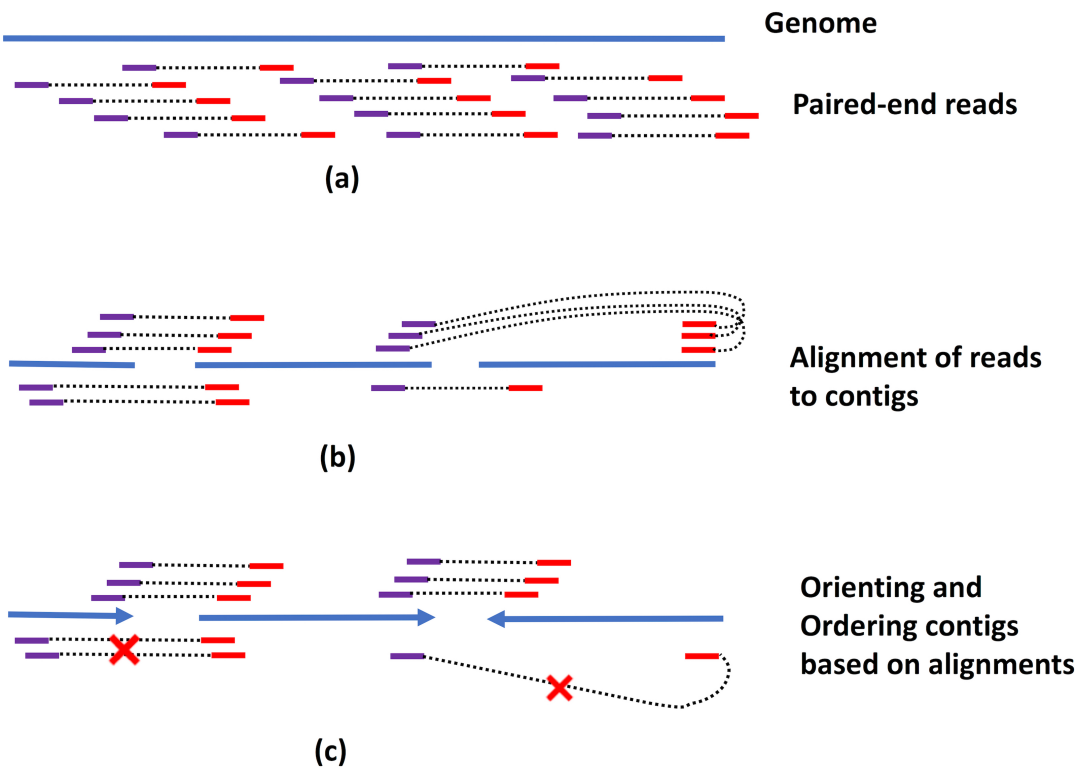


FIGURE 3.12: **Use of pairwise linkage information for scaffolding.** **(a)** Paired-end reads are sequenced from the genome. Depending on the technology, the approximate distance and/or relative orientation of the paired reads may not be known. **(b)** The reads are aligned to contigs. Reads with their ends aligned to two different contigs provide linkage information useful for scaffolding. **(c)** Linkage information is used to orient and order the contigs into scaffolds. Usually not all constraints can be preserved, and algorithms attempt to minimize inconsistencies (marked with X). [from (Ghurye and Pop, 2019)]

On the base pair level, the difference between a contig and a scaffold is mainly distinguished by the presence of "N" strings representing the "Non available" signal of the corresponding sequence, which reflects a scaffolding gap linking two contigs. As it is shown in Figure 3.13, there is also a part of such gap that is represented by the absence of any letter, which indicates that we also miss the gap length information.

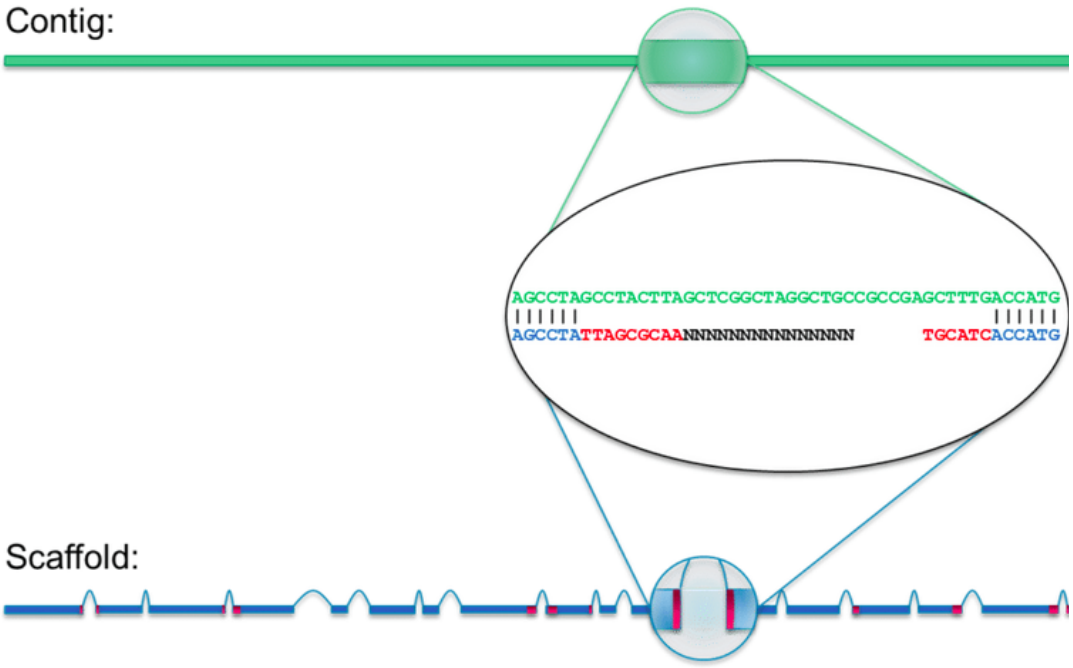


FIGURE 3.13: **Illustration of the difference between contigs and scaffolds in genome assemblies.** Scaffolds are created by chaining contigs together using additional information about the relative position and orientation of the contigs in the genome. Contigs in a scaffold are separated by gaps, which are designated by a variable number of "N" letters. [from https://www.pacb.com/blog/genomes-vs-gennnnes-difference-contigs-scaffolds-genome-assemblies/]

In what follows, since our purpose is mostly to exploit already available sequencing data to improve over-fragmented genomes, we focus on a scaffolding method using paired-end short reads. Formally, it is possible to extract from these information a set of relationships between the contigs, that might be inconsistent.

The *scaffold graph* is defined as follows: vertices represent contig extremities, while edges are of two kinds: (1) contig edges, linking both extremities of a contig, and (2) inter-contig edges relating the pairing-information. A weight function on the inter-contig edges indicates how many pairs are supporting this edge (see Figure 3.14). Due to repeats, some of the inter-contig edges are erroneous and have to be removed from the graph. In other cases, they are supported by the Repeated Regions (RRs).

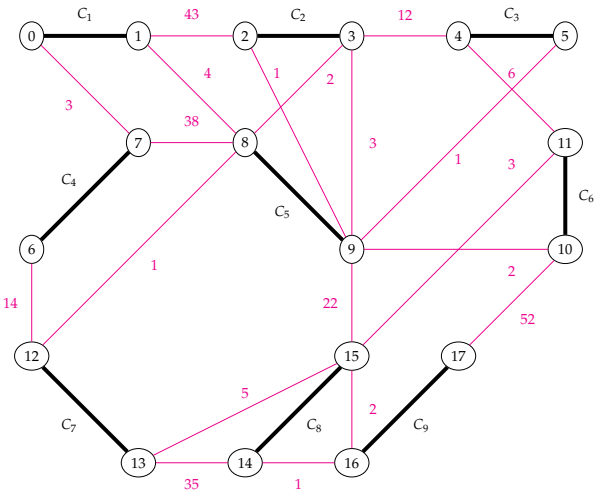


FIGURE 3.14: A scaffold graph with nine contigs (bold edges) and twenty inter-contig edges. Vertices are contig extremities. For instance, contig C_1 is figured by vertices labelled by 0 and 1, the (0,1) direction corresponds to the forward reading of the contig in the assembly file, and the (1,0) direction corresponds to the reverse direction. Inter-contig edges are labelled by the number of pairs of reads connecting one contig extremity to another.

The scaffolding step is also touched by the RR issue. RRs location along contigs, especially when they are near the extremities, can lead to ambiguities at the scaffolding step. Indeed, most scaffolders use a graph structure establishing relationships between contigs sharing a piece of information. This information may come from a set of long reads (if available), or pairs of short reads, one read mapping on the first contig, and the mate mapping on the other contig. Typically, in the latter case, when the reads come from an RR, they may map ambiguously, and a choice has to be made during the processing of the graph. Here we propose, instead of just suffering from their presence, to use RR sequences as an attempt to enhance the scaffolding.

3.1.4 Correcting short read assembly errors

While the efficiency of bioinformatic tools used for assembly is increasing, errors of sequence construction from contiguous short reads persist. One way to untangle ambiguous parts of such assemblies is to use long reads, produced by TGS technologies. However, this is not always possible due to the high cost and high error rate factors.

Recent state-of-the-art on error correction tools targeting Illumina short reads shows that it is possible to enhance De Bruijn Graphs (Heydari et al., 2019), in particular when the correction targets reads near highly repetitive DNA regions (Heydari et al., 2017). However, such correction is proposed between the sequencing step and the assembly step, using analysis on the k-mers. Here, we propose an approach addressing a correction between the contig production step and the scaffolding step.

3.2 New approach: From classic to enhanced scaffolding

In this section, the main question we address is: How to improve the quality of genome assembly using RRs? A secondary question is raised about the *type* of repeats that are the most involved in misassemblies. Here, we focus on the improvement of genomes produced through a *de novo* approach using short reads (improvement of existing assemblies in databases), with a relatively well-defined repeat landscape (repeats documented in the Repbase database). We propose a pipeline progressively refining inter-contig edges through RR analysis.

3.2.1 Method description

We implemented a snakemake (Mölder et al., 2021) pipeline summarized in Figure 3.15. The first four steps aim to produce datasets composed of both a reference genome and a contig set that can be compared to the reference. Further steps are separated in two paths: (1) the first path corresponds to a classic scaffolding with paired-end reads information leading to generation of paired-end scaffolding graph (PE graph), (2) whereas the second path includes repeated regions analysis. The original part of our work lies in the second path, which we describe in detail in Paragraph Repeated Regions analysis.

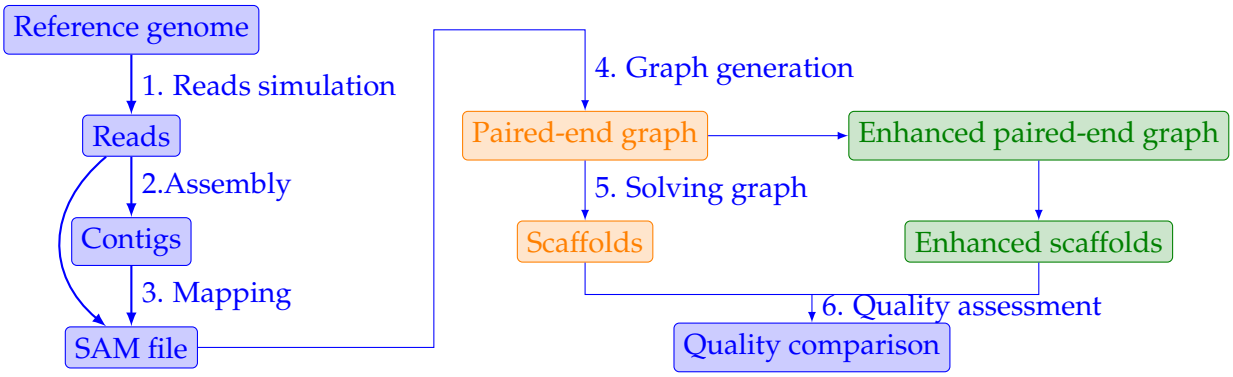


FIGURE 3.15: Overview of the pipeline.

Data production

Simulation We validated our approach on simulated data. The first step was to generate *paired-end reads* as basic data for the assembly and then the scaffolding. To simulate short reads, we chose the ART (Huang, Burns, and Boeke, 2012) software (version 2.5.8;), which produces reads close to the commonly used technologies, and because of its simplicity of use, while allowing a large choice of options.

Assembly We chose to build the contigs with: (1) SPAdes (version 3.13.0 ; `http://cab.spbu.ru/software/spades/` ; (Bankevich et al., 2012)), which is one the mostly used assembly tools and proposes an iterative DBG approach, and (2) Minia (version 3.2.1 ; `https://github.com/GATB/minia` ; (Chikhi and Rizk, 2013)), which is

very light in terms of memory consumption, thanks to its use of Bloom filters. We therefore obtain two separate contig files from different assembly programs which will each be used in all the following steps of the *pipeline* so that we can compare their qualities at the end.

Mapping The next step is to map the *paired-end reads* to the contigs obtained in the previous step. The contigs were mapped on the reference sequences using: (1) Minimap2 (version 2.17 ; `https://github.com/lh3/minimap2` ; (Li, 2018)), and (2) BWA-MEM (version 0.7.17-r1188 ; `https://github.com/lh3/bwa` ; (Li and Durbin, 2009)). Both mapping tools are also famous for their interesting performances and reliability.

The initial protocol used BWA (Li and Durbin, 2009), an alignment tool using "reverse search" (*backward search*) with the Burrows-Wheeler transform. We chose to use BWA-MEM, improvement of BWA, because the latter did not take into account the information in *paired-end reads*. We decided to compare it with Minimap2 for the speed of execution of the latter.

Graph generation Generating *paired-end* scaffold graphs is performed with the Scaftools tool (Weller, Chateau, and Giroudeau, 2015), from the mapping of paired-end reads to the contigs. The graphs generated in each of the four cases (both assembly tools and both mapping tools) will then be passed onto our graph improvement tool.

Repeated Regions analysis

Repeated Region detection The consensus sequences of the repeated regions were obtained from the Repbase Update (RU) database (Bao, Kojima, and Kohany, 2015). RU contains more than 38,000 sequences of different families or subfamilies. The RRs present within the contigs were then detected by aligning the RU consensus sequences using BLAST (McGinnis and Madden, 2004) using megablast default parameters. Consequently, we obtain an alignment file used to label the contigs.

Clustering contigs according to repetitions family Two contigs carrying repetitions of different families can be linked within the PE graph. This link is induced due to the similarities between these RRs, however, it is not coherent with the biological reality. It is therefore necessary to separate the contigs according to the repetitions they carry, in order to limit such incoherent links and, instead, favor them in case of contigs carrying the same RR. The classification and clustering of repetitions can be done at different levels/scales: (1) clusters that are too small would be less informative, (2) while clusters that are too large would make the further processing heavier/more complicated. We performed the clustering at the subfamily level.

Building the RR graph At this stage, each contig is defined by the following values: its name, its length (ℓ), the name of the repetition family carried, the identifier of the original repetition (*repid*), the start bound (*start*), and the end bound (*end*) of the RR on the contig. If one of the bounds is equal to 1 or ℓ, the RR is considered *external*, otherwise it is qualified as *internal*. Within each cluster, the position of the RRs

on each contig is evaluated and then exploited in order to join the contigs carrying the same RR. The purpose of these junctions is to orient the contigs according to the RR information they carry. These information allow, for each cluster, to generate a graph using Graphviz format(Gansner and North, 2000). The set of all these graphs is called the *RRs graph*. The process leading to the RR graph is described in Figure 3.16.

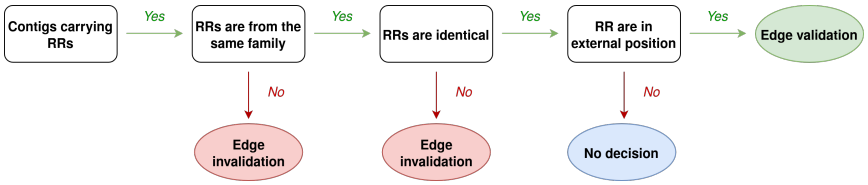


FIGURE 3.16: RRs detection and characterization.

Using RR graphs to correct a PE graph We use the edges from the RR graph to apply corrections to the PE graph. We remind that in both graphs, vertices are contig extremities and edges are links between these extremities. It is obvious that the applied corrections only concern the edges implied as for repetitions. However, we can assume that the edges not affected by RRs are less likely to cause problems because they are not impacted by them. These corrections can be of several types:

- **Edges common in the PE graph and the RR graph**. We are *a priori* assured of the validity of an edge if it is present within both graphs. In this case, we add an additional weight to the weight of the PE edge, to strengthen this edge during the final scaffolding. This weight is relative to the size of the cluster from which the RR comes, with an additional weight of one per hundred elements in the cluster.
- **PE edges between contigs carrying RRs from different families**. In this case, the PE edges are removed from the PE graph, since the similarity yielding this edge has been invalidated by the RR sequences.
- **PE edge with only one contig carrying RRs**. In this case, the validation process depends on the way the RR is mapped on the contig. The invalidation is performed only when the RR should be present on both contigs (see Figure 3.17).

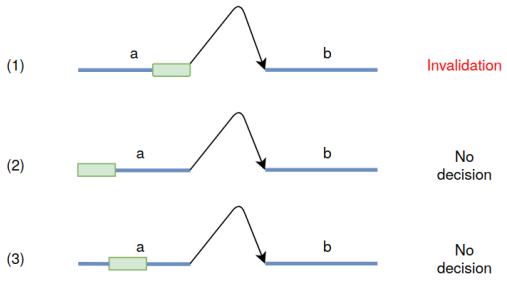


FIGURE 3.17: PE Edge validation for case with only one contig carrying RR. Validation depends on position of RR within the contig.

After the RR analysis: solving and quality analysis

Solving the graphs The resolution of the graphs obtained is also carried out with Scaftools, for the graphs of *paired-end* as well as for the improved graphs. By solving the graph, we mean extracting from the scaffold graph a set of paths of maximum total weight, corresponding to the scaffolds. Knowing that they cause incoherent alignments, the RRs will induce a bias in the scores of inter-contig edges, which will result in poor resolution of the graph. From each original reference genome, we obtain at the end of the *pipeline*, 8 different genomes.

Quality assessment Each assembly was validated with QUAST-LG (version 5.0.2 ; `http://cab.spbu.ru/software/quast-lg./` ; (Gurevich et al., 2013), (Mikheenko et al., 2018)). We expected to get a reduction of *misassemblies* in the tests performed with RRs-corrected PE graphs (PE+RR graph).

3.2.2 Data simulated

We decided to take as reference genomes (1) *D. melanogaster* for the very high quality of its sequenced genome as well as the knowledge of its repeated regions(Hoskins et al., 2015), and (2) *Caenorhabditis elegans* for its small genome, containing little repetitions, and also for its sequencing quality. We simulated sequencing data using *D. melanogaster* and *C. elegans* with the following common specifications:

- Simulated technology: Illumina HiSeq 2000
- Coverage: 20X
- Reads size: 100bp
- Insert size 300bp
- Standard deviation: 10%

3.3 Results

3.3.1 Impact on the assembly quality

For each dataset, the eight scaffold sets produced by the pipeline are compared to the reference using QUAST. We selected the following criteria to analyse the efficiency of the approach: number of contigs (in this case, number of final scaffolds), number of unaligned contigs (scaffolds), percentage of the genome covered by the scaffolds, NG50 (corresponding to the scaffold size such that 50% of the known or estimated genome size are supposed to be of the NG50 length or longer), and the number of misassemblies.

D. melanogaster

Table 3.3 shows the results for the eight genomes produced on the *D. melanogaster* dataset.

D.melanogaster	SPAdes				Minia			
	Minimap2		BWA-MEM		Minimap2		BWA-MEM	
	PE only	PE + RR	PE only	PE + RR	PE only	PE + RR	PE only	PE + RR
Scaffolds	1894	2019	1861	2032	2212	2158	2307	2249
Unaligned scaffolds	8	8	8	**6**	103	**66**	140	**109**
Coverage (%)	83.586	83.147	83.564	83.163	82.691	82.357	82.749	82.42
NG50	138 662	129 502	141 803	133 722	120 493	115 298	115 249	114 878
Nb of *misassemblies*	708	**552**	770	**567**	159	164	252	261
Improvement rate %	**22.03**		**26.36**		-3.14		-3.57	

TABLE 3.3: Result on the *D. melanogaster* dataset. Bold figures shows the improvement achieved by the method. The improvement rate on last row is calculated using the number of *misassemblies* (100×(PE only - (PE+RR))/PE only).

C. elegans	SPAdes				Minia			
	Minimap2		BWA-MEM		Minimap2		BWA-MEM	
	PE only	PE + RR	PE only	PE + RR	PE only	PE + RR	PE only	PE + RR
Scaffolds	4266	4597	4244	4541	5236	5286	5230	5301
Unaligned scaffolds	0	0	0	0	1	0	3	4
Coverage (%)	89.686	89.26	89.673	89.325	87.804	87.501	87.825	87.549
NG50	33 576	29 337	33 157	29 884	24 884	24 437	24 864	24 275
Nb of *misassemblies*	1770	1783	1893	1921	981	1009	1258	1305
Improvement rate %	-0.73		-1.48		-2.85		-3.89	

TABLE 3.4: Result on the *C. elegans* dataset (whole genome). The improvement rate on last row is calculated using the number of *misassemblies* (100×(PE only - (PE+RR))/PE only).

For *D. melanogaster*, the results show a slight decrease in the genome's unaligned scaffolds and NG50 coverage (length for which the collection of contigs of this length cover at least half of the reference genome), while an improvement in the number of *misassemblies* up to 26% for SPAdes (but no improvement with Minia). SPAdes provides fewer contigs than Minia, and produces far fewer unaligned contigs. It also provides greater genome coverage. Our hypothesis to explain this difference between both assembly tools is that Minia, due to its decision process to cut nodes with a large in-degree or out-degree in the DBG, may isolate more drastically RRs as contigs, thus RRs could not help connecting them to other contigs. Difference between the use of Minimap2 *vs.* the use of BWA-MEM in the mapping does not appear to be significant.

C. elegans

Table 3.4 shows the results for the eight assemblies produced on the *C. elegans* dataset.

The results are not very positive for the *C. elegans* genome, when applied on the whole genome. *Misassemblies* are more numerous with the application of the method, contrary to the expectations. Improvement rates are negative, but small. Again, results are better with SPAdes than with Minia.

On the contrary, when the method is applied separately on each chromosome, results are far better, as shown in Table 3.5 (only the number of *misassemblies* are reported here, for a better readability), for SPAdes.

C. elegans	SPAdes				Minia			
	Minimap2		BWA-MEM		Minimap2		BWA-MEM	
	PE only	PE + RR	PE only	PE + RR	PE only	PE + RR	PE only	PE + RR
Chr. 1	292	**276**	307	**298**	163	173	225	230
Chr. 2	282	**230**	283	**251**	124	**119**	151	151
Chr. 3	225	**213**	230	**220**	119	120	161	170
Chr. 4	378	**345**	389	**367**	210	**194**	272	**253**
Chr. 5	358	**344**	388	**370**	186	192	243	**238**
Chr. X	233	**193**	248	**212**	125	**119**	153	**148**
Improvement rate (min/mean/max)%	3.91 / **9.84** / 18.44 p-value=0.88%		2.93 / **7.25** / 14.52 p-value=0.56%		-6.13 / **1.05** / 7.62 p-value=68.07%		-5.60 / **0.75** / 6.99 p-value=55.92%	

TABLE 3.5: Number of *misassemblies* on the *C. elegans* dataset, chromosome per chromosome. Results are significantly better with SPAdes, and equivalent with Minia (see p-values).

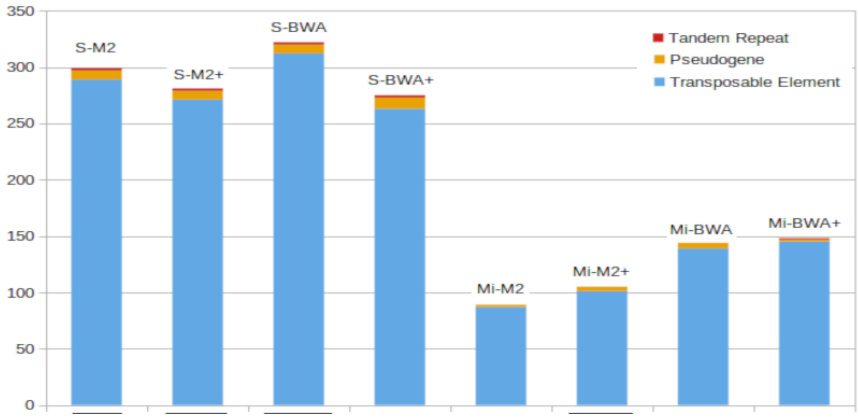


FIGURE 3.18: Analysis of the number of repeats on the extremities of *misassemblies* along the 2R chromosomal arm of *D. melanogaster*. For the assembly: "S" stands for *SPAdes* and "Mi" for *Minia*. For the mapping: "M2" stands for *Minimap2*. For scaffolding graphs, the "+" sign indicates an enhanced graph (PE+RR).

3.3.2 RR within the misassemblies

To analyze the *misassemblies* detected by QUAST, we mapped them on the reference genome. We crossed this mapping with a GFF file of *D. melanogaster* genome, with RRs (tandem repetitions, pseudo-genes and transposable elements), and detected the RRs present at the ends of the *missasemblies*. We have observed that RRs are involved in 60% to 70% of the remaining *misassemblies*. Even if we detect some tandem repetitions and pseudo-genes, the vast majority is composed of transposable elements. We can therefore deduce that transposable elements are the most disturbing for the reconstruction of genomes, because of their numerous specificities (size, activity, age). We performed this analysis on the genome of *D. melanogaster* using the latest available version of its sequenced genome (*release* 6.26), which lists all of the annotated regions known to date. Results of this analysis on the eight scaffoldings, for the 2R chromosomal arm, are presented on Figure 3.18. Results are very similar on other chromosomal arms and chromosomes.

To complete this analysis and find out if one type of TE is particularly involved in the assembly disturbance, we also considered an "historical" approach, and had a look at the the first release of the *Drosophila melanogaster* genome. This previous release is more fragmented, and the gaps are essentially due to repeat-rich regions (Hoskins

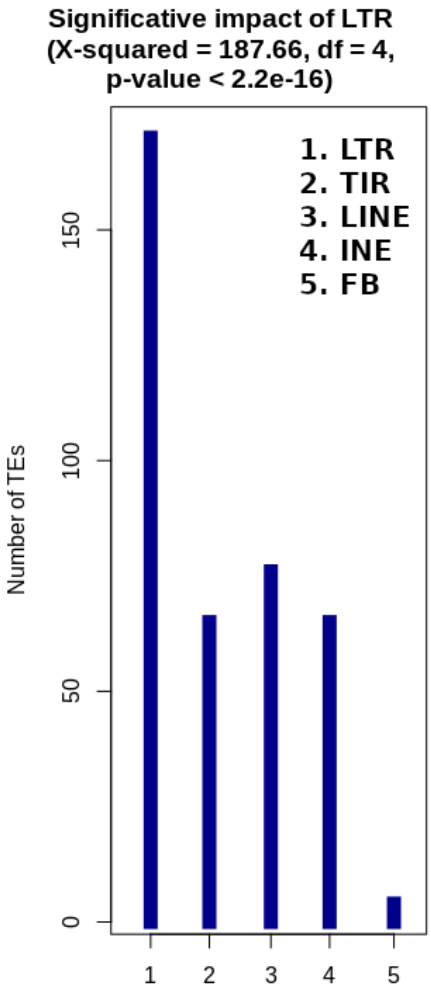


FIGURE 3.19: Number of TEs related to gaps, classified by type for *D. melanogaster R1 vs. R6*

et al., 2015) which necessitate non-trivial techniques to be partially desintricated. We mapped the drosophila known TEs on the gaps constated when aligning release 1 against release 6 and examined each categories. Result is shown on Figure 3.19, revealing that essentially LTR are responsible for these misassemblies.

3.4 Discussion and conclusion

Improving the quality of sequence reconstructions is necessary for a better understanding of the evolution of genomes and their dynamics. Repeated regions present challenges for genome assembly and scaffolding. We presented a pipeline based on scaffold graph enhancement when combining classic paired-end reads information with repeated regions information.

This pipeline shows promising results when used with the SPAdes assembly tool. Probably due to the fact that we based our analysis on reference genomes, which are well-assembled but escape repeat-rich regions, the result may not appear spectacular, however it opens a window on assembly improvement. We also showed that repeated regions are involved on the misassemblies, and that they are essentially transposable elements, which is not surprising but allows us to concentrate on these particular repeats. Amongst those transposable elements, LTR were responsible for the vast majority of gaps observed on the *D. melanogaster* previous releases.

A lot of pending questions remain however. First, it would be interesting to exploit other options when using the pipeline. For the moment, the re-weighting of the consistent edges is quite arbitrary, and depends on the size of the clusters. It would be interesting to study the robustness of this criteria, with respect to the clustering scale for instance, as well as it possible improvement using distance information. Indeed, distance between contigs may be estimated using the pairing information together with the insert size between mate fragments in the short reads sequencing. This estimation is not really precise, but may help refining the consistency in ambiguous cases, when compared to the length of the detected RRs. In the presented version, the removal of intercontig edges is a binary decision process: we decide to keep or to remove edges. This process could be done with more subtlety by introducing a continuous measure on the edges reliability, which would influence the weight of the edge positively ("keep the edge" case) or negatively ("remove the edge" case). For instance we could try to quantify how we can come across these RRs randomly, and consequently to establish probability of decision. Of course, another natural perspective of our work is to extend it to a larger variety of genomes and assembly tools.

Chapter 4

Conclusion and Perspectives

Contents

4.1 BREC : A user-friendly tool for accurate recombination rate and chromatin boundary estimates 107
4.1.1 BREC's limitations 108
4.1.2 Ongoing deployment of BREC for an install-free web access 108
4.1.3 BREC 2.0 is on the way 109
4.1.4 How can BREC serve the community? 110
4.2 A new assembly pipeline to improve the assembly of repeat rich regions 111
4.3 Conclusion/Discussion : Towards mosquito genomes 112

In an attempt to explain, as precisely as possible, the impact of TEs on the evolution of genomes, we needed to produce various types of information regarding their abundance, distribution and dynamics, at the genome-wide scale. To do so, we started with developing a set of computation methods and tools which are non-genome specific. We distinguish two contributions, the first one providing analysis on the chromatin structure of genomes, the second one focusing on the production of high quality genomes.

4.1 BREC : A user-friendly tool for accurate recombination rate and chromatin boundary estimates

In chapter 2 , we propose an automated computational tool, based on the Marey maps method, allowing to identify heterochromatin boundaries along chromosomes and estimating local recombination rates: called **BREC** for **B**oundaries and **REC**ombination rate estimates. BREC is provided within an R-package and a Shiny web-based graphical user interface. BREC takes as input the same genomic data, genetic and physical distances, as in previous tools. It follows a workflow that, first, tests the data quality and offers a cleaning option, then estimates local recombination rates and identify HCB. Finally, BREC re-adjusts recombination rate estimates along heterochromatin regions, the centromere and telomere(s).

BREC is non-genome-specific, running even on non-model genomes as long as genetic and physical maps are available. BREC handles different markers' density and distribution issues.

BREC's heterochromatin boundaries have been validated with cytological equivalents experimentally generated on the fruit fly *D. melanogaster* genome, for which BREC returns congruent corresponding values. Also, BREC's recombination rates have been compared with previously reported estimates. Based on the promising results, we believe our tool has the potential to help bring data science into the service of genome biology and evolution. We introduce BREC within an R-package and a Shiny web-based user-friendly application yielding a fast, easy-to-use, and broadly accessible resource. BREC R-package is available at the GitHub repository `https://github.com/GenomeStructureOrganization/BREC`.

Identifying the boundaries delimiting euchromatin and heterochromatin allows investigating recombination rate variations along the whole genome, helping to compare recombination patterns within and between species. Furthermore, such functionality is fundamental for identifying the position of the centromeric and telomeric regions. Indeed, the position of the centromere along the chromosome has an influence on the chromatin environment, and recent studies are interested in investigating how genome architecture may change with centromere organization (Muller, Gil, and Drinnenberg, 2019).

Throughout this thesis project, and especially for our software development, our vision has been to not only share our computational solutions with the scientific community, but also to cover as much as possible the minimum requirements which would allow other researchers to easily find, access, and reuse our software and data resources. We tried to ensure some of the increasingly demanded FAIR requirements, which aim for providing Findable, Accessible, Interoperable, and Reusable research software and datasets (Katz, Gruenpeter, and Honeyman, 2021).

4.1.1 BREC's limitations

We identified some limitations that may make the use of BREC less relevant, and which can be handled in a future version, such as:

- The choice of the best regression model and span value in case of the Loess.
- Taking into account the non-zero recombination rates in (sub)telomeric regions as well as the sex-biased recombination landscape which in some cases would not be precisely representative of such variation in the species (Sardell and Kirkpatrick, 2020).
- Handling the issue of the overlapping heterochromatin boundaries (see Figure C.1).

4.1.2 Ongoing deployment of BREC for an install-free web access

Figure 4.1 shows a screenshot of the self-service platform `https://www.shinyapps.io/` that we chose to test a first deployment of the BREC shiny app. This will allow to switch to an install-free alternative with a direct online access in order to improve the user experience and avoid most of the technical issues related to portability and scalability. This process is a work in progress as the R-package should be adapted first before it can be correctly deployed on the server.

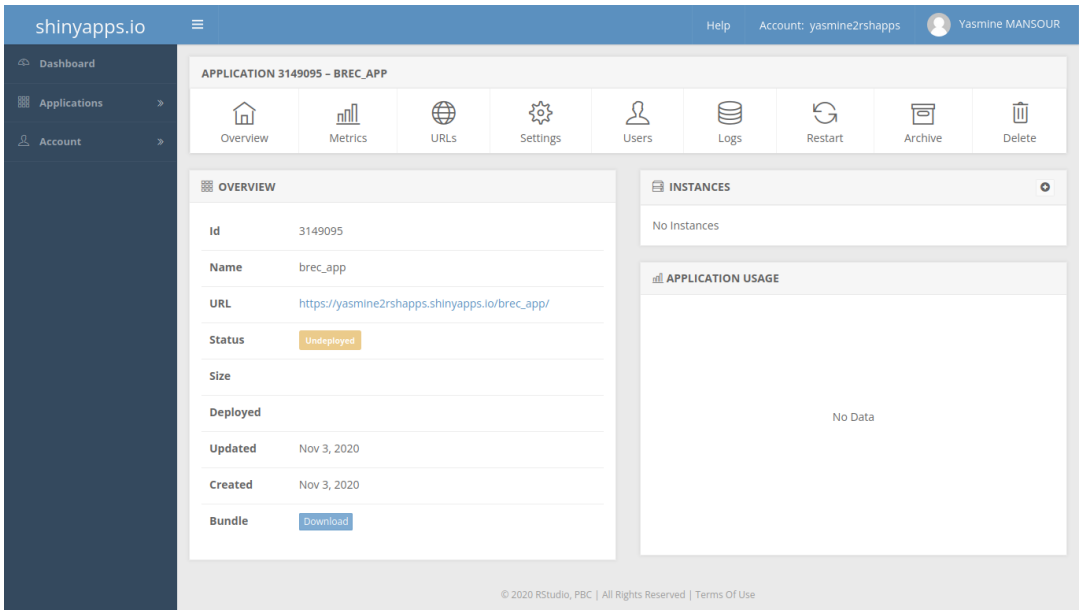


FIGURE 4.1: Screenshot of the ongoing deployment of BREC on the shinyapps.io platform.

4.1.3 BREC 2.0 is on the way

The new version of BREC is a work in progress, and it will mainly provide the significant update of running in the whole-genome mode, where BREC will automatically run on all the available chromosomes of a specific genome. The Figure 4.2 shows a screenshot of the current development status, where the identified centromeric and telomeric regions are represented by the corresponding ideograms, and the centromere is distinguished by a red dot for more clarity.

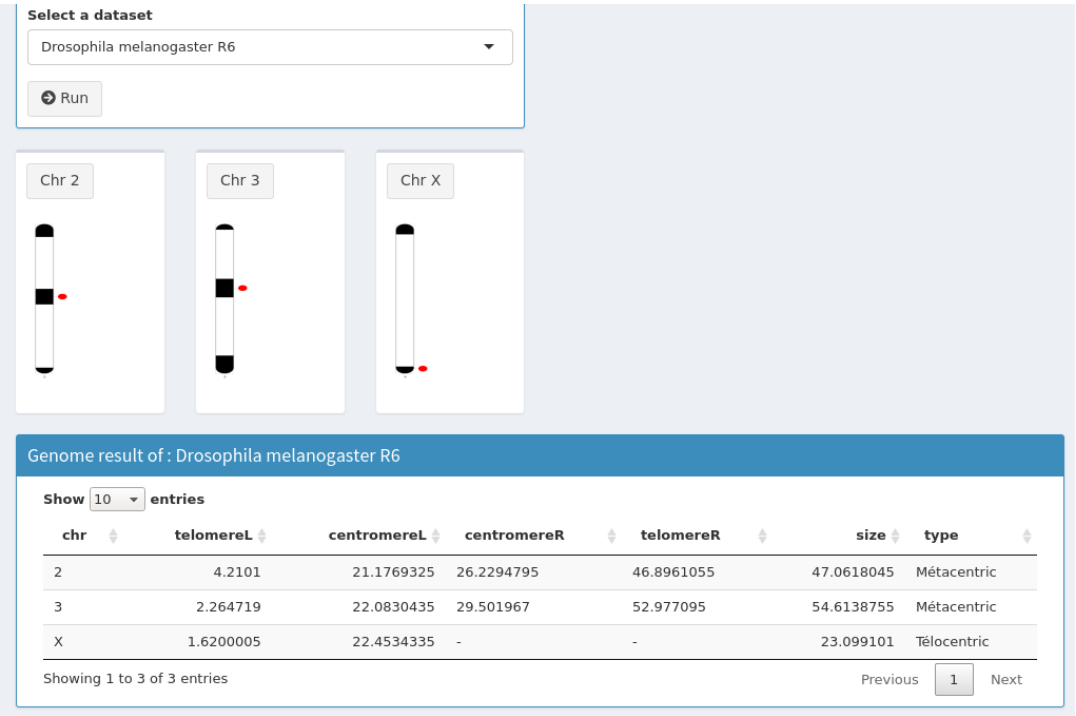


FIGURE 4.2: Screenshot of current BREC interface status.

In addition to ergonomic improvements, we consider methodological evolutions. As short-term perspectives for this work, we may consider extending the robustness tests to additional datasets with high quality and mandatory information (*e.g.* boundaries identified with the cytological method, high quality maps). Retrieving such datasets seems to become less and less complicated. We may also improve the identification of boundaries with a more refined analysis around them, using an iterative multi-scale algorithm for instance. As mid-term perspectives, we underline that BREC could integrate other algorithms aiming to provide further analysis options such as the comparison of heterochromatin regions between closely related species. Also, we are aware that it would be interesting to compare BREC results with more existing methods. Thus, we plan to properly do so in the near future.

4.1.4 How can BREC serve the community?

Finally, we are highly interested in the different facets of applying BREC. Figure 4.3 gives a glance at the type and scale of studies that would benefit form our BREC package, in order to further advance the understanding of how and why recombination rates vary within and between species, and this impacts the architecture and evolution of eukaryotic genomes.

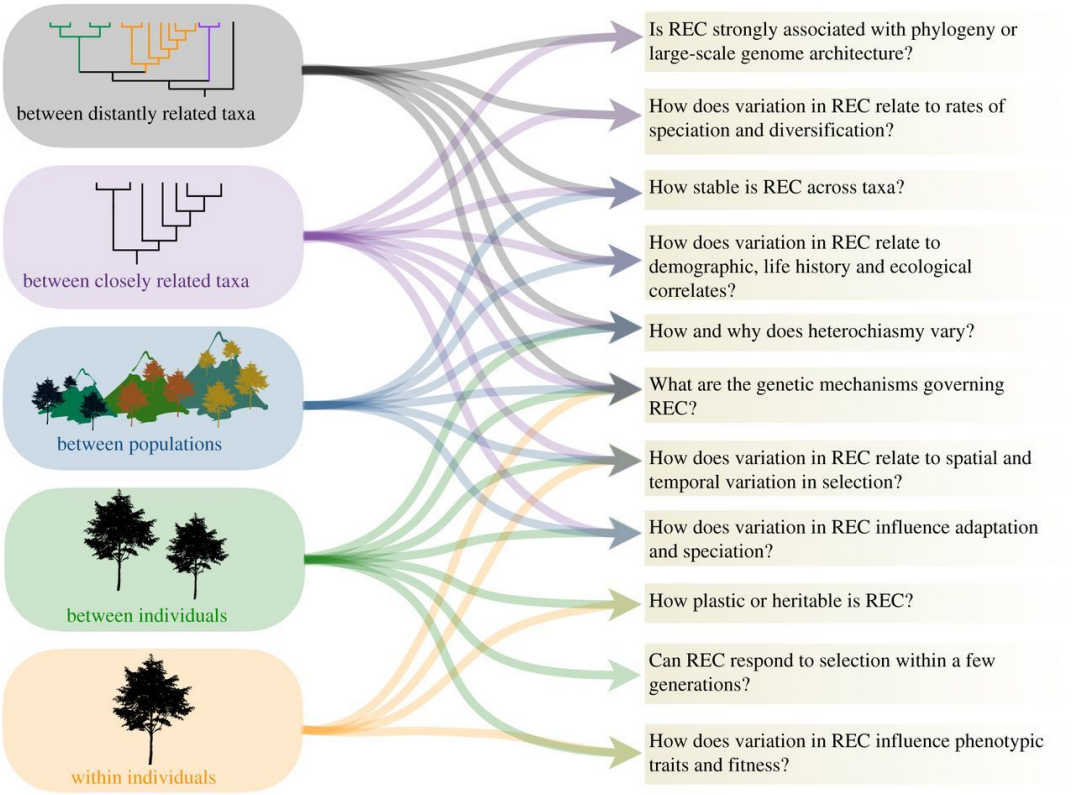


FIGURE 4.3: Comparing recombination landscape and frequency (REC) across different taxonomic and spatial scales (boxes on the left) provides complementary data to address outstanding questions about how and why recombination varies (boxes on right). [From (Stapley et al., 2017)]

Amongst the most accessible questions, we can highlight that studies relative to specific transposable elements, and addressing their association with genomic features and evolution, need data provided by BREC as inputs. This is the case for instance in (Chen et al., 2020), where they combine TEs insertion landscape, recombination rate estimates and methylation data. In this paper, the recombination rate estimate is qualified as "low" and is not really recent (2006), thus we may hope that BREC could provide a better input data for such studies.

4.2 A new assembly pipeline to improve the assembly of repeat rich regions

In Chapter 3 we addressed the impact of DNA repeats on the quality of the genome assemblies. We proposed a pipeline to improve genome assembly at the scaffolding step, by taking into account the information provided by the annotation of contigs with known repeated regions. We got encouraging results on well-referenced genomes. This work was a proof of concept, which needs to be enriched in the future.

First, as a short-term perspective, it would be interesting to explore the robustness of the method with respect to several factors like input data quality and features and solving methods. We would like to apply them on real datasets as well, and cross

the results with other scaffolding methods using external information. A second short-term perspective concerns the core-algorithm, which may be optimized and generalized to accept a wider range of source information. For instance, it would have been interesting to enrich the pipeline with an automatic TE detection step on sufficiently large contigs, and with TE from other databases.

Also, we focused in this study on short read assemblies, which concern a vast majority of fragmented existing genomes on databases. Other sequencing data are also exploited, yielding less fragmented genomes, like long-read dedicated or hybrid methods. This would imply to adapt the decision algorithm to hybrid or long-read dedicated scaffolding method, which are based on different kinds of graphs. For instance, we could map RRs directly on long reads, and exploit those which are overlapping extremities to prevent misassemblies.

A mid-term perspective would be to enlarge the application field of our tool by testing whether some TEs are more disruptive elements in the face of genome assembly process and if this due the TE biology or the TE age. And if this appears to be true, how could we infer information on TEs obtained before the assembly to limit this disruptive effect during the assembly. For instance, crossing the TE landscape of related species with the TE contents on contigs could yield evidences on their putative localisation on the genome.

4.3 Conclusion/Discussion : Towards mosquito genomes

Overall, our preliminary results of BREC seem encouraging as we are able to reidentify with accuracy the pericentromeric regions. We believe that the Shiny interface will be very useful for non computer scientists or users working on non-model organisms to appreciate the BREC outputs and choose the best models.

Though previous contributions have been thought as non genome-specific tools, we saw in Chapter 2 that crossing generic treatments with specific information provide useful insight to understand how are organised those genomes. A lot of questions still remain. One research aim in the ISEM team is to establish links between TE dynamics and the chromatin landscape. As present in the introduction, the strong genome size and repeat content variation across mosquitoes species make them good models to investigate such association. The preliminary results obtained using BREC on *Ae. aegypti* support the veracity of our approach as BREC is able to define with accuracy the pericentromeric regions. In *Cx. pipiens*, our preliminary analyses show a clear association between the chromatin structure and the distribution of some TE elements : While MITE elements are enriched in euchromatic regions, LINE elements appear active and dense in pericentromeric regions (see Appendix C). A comparative genomics analysis between *Ae. aegypti* and *Cx. pipiens* suggests that genome size variation is partially explained by large insertions/deletions in the pericentromeric regions where TEs, and more specially LINEs, are known to appear highly dynamic (see Appendix C). Such observation suggests the implication of the TE dynamics in the genome size variation through pericentromeric expansion/contraction. To test this hypothesis, it would be interesting to conduct the same kind of study on other mosquito genomes such as *An. gambiae*.

More and more studies are conducted using large genomic dataset, like for instance in (Melo and Wallau, 2020), where 24 mosquitoes genomes are analysed to highlight

the horizontal transfers of TEs between species. This yield a wider perspective on how genomes are evolving and use TEs as vectors to propagate adaptation.

Appendix A

Communications

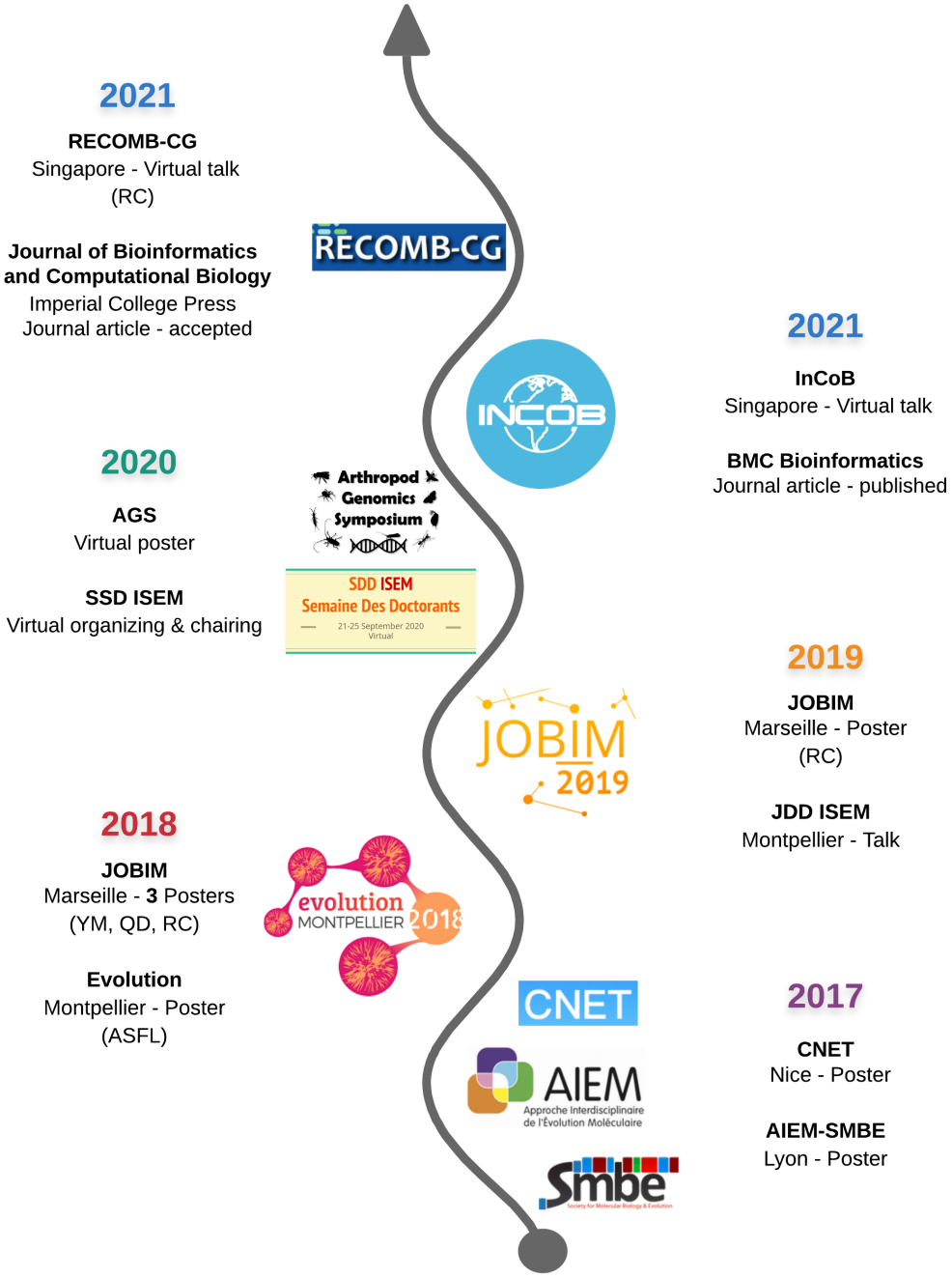


FIGURE A.1: Vertical timeline for the main scientific communications related to this thesis project.

A.1 Conferences

A.1.1 Automatic identification of heterochromatin boundaries through recombination rate estimates

InCoB 2021 International conference of Bioinformatics https://incob.apbionet.org/incob20/ (TALK).

Doctiss 2019 PhD days organized by the I2S doctoral school in Montpellier https://seminaire.inrae.fr/doctiss2019/ (submitted abstract, no vacancy for a talk).

JDD 2019 PhD days organized by the ISEM (TALK).

JOBIM 2018 Open Days of Biology, Computer Science, and Mathematics (Journées Ouvertes de Biologie Informatique et Mathématiques) organized in Marseille. https://jobim2018.sciencesconf.org/ (POSTER)

AIEM-SMBE 2017 "Approche Interdisciplinaire de l'Evolution Moléculaire" joint annual meeting organized with the Society for Molecular Biology and Evolution in Lyon https://project.inria.fr/aiem2017/fr/ (POSTER).

A.1.2 How transposable elements shape mosquito genomes

CNET 2017 National conference on transposable elements organized in Nice https://cnet2017.sciencesconf.org/ (POSTER).

A.1.3 Organization of insect genomes driven by some transposable element families

2nd Joint Congress on Evolutionary Biology 2018 organized in Montpellier https://www.labex-cemeb.org/en/ii-joint-congress-evolutionary-biology-montpellier-2018 (POSTER presented by Anna-Sophie FISTON-LAVIER)

A.1.4 How to improve genome assembly using repetitive elements

AGS 2020 Arthropods Genomics Symposium. http://i5k.github.io/ags2020 (VIRTUAL POSTER)

JOBIM 2019 Open Days of Biology, Computer Science, and Mathematics (Journées Ouvertes de Biologie Informatique et Mathématiques) organized in Nantes. https://jobim2019.sciencesconf.org/ (POSTER presented by Remy COSTA)

JOBIM 2018 Open Days of Biology, Computer Science, and Mathematics (Journées Ouvertes de Biologie Informatique et Mathématiques) organized in Marseille. https://jobim2018.sciencesconf.org/ (POSTER presented by Quentin DELORME)

A.1.5 In-depth analysis of the impact of transposable elements on genome assembly quality

JOBIM2018 Open Days of Biology, Computer Science, and Mathematics (Journées Ouvertes de Biologie Informatique et Mathématiques) organized in Marseille. https://jobim2018.sciencesconf.org/ (POSTER presented by Remy COSTA)

A.2 Journal Articles / Conference supplements

A.2.1 Delorme, Q., Costa, R., Mansour, Y., Fiston-Lavier, AS., and Chateau, A. Involving Repetitive Regions in Scaffolding Improvement. To appear in Journal of Bioinformatics and Computational Biology special issue on RECOMB-CG2021.

A.2.2 Mansour, Y., Chateau, A. and Fiston-Lavier, AS. BREC: an R package/Shiny app for automatically identifying heterochromatin boundaries and estimating local recombination rates along chromosomes. BMC Bioinformatics 22, 396 (2021).

https://doi.org/10.1186/s12859-021-04233-1

Pre-print version in bioRxiv (2020) https://www.biorxiv.org/content/10.1101/2020.06.29.178095v3

Automatically identifying eu-hetero-chromatin boundaries through recombination rate estimates

Yasmine Mansour *† [1,2], Annie Chateau [1,3], Anna-Sophie Fiston-Lavier [2]

[1] Laboratoire dÍnformatique de Robotique et de Microélectronique de Montpellier (LIRMM) – Université de Montpellier : UMR5506, Centre National de la Recherche Scientifique : UMR5506 – 161 rue Ada - 34095 Montpellier, France

[2] Institut des Sciences de lÉvolution de Montpellier (ISEM) – Université de Montpellier, Institut de recherche pour le développement [IRD] : UR226, Centre National de la Recherche Scientifique : UMR5554 – Place E. Bataillon CC 064 34095 Montpellier Cedex 05, France

[3] Institut de Biologie Computationnelle (IBC) – Centre de Coopération Internationale en Recherche Agronomique pour le Développement, Institut National de la Recherche Agronomique, Institut National de Recherche en Informatique et en Automatique, Université de Montpellier, Centre National de la Recherche Scientifique – 95 rue de la Galéra, 34095 Montpellier, France

Meiotic recombination is a vital biological process which plays an essential role for investigating genome-wide structural as well as functional dynamics. Various methods for estimating recombination rates exist in the literature. Population genetic based-methods [Stumpf and McVean, 2003] provide accurate fine-scale estimates. Nevertheless, these methods are very expensive, time-consuming, require a strong expertise and, most of all, are not applicable on all kinds of organisms. Moreover, the sperm-typing method [Jeffreys et al., 2000], which is also extremely accurate providing high-density recombination maps, is male-specific and share the same experimental requirements as population genetic methods. On the other hand, a purely statistical approach, the Marey Maps [Chakravarti, 1991], could avoid some of the above issues based on other available genomic data : the genetic and physical distances. The Marey maps for recombination estimates consist on correlating, for the same chromosome, the physical map with the genetic map containing respectively physical distances and genetic distances for a set of genetic markers. Despite the efficiency of this method and mostly the availability of physical and genetic maps, generating recombination maps rapidly and for any organism is still challenging. Hence, the increasing need of an automatic, portable and easy-to-use tool.

Here, we propose an automated bioinformatic solution based on the Marey maps method in order to provide local recombination rate estimates for various organisms. Furthermore, our approach allows to determine the eu-hetero-chromatin boundaries along chromosomes. This functionality is fundamental for identifying the location of the peri/centromeric and telomeric regions known to present a reduced recombination rate in most genomes. Most importantly for genomes which are provided as whole chromosomes instead of two arms per chromosome. We implemented our recombination tool by fitting a third-order polynomial to each chromosome based on genetic and physical maps. Compared to previous tools [Fiston-Lavier et al., 2010, Rezvoy et al., 2007], we have add a couple of new modules as to assess the quality of the data

*Speaker

†Corresponding author: yasmine.mansour@umontpellier.fr

(i.e. number and distribution of the markers along the genome) and to remove low-quality data according to the user's preference. Our approach automatically re-adjusts estimates in regions with a depletion of fitness between the polynomial and the data to detect the eu-hetero-chromatin boundaries for centromeric and telomeric regions in order to keep the estimates as authentic as possible to the biological process. Identifying these boundaries allows investigating recombination variations along the whole genome which will help comparing recombination patterns within and between species, especially insects in our case.

Our approach for the eu-hetero-chromatin boundaries detection has been primarily validated with cytological results that are experimentally generated on the *Drosophila melanogaster* genome [Comeron et al., 2012]. Moreover, since the pipeline we are proposing is non-genome-specific, our study is efficiently portable on other model as well as non-model genomes for which both genetic and physical maps are available. We have started interpreting the results on the mosquito specie *Culex pipiens*. We estimated the recombination rate along this genome and identified the heterochromatin boundaries on its three chromosomes. Also, after annotating its TEs, we have analyzed the correlation between TEs and recombination patterns. As in *D. melanogaster*, we observed non-homogenous distribution for active TE families such as LINEs and MITEs. In *Cx. pipiens*, while LINEs are enriched in pericentromeric regions, MITEs exhibit a higher density in euchromatin. In an attempt to explain such distribution bias, we investigated the dynamics for these two TE families through a comparative genomic approach carried out on other insect genomes.
We find our preliminary results quite promising since the TE distribution patterns across genomes generally show enrichment in specific regions such as constitutive heterochromatic exhibiting low recombination and low gene density. Therefor, we aim to take advantage of genome-wide recombination landscape to seek an explanation to the cause/effect association between recombination rate and TEs.

Keywords: Recombination rate, heterochromatin, transposable elements, comparative genomics, bioinformatics

Automatic identification of heterochromatin boundaries through recombination rate estimates

Yasmine MANSOUR [1,3*], Annie CHATEAU [1,2], Anna-Sophie FISTON-LAVIER [3]

[1] Laboratoire d'Informatique de Robotique et de Microélectronique de Montpellier (LIRMM) – Université de Montpellier, CNRS - Montpellier, France

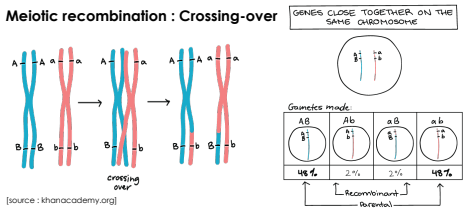
[2] Institut de Biologie Computationnelle (IBC) - Montpellier, France

[3] Institut des Sciences de l'Evolution - Montpellier (ISEM) – Université de Montpellier, CNRS - Montpellier, France

***Contact :** yasmine.mansour@umontpellier.fr

Motivation

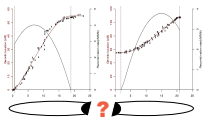
Meiotic recombination : Crossing-over



[source : khanacademy.org]

Meiotic recombination is a vital biological process which guarantees the **diversity** of genetic material over generations. This process consists on the **exchange of DNA fragments** within and between chromosomes.

Various experimental (biological) methods for estimating recombination rate exist. They provide accurate fine-scale estimates, yet, they are **very expensive, time-consuming, require a strong expertise** and, most of all, **are not applicable on all kinds of organisms** [1, 2]. A purely statistical approach, **the Marey Maps** [3], could avoid some of the above issues based on other **available genomic data : the genetic** and **physical distances**.



Fiston-Lavier, et al. 2010 adapted [5]

Several Marey Map-based tools are available for different specefic genomes [4, 5, 6]. However, **more adapted** tools are required to better handle New Generation Sequencing (**NGS**) data, which are providing new insights for :

- Enhancing **whole-genome assembly quality**
- Producing **high density** genetic maps

Here, we propose a **non-genome-specific tool for estimating recombination rates** with an **automated identification of heterochromatin boundaries (HCB)**.

Computational tool and results : *Drosophila melanogaster* genome - Release 5 - chromosome 2

Data quality control
test the quality of input data according to the number and distribution of markers along the chromosome

Optional data cleaning
detect and eliminate outliers using boxplot statistics

Marey Map-based recombination rate estimates

(1) Build the Marey Map correlating between genetic and physical maps (distances) along the chromosome

(2) Interpolation : generate a 3rd degree polynomial that better fits the genomic data

(3) Local recombination rate estimates are represented by the prime derivative of the polynomial

Identification of **HeterChromatin Boundaries (HCB)**

(4) Use the coefficient of determination R^2 to measure how well the generated polynomial fits the input data

Compute R^2 for each marker then cumulative R^2 curves which represent values accumulated along the chromosome in 2 directions in order to eliminate the bias induced by the accumulation process

(4) R^2-forward : starting from the beginning of the chromosome to provide the left HCB

(5) Using a sliding window, apply a statistical test on each cumulative R^2 curve to identify **centromeric** boundaries

Pseudo algorithm for HCB identification

```
starting_coordinate_of_sliding_window ← min(recombination_rate)
sliding_window_size ← max(inter – marker_ranges_of_physical_map)
sliding_step ← sliding_window_size
found_HCB ← false
λ ← room(nbsegs_of_segments_inside_sliding_window)
while not(end_of_chromosome) AND not(found_HCB) do
    if λ > 0 then
        move_one_step_forward(sliding_window)
    else
        found_HCB ← true
    end
end
HCB ← ending_coordinate_of_sliding_window)
return HCB
```

(6) R^2-backwards : starting from the end of the chromosome to provide the right HCB

(7) Detect the minimum value on both R^2 curves to identify **telomeric** boundaries

(8) Extrapolation
update recombination rate estimates of centromeric and telomeric regions according to the identified heterochromatin boundaries

Data quality control

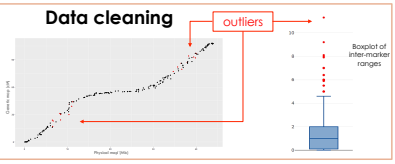
Using inter-marker ranges of physical distances to assess :

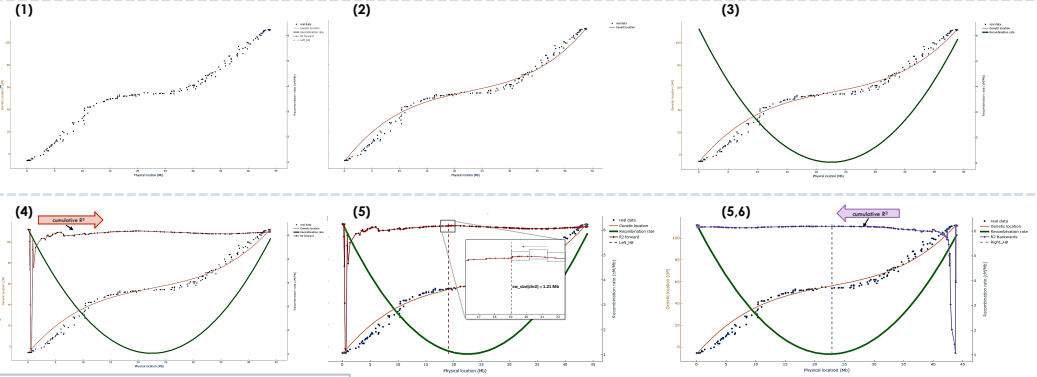
- The number of data points per chromosome
- Their distribution along the chromosome : Chi-squared test χ^2

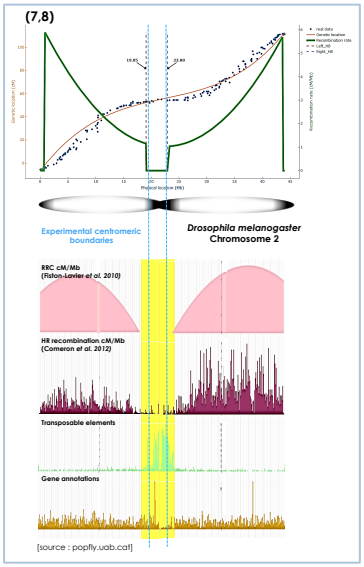
- Boolean function to alert the user of his data quality

Data cleaning

Detecting outliers using boxplot statistics based on inter-marker ranges on the genetic map. The user is given the option of deleting all, a part or none of the detected outliers.







[source : popfly.uab.cat]

Validation of centromeric HCB with experimental results

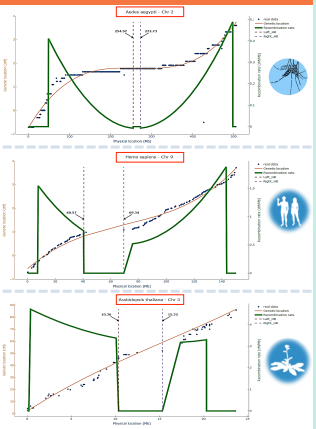
Fruit Fly *Drosophila melanogaster*

Coordinates of heterochromatin boundaries on each chromosome	Cytological results from [7] converted* to Release 5 (Mb)	Our results (Mb)
2	[19.67 - 22.74]	[19.05 - 22.88]
3	[20.55 - 24.62]	[21.17 - 23.39]
X	20.37	19.53

[* flybase.org/convert/coordinates]

Non-genome specific tool

Species	Common Name	Kingdom	Subgroup
Aedes aegypti	**Yellow fever mosquito**	animal	invertebrate
Anopheles gambiae	**African malaria mosquito**	animal	invertebrate
Apis mellifera scutellata	**Honeybee**	animal	invertebrate
Bombyx mandarina	Silkworm	animal	invertebrate
Caenorhabditis briggsae	Roundworm	animal	invertebrate
Caenorhabditis elegans	Roundworm	animal	invertebrate
Culex pipiens	**Common house mosquito**	animal	invertebrate
Drosophila melanogaster	**Fruitfly**	animal	invertebrate
Drosophila pseudoobscura	**Fruitfly**	animal	invertebrate
Heliconius melpomene melpomene	Postman butterfly	animal	invertebrate
Bos taurus	Cow	animal	vertebrate
Canis lupus	Wolf	animal	vertebrate
Cynoglossus semilaevis	Tongue sole	animal	vertebrate
Danio rerio	Zebrafish	animal	vertebrate
Equus ferus przewalskii	Prewalksi's horse	animal	vertebrate
Ficedula albicollis	Collared flycatcher	animal	vertebrate
Gallus gallus	Chicken	animal	vertebrate
Gasterosteus aculeatus	Stickleback	animal	vertebrate
Homo sapiens	**Human**	animal	vertebrate
Lepisosteus oculatus	Spotted gar	animal	vertebrate
Macaca mulatta	Rhesus macaque	animal	vertebrate
Meleagris gallopavo	Turkey	animal	vertebrate
Mus musculus castaneus	**House mouse**	animal	vertebrate
Oryzias latipes	Medaka	animal	vertebrate
Ovis canadensis	Bighorn sheep	animal	vertebrate
Papio anubis	Olive baboon	animal	vertebrate
Sus scrofa	Wild boar	animal	vertebrate
Citrus reticulata	Mandarin Orange	plant	woody
Gossypium raimondii	New world cotton	plant	woody
Populus trichocarpa	Black cottonwood	plant	woody
Prunus davidiana	David's peach	plant	woody
Arabidopsis thaliana	**Thale cress**	plant	herbaceous
Brachypodium distachyon	Purple false brome	plant	herbaceous
Capsella rubella	Pink Shepherd's Purse	plant	herbaceous
Citrullus lanatus lanatus	Watermelon	plant	herbaceous
Cucumis sativus var. hardwickii	Cucumber	plant	herbaceous
Glycine soja	Wild soybean	plant	herbaceous
Medicago truncatula	Barrel medic	plant	herbaceous
Oryza rufipogon	**Wild rice**	plant	herbaceous
Setaria italica	Foxtail millet	plant	herbaceous
Sorghum bicolor subsp. Verticilliforum	Wild Sudan grass	plant	herbaceous
Zea mays ssp parviglumis	Teosinte	plant	herbaceous



A sample of available genetic and physical maps for **40 species** [6] **updated** with new genome versions and enriched with **2 recently assembled mosquito genomes : Culex pipiens** and **Aedes aegypti** [8]

Conclusions and Perspectives

- A **non-genome-specific computational** tool for the estimation of recombination rates along chromosomes.
- **Automatic identification of heterochromatin boundaries**.
- **Easy-to-use** tool providing **reliable broad-scale** results with interactive plots.
- **Validation** of obtained results with other **genomic features** of *D. melanogaster genome* : **transposable elements distribution** and **gene density** among others.
- **Strong correlation** of recombination rate and transposable elements distribution along *D. melanogaster* [9] as well as *Culex pipiens genomes*.
- ***Next step*** *: identifying genome* ***assembly errors*** based on genetic map outliers which are expected to be related to physical map errors, thus, misassemblies.

References

1. Jeffreys, A. J. High resolution analysis of haplotype diversity and meiotic crossover in the human TAP2 recombination hotspot. Hum. Mol. Genet. 9, 725–733 (2000).
2. Stumpf, M. P. H. & McVean, G. A. T. Estimating recombination rates from population-genetic data. Nat. Rev. Genet. 4, 959–968 (2003).
3. Chakravarti, A. A graphical representation of genetic and physical maps: The Marey map. *Genomics* **11**, 219–222 (1991).
4. Rezvoy, C., Charif, D., Guéguen, L. & Marais, G. A. B. MareyMap: An R-based tool with graphical interface for estimating recombination rates. Bioinformatics 23, 2188–2189 (2007).
5. Fiston-Lavier, A. S., Singh, N. D., Lipatov, M., & Petrov, D. A. (2010). *Drosophila melanogaster* recombination rate calculator. Gene, 463(1–2), 18–20.
6. Corbett-Detig, R. B., Hartl, D. L. & Sackton, T. B. Natural Selection Constrains Neutral Diversity across A Wide Range of Species. *PLoS Biol.* **13**, e1002112 (2015).
7. Bergman, C. M., Quesneville, H., Anxolabéhère, D. & Ashburner, M. Recurrent insertion and duplication generate networks of transposable element sequences in the Drosophila melanogaster genome. Genome Biol. 7, R112 (2006).
8. Dudchenko, O., Batra, S. S., Omer, A. D., Nyquist, S. K., Hoeger, M., Durand, N. C. et al. (2017). *De novo* assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science*, *356*(6333), 92–95.
9. Petrov, D. A., Fiston-Lavier, A.-S., Lipatov, M., Lenkov, K. & Gonzalez, J. Population Genomics of Transposable Elements in Drosophila melanogaster. Mol. Biol. Evol. 28, 1633–1644 (2011).

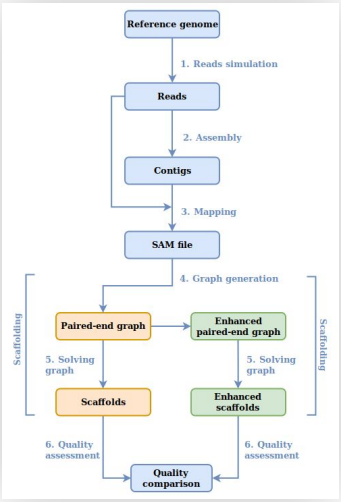
Acknowledgements

We thank the **Algerian Government** and particularly the **Ministry of Higher Education and Scientific Research** for funding this thesis, and also the **LabEx CeMEB (ERJ)** for funding a part of this study.

Yasmine MANSOUR, 4th and final year PhD student

Arthropod Genomics Symposium - July 2020 -

Ongoing sub-project: How to improve mosquito genome assembly using repetitive DNA?



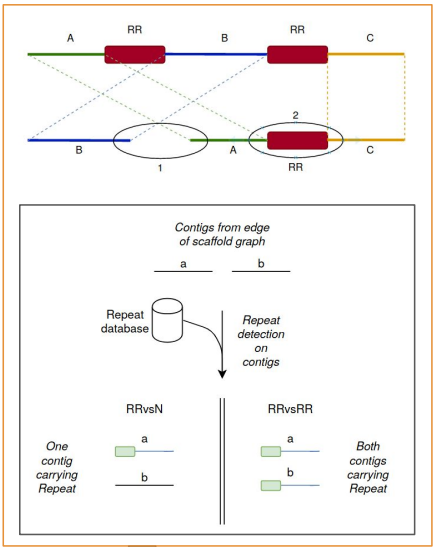
Our pipeline for enhancing the paired-end graph (in orange) with an additional RR-based scaffolding step (in green)

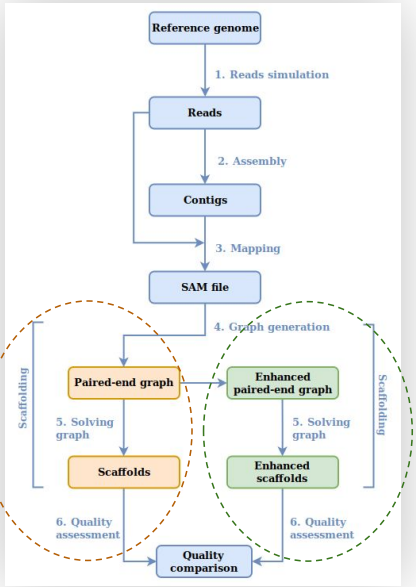
- **Aim**
  - Improve assemblies by taking into account the presence of repetitive DNA sequences in general (RR), and transposable elements (TEs) in particular
- **Methods**
  - Analysis of the relationship between the presence and nature of RRs, and genome assembly errors
  - Pipeline proposed: the use of RR-based graphs to correct the paired-end scaffolding graph and enhance the assembly quality
  - See next slides for further details...
- **Results**
  - Among all RRs, TEs are the most disruptive in *D. melanogaster* and *C. elegans*
  - Our solution yields promising results
- **Discussion and next step**
  - What about targeting only TEs, or specific TE families?
  - How to adapt and apply this pipeline on three mosquito genomes
- **Team**
  - Yasmine MANSOUR, Quentin DELORME, Rémy COSTA, Anna-Sophie FISTON-LAVIER and Annie CHATEAU

Availability: 9-10 PM CET, Friday 24/07, sometime next week, and anytime offline

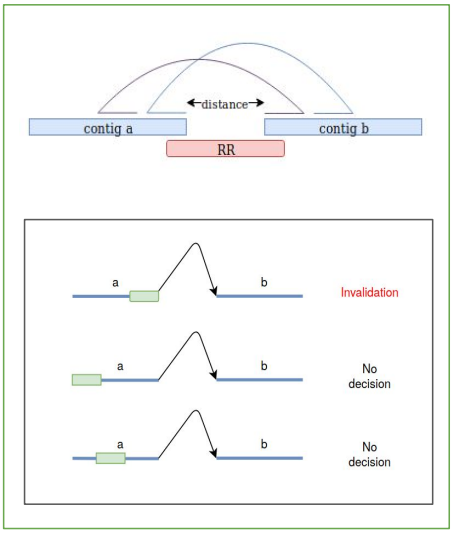
Institute of Evolution Science (ISEM), Lab of Computer Science (LIRMM), University of Montpellier, France
The Mediterranean Centre for Environment and Biodiversity Laboratory of Excellence (CeMEB LabEx)
Algerian Excellence Scholarship Program "Averroes", Ministry of Higher Education and Scientific Research

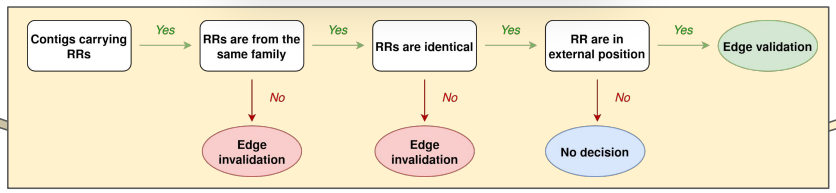
RRs are the challenge





RRs are the solution





Arthropod Genomics Symposium - July 2020 -

→ Stay tuned...

[Y. MANSOUR et al., How repetitive DNA disrupt genome assembly, preprint to be available soon]

How to efficiently adapt our pipeline to mosquitoes, with a focus on TEs?

Pipeline steps / species	1. Simulating reads	2. Assembly	3. Mapping	Scaffolding: 4. Graph generating	Scaffolding: 5. Graph solving	6. Quality assessment	Current progress state
Drosophila melanogaster and *Caenorhabditis elegans*	Tool: ART Technology: Illumina HiSeq 2000 Coverage: 20X Reads size: 100bp Insert size: 300bp Standard deviation: 10%	Spades Minia	BWA MEM Minimap2	Scafftools		QUAST-LG BUSCO	Done
Anopheles gambiae	• Starting point: the same tools above • Any advice is very much welcomed and highly appreciated regarding: ○ The tools have been used ○ Data and simulation technology ○ The pipeline workflow management system: currently Snakemake ○ Implementation / programming language: currently Python ○ Further aspects not yet considered ○ ...					Arthropod Genomics Symposium - July 2020 -	On going
Culex pipiens							To do
Aedes aegypti							

Our work in the context of my doctoral thesis project

Arthropod Genomics Symposium - July 2020 -

1. **Y. MANSOUR**, M. HAMOUMA, A. CHATEAU, A.-S. FISTON-LAVIER, 2017. *How transposable elements shape mosquito genomes, **CNET 2017,** Nice, France,* pages 41-42 [link]
2. **Y. MANSOUR**, A. CHATEAU, A.-S. FISTON-LAVIER 2018. *Automatically identifying eu-heterochromatin boundaries through recombination rate estimates*, ***JOBIM 2018**, Marseille, France,* pages 382-383 [link]
3. R.COSTA, **Y. MANSOUR**, A. CHATEAU, A.-S. FISTON-LAVIER 2018. *In-depth analysis of the impact of transposable elements on genome assembly quality*, ***JOBIM 2018**, Marseille, France,* pages 290-291 [link]
4. Q. DELORME, **Y. MANSOUR**, A. CHATEAU, A.-S. FISTON-LAVIER 2018. *How to improve genome assembly using repetitive elements*, ***JOBIM 2018**, Marseille, France,* pages 98-99 [link]
5. A.-S. FISTON-LAVIER, **Y. MANSOUR**, A. CHATEAU, M. HAMOUMA 2018. *Organization of insect genomes driven by some transposable element families*, ***Evolution 2018,** Montpellier, France* [link]
6. R. COSTA, Q. DELORME, **Y. MANSOUR**, A.-S. FISTON-LAVIER and A. CHATEAU. 2019. *How to involve repetitive regions in scaffolding improvement*, ***JOBIM 2019** - Nantes, France,* page 271 [link]

@2YasmineMANSOUR
yasmine.mansour@umontpellier.fr
yasmine.mansour21@gmail.com

Also, our recent preprint.. and my first :)

★ **Y. MANSOUR**, A. CHATEAU, A.-S. FISTON-LAVIER 2020**.** *BREC: An R package/Shiny app for automatically identifying heterochromatin boundaries and estimating local recombination rates along chromosomes*, **bioRxiv 2020.06.29.178095**; doi: https://doi.org/10.1101/2020.06.29.178095

Institute of Evolution Science (ISEM), Lab of Computer Science (LIRMM), University of Montpellier, France
The Mediterranean Centre for Environment and Biodiversity Laboratory of Excellence (CeMEB LabEx)
Algerian Excellence Scholarship Program "Averroes", Ministry of Higher Education and Scientific Research

SMBE regional meeting in Lyon, Interdisciplinary Approaches for Molecular Evolution

POSTER 9

Yasmine Mansour

Title : **How transposable elements shape mosquito genomes**

Authors : **Yasmine Mansour**[1,2,3*], Mickaël Hamouma[1,2,3], Annie Chateau[1,2] and Anna-Sophie Fiston-Lavier[3]

Address :
1 : Laboratoire d´Informatique de Robotique et de Microélectronique de Montpellier (LIRMM) - Université de Montpellier, CNRS, Montpellier, France.
2 : Institut de Biologie Computationnelle (IBC), Montpellier, France.
3 :Institut des Sciences de l'Evolution de Montpellier (ISEM) Université de Montpellier, CNRS, Montpellier, France.
*Corresponding author: yasmine.mansour@umontpellier.fr

Abstract :
Mosquitoes are human infectious disease vectors that have been extensively studied, not only because of the high genetic diversity their genomes manifest, but also for their remarkably strong capacity of fast adaptation, such as climate changes or insecticide resistance. While several studies of genes known to be involved in adaptation help to shed light on the putative role of transposable elements (TEs) in such evolution process, the impact of TEs on mosquito genome structure and evolution are still poorly tackled (Assogba et al., 2016).
Here, we carry out the study on TE abundance and distribution in mosquito genomes. In March 2017, a research group ended up with the first chromosome-length scaffolds in Culex pipiens quinquefaciatus and Aedes aegypti. We decided to start focusing on the new version of the Cx. pipiens genome assembly (CpipJ3) (Dudchenko et al., 2017). We started developing a new tool to estimate the recombination rates along chromosomes based on Marey maps (Fiston-Lavier et al., 2010) (Rezvoy et al., 2007). Our tool includes a statistical-based approach for the detection of the heterochromatin boundaries that automatically re-adjusts estimates in regions with a depletion of fitness between the polynomial and the data. After assessing the veracity of the tool with experimental data from Anopheles gambiae (Sharakhova et al., 2010), we estimated the recombination rate along the Cx. pipiens new assembly.
On the other hand, we annotated individual TE insertions in Cx. pipiens. We built a Culex specific TE library, a set of canonical sequences representative of TE families in this genome. We then annotated them combining results from homology-based (TEfam database:

https://tefam.biochem.vt.edu/) and signature-based approaches. We reported a high diversity with TE families from the three main types of TEs (DNA, LTR, non-LTR).
The annotation of individual TE insertions in CpipJ3 reveals a higher TE content compared with previous studies (33% instead of 29% for CpipJ2). Our results also showed a nonhomogenous distribution of TEs along the Cx. pipiens chromosomes with an enrichment of TEs in the heterochromatin. In-depth analysis of the TE organization is currently in process. Our results should help explaining the Cx. pipiens genome structure but also assessing the quality of the new release of the assembly.

Automatically identifying chromatin boundaries through recombination rate estimates

Yasmine MANSOUR[1,2], Annie CHATEAU[1,3], Anna-Sophie FISTON-LAVIER[2]

yasmine.mansour@umontpellier.fr
[1]*The Montpellier Laboratory of Informatics, Robotics and Microelectronics* (LIRMM)
[2]*Institute of Evolution Science of Montpellier* (ISEM)
[3]*The Computational Biology Institute* (IBC)*, Montpellier*

Key words — Bioinformatics, Genomics, Recombination rate, Chromatin boundaries, Marey maps.

Résumé :

Meiotic recombination is a vital biological process which guarantees the diversity of genetic material over generations. This process consists on the exchange of DNA fragments within and between chromosomes. Recombination rate is a metric to estimate the frequency of the DNA fragment exchange along the chromosome. Various experimental (biological) methods for estimating recombination rate exist. They provide accurate fine-scale estimates, yet, they are very expensive, time-consuming, require a strong expertise and, most of all, are not applicable on all kinds of organisms [1, 2]. A purely statistical approach, the Marey Maps [3], could avoid some of the above issues based on other available genomic data : the genetic [1] and physical distances [2]. The Marey maps for recombination rate [3] estimates consist on correlating, for the same chromosome, the physical map with the genetic map containing respectively physical distances and genetic distances for a set of genetic markers [4]. Despite the efficiency of this method and mostly the availability of physical and genetic maps, generating recombination maps rapidly and for any organism is still challenging. Thus, there is an increasing need for an automatic, portable and easy-to-use tool.

Here, we propose an automated bioinformatic non-genome-specific solution based on the Marey maps method in order to provide local recombination rate estimates. Furthermore, our approach allows to determine the eu-hetero-chromatin boundaries along chromosomes. This functionality allows identifying the location of the peri/centromeric and telomeric regions known to present a reduced recombination rate in most genomes. We implemented our recombination tool by fitting a third-order polynomial for each chromosome based on genetic and physical maps. Also, we used the R^2 statistic in order to automatically re-adjust estimates in regions with a depletion of fitness between the polynomial and the data. A sliding window on the R^2 curves allows to identify eu-hetero-chromatin boundaries with a reliable accuracy. Compared to previous tools [4, 5], we have added new modules as to assess the quality of the data (i.e. number and distribution of the markers along the genome) and to remove low-quality data according to the user's preference. Our tool is implemented using the R-programming language [5] and thus is simple to run on any platform.

Our results has been primarily validated with experimentally generated equivalents on the fruit fly genome *Drosophila melanogaster* [6]. Moreover, the pipeline we are proposing is efficiently portable on other model as well as non-model genomes for which both genetic and physical maps are available. We find our preliminary results quite promising. Therefore, we aim to take advantage of genome-wide recombination landscape to seek an explanation to the cause/effect association between recombination

1. Genetic distance is a measure that statistically estimates how far apart are two markers on the chromosome, it's unit is CentiMorgan (cM).
2. The physical position of the genetic marker on the chromosome, it's measured in Base pair (bp).
3. Measured in cM/Mb.
4. DNA sequences with known physical location on the genome.
5. https ://www.r-project.org/

rate and genome structure and evolution.

Références

[1] A. J. Jeffreys. High resolution analysis of haplotype diversity and meiotic crossover in the human TAP2 recombination hotspot. *Human Molecular Genetics*, 9(5) :725–733, mar 2000.

[2] Michael P.H. Stumpf and Gilean A.T. McVean. Estimating recombination rates from population-genetic data. *Nat. Rev. Genet.*, 4(12) :959–968, 2003.

[3] Aravinda Chakravarti. A graphical representation of genetic and physical maps : The Marey map. *Genomics*, 11(1) :219–222, 1991.

[4] Anna Sophie Fiston-Lavier, Nadia D Singh, Mikhail Lipatov, and Dmitri A Petrov. Drosophila melanogaster recombination rate calculator. *Gene*, 463(1-2) :18–20, 2010.

[5] Clément Rezvoy, Delphine Charif, Laurent Guéguen, and Gabriel A.B. Marais. MareyMap : An R-based tool with graphical interface for estimating recombination rates. *Bioinformatics*, 23(16) :2188–2189, 2007.

[6] Josep M. Comeron, Ramesh Ratnappan, and Samuel Bailin. The Many Landscapes of Recombination in Drosophila melanogaster. *PLoS Genet.*, 8(10) :e1002905, oct 2012.

How transposable elements shape mosquito genomes

Yasmine Mansour*[1,2,3], Michael Hamouma[1,2,3], Annie Chateau[1,2], and Anna-Sophie Fiston-Lavier†[3]

[1]Laboratoire dÍnformatique de Robotique et de Microélectronique de Montpellier (LIRMM) – Université de Montpellier : UMR5506, Centre National de la Recherche Scientifique : UMR5506 – CC 477, 161 rue Ada, 34095 Montpellier Cedex 5, France

[2]Institut de Biologie Computationnelle (IBC) – Centre de Coopération Internationale en Recherche Agronomique pour le Développement, Institut National de la Recherche Agronomique, Institut National de Recherche en Informatique et en Automatique, Université de Montpellier, Centre National de la Recherche Scientifique – 95 rue de la Galéra, 34095 Montpellier, France

[3]Institut des Sciences de lÉvolution [Montpellier] (ISEM) – Université de Montpellier, Institut de recherche pour le développement [IRD] : UR226, Centre National de la Recherche Scientifique : UMR5554 – Place E. Bataillon CC 064 34095 Montpellier Cedex 05, France

Abstract

Mosquitoes are human infectious disease vectors that have been extensively studied, not only because of the high genetic diversity their genomes manifest, but also for their remarkably strong capacity of fast adaptation, such as climate changes or insecticide resistance. While several studies of genes known to be involved in adaptation help to shed light on the putative role of transposable elements (TEs) in such evolution process, the impact of TEs on mosquito genome structure and evolution are still poorly tackled (Assogba *et al.*, 2016). Here, we carry out the study on TE abundance and distribution in mosquito genomes. In March 2017, a research group ended up with the first chromosome-length scaffolds in *Culex pipiens quinquefaciatus* and *Aedes aegypti*. We decided to start focusing on the new version of the *Cx. pipiens* genome assembly (*CpipJ3*) (Dudchenko *et al.*, 2017).

According to TE evolutionary models, we may expect to observe an enrichment of TEs in regions poor in genes and regions of reduced recombination. To test this hypothesis, we started developing a new tool to estimate the recombination rates along chromosomes based on Marey maps (Fiston-Lavier *et al.*, 2010) (Rezvoy *et al.*, 2007). Our tool includes a statistical-based approach for the detection of the heterochromatin boundaries that automatically re-adjusts estimates in regions with a depletion of fitness between the polynomial and the data. After assessing the veracity of the tool with experimental data from *Anopheles gambiae* (Sharakhova et *al.*, 2010), we estimated the recombination rate along the *Cx. pipiens* new assembly.

On the other hand, we annotated individual TE insertions in *Cx. pipiens*. We built a *Culex* specific TE library, a set of canonical sequences representative of TE families in this genome. We then annotated them combining results from homology-based (TEfam database: https://tefam.biochem.vt.edu/) and signature-based approaches. We reported a high diversity with TE families from the three main types of TEs (DNA, LTR, non-LTR).

*Speaker

†Corresponding author: anna-sophie.fiston-lavier@umontpellier.fr

The annotation of individual TE insertions in *CpipJ3* reveals a higher TE content compared with previous studies (33% instead of 29% for *CpipJ2*). Our results also showed a non-homogenous distribution of TEs along the *Cx. pipiens* chromosomes with an enrichment of TEs in the heterochromatin. In-depth analysis of the TE organization is currently in process. Our results should help explaining the *Cx. pipiens* genome structure but also assessing the quality of the new release of the assembly.

How transposable elements shape mosquito genomes

Yasmine Mansour[1,2,3], Mickaël Hamouma[1,2,3], Annie Chateau[1,2], Anna-Sophie Fiston-Lavier[3]*

[1] Laboratoire d´Informatique de Robotique et de Microélectronique de Montpellier (LIRMM) – Université de Montpellier, CNRS - Montpellier, France

[2] Institut de Biologie Computationnelle (IBC) – CIRAD, INRA, INRIA, CNRS - Montpellier, France

[3] Institut des Sciences de l'Evolution - Montpellier (ISEM) – Université de Montpellier, IRD, CNRS - Montpellier, France

*Contact : anna-sophie.fiston-lavier@umontpellier.fr

Introduction

Mosquitoes are human infectious **disease vectors** that have been extensively studied, not only because of the high genetic diversity their genomes manifest, but also for their remarkably strong capacity of fast adaptation, such as climate changes or **insecticide resistance**.

While several studies of genes known to be involved in **adaptation** help to shed light on the putative role of **transposable elements (TEs)** in such **evolution** process, the impact of TEs on mosquito **genome structure** and evolution are still poorly tackled.

Here, we carry out the study on **TE abundance** and **distribution** in mosquito genomes.

What are TE abundance and distribution in mosquito genomes ?

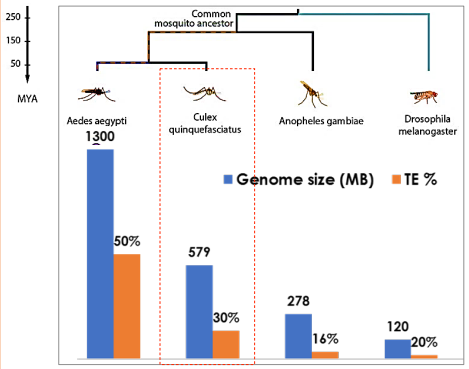


Figure 1. Dudchenko, et al. 2017 adapted [1]

We start focusing on the new version of the ***Culex pipiens quinquefaciatus*** genome assembly (*CpipJ3*) which provides the first chromosome-length scaffolds[1].

Methods

Recombination Rate estimation : Marey maps approach

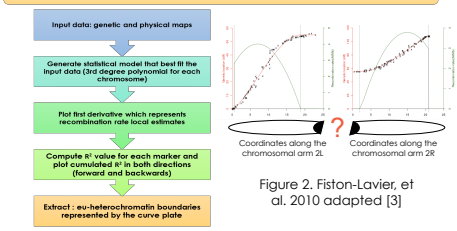
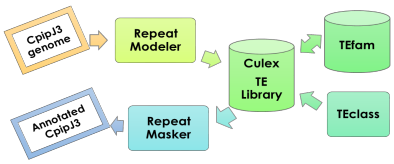


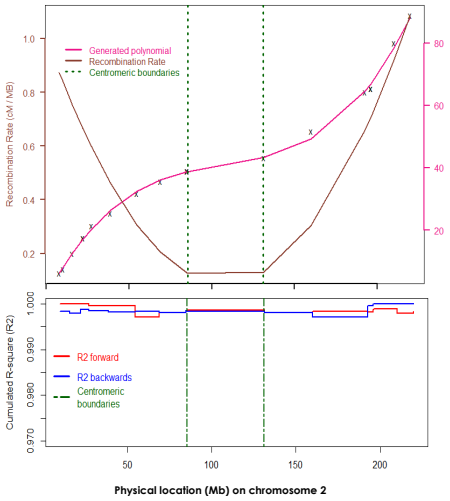
Figure 2. Fiston-Lavier, et al. 2010 adapted [3]

TE Annotation pipeline

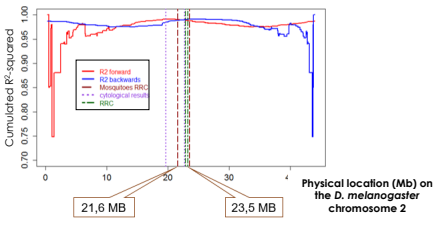


Results

Recombination rates estimation along Cx. *Pipiens* chromosomes



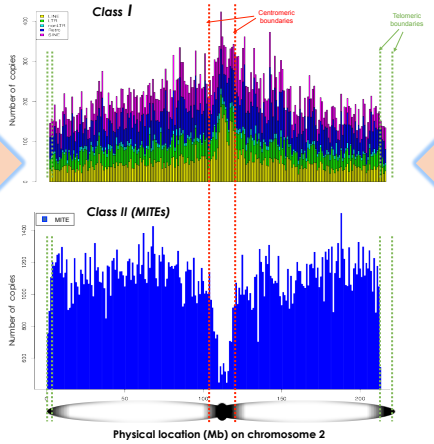
Validation of the heterochromatc boundary detection



Comparison of our method with previous ones in *Drosophila melanogaster*:

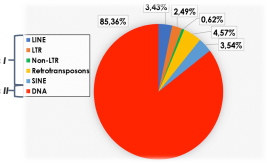
Methods for the heterochromatin boundary estimates (Mb)	Cytological results [5]	RRC [3]	Mosquitoes RRC (Our method)
2L	19.67	22.88	21.60
2R	22.74	23.28	23.50

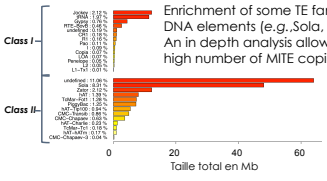
The TE distribution highlights centromeric boundaries



A genome enriched in DNA elements

Using our pipeline with the new release of the genome assembly (*CpipJ3*), we re-estimated the TE content in this genome. We ended up with 33% of TEs that is greater than previous estimates (30%).





Enrichment of some TE families mostly DNA elements (e.g. *Sola*, Zator or TC1). An in depth analysis allows identifying a high number of MITE copies (>10 000).

Conclusion

- **Automated and optimized statistical tool for the estimation of the recombination rate along the chromosomes:**
  - Validation of our approach with experimental results on *Drosophila melanogaster*
  - TE distribution confirms our eu-heterochromtic boundaries
- **A genome enriched in DNA elements:**
  - New TE content estimate (33%)
  - More than two-third of the genome is composed of DNA elements
  - Around 75% of the DNA elements are MITEs
  - Some families known to be active are more highly represented than non-active families
- **TE family distribution suggests insertion bias in Cx. *pipiens* genome:**
  - RNA elements are enriched in heterochromatin while DNA elements are preferentially located in euchromatin
  - LINE element distribution shows a strong bias in centromeric regions
  - MITE elements are the main TEs in euchromatin

Taking all together, our preliminary work stands in stark contrast with previous estimates and suggests that TEs have played a much larger role in shaping Cx. *pipiens* genome than previously believed.

References

1. Dudchenko, O., Batra, S. S., Omer, A. D., Nyquist, S. K., Hoeger, M., Durand, N. C. et al. (2017). *De novo* assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science*, *356*(6333), 92–95.
2. Ment Rezvoy, C., Charif, D., Gué Guen, L., & Marais, G. A. B. (2007). MareyMap: an R-based tool with graphical interface for estimating recombination rates, 23(16), 2188–2189.
3. Fiston-Lavier, A. S., Singh, N. D., Lipatov, M., & Petrov, D. A. (2010). *Drosophila melanogaster* recombination rate calculator. Gene, 463(1–2), 18–20.
4. Comeron, J. M., Ratnappan, R., & Bailin, S. (2012). The Many Landscapes of Recombination in *Drosophila melanogaster*. PLoS Genetics, 8(10), e1002905.
5. Bergman, C. M., Quesneville, H., Anxolabéhère, D., & Ashburner, M. (2006). Recurrent insertion and duplication generate networks of transposable element sequences in the Drosophila melanogaster genome. Genome Biology, 7(11), R112.
6. Abrusán, G., Grundmann, N., Demester, L., & Makalowski, W. (2009). TEclass - A tool for automated classification of unknown eukaryotic transposable elements. *Bioinformatics*, *25*(10), 1329–1330.
7. http://www.repeatmasker.org/
8. http://www.repeatmasker.org/RepeatModeler/
9. https://tefam.biochem.vt.edu/tefam/

Acknowledgements

We thank the Algerian Government and particularly the Ministry of Higher Education and Scientific Research for funding this thesis.

Thanks to all the members of the EVAS (ISEM-CIAWOL) and MAB (LIRMM) teams.

Organization of insect genomes driven by active transposable element families

Yasmine MANSOUR[1,2], Mickael HAMOUMA[1,2], Annie CHATEAU[1,3], and Anna-Sophie FISTON-LAVIER[2]

[1]The Montpellier Laboratory of Informatics, Robotics and Microelectronics *(LIRMM)*, Montpellier, France
[2]Institute of Evolution Science of Montpellier *(ISEM)*, Montpellier, France
[3]The Computational Biology Institute *(IBC)*, Montpellier, France

Transposable elements (TEs) have been rapidly gained in insect species such as the P-element, which invaded the worldwide Drosophila melanogaster populations in less than 50 years. Such feature makes TEs as good markers of recent evolution processes, such as adaptation. Unfortunately, the impact of TEs on the structure and the evolution of insect genomes is still poorly tackled mostly because of the low-quality genome assemblies. Here, we investigated the TE organization in two insect genomes: Culex pipiens, a recent genome assembly and, D. melanogaster, offering high-quality genome assemblies and annotations. The TE distribution patterns across genomes generally show enrichment in particular areas such as constitutive heterochromatic showing a low recombination and low gene density. As no recombination rate estimates were available in Cx. pipiens as in most of the insect genomes, we developed a statistical approach that generates broad-scale maps of recombination by fitting a third-order polynomial to each chromosome arm based on genomic and physical maps. This new approach offers several functionalities to remove low-quality genomic map data, assess the quality of the data (i.e. number and repartition of the markers along the genome). Our approach automatically re-adjusts estimates in regions with a depletion of fitness between the polynomial and the data to estimate the heterochromatin boundaries. We validated our approach in D. melanogaster. We then estimated the recombination rate along the Cx. pipiens genome and identifed the heterochromatin boundaries. After the annotation of TEs in Cx. pipiens, we analyzed the relationship between TEs and recombination. In both species, we observed non-homogenous distributions for active TE families such LINE and MITE. In Cx. pipiens, while LINEs are enriched in pericentromeric regions, MITEs are richer in euchromatin. To attempt to explain such distribution bias, we investigated the TE dynamics for these two TE families launching a comparative genomic approach in other insect genomes.

Organization of insect genomes driven by some transposable element families

ine Mansour[1,2,3], Mickaël Hamouma[1,2,3], Annie Chateau[2,3], Anna-Sophie Fiston-Lavier[1*]

[1]Institut des Sciences de l'Evolution - Montpellier (ISEM) – Université de Montpellier, IRD, CNRS - Montpellier, France

oire d´Informatique de Robotique et de Microélectronique de Montpellier (LIRMM) – Université de Montpellier, CNRS - Montpellier, France

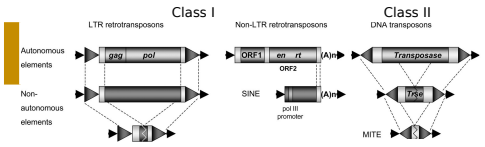
[3]Institut de Biologie Computationnelle (IBC) – CIRAD, INRA, INRIA, CNRS - Montpellier, France

*Contact : anna-sophie.fiston-lavier@umontpellier.fr

Co-Evolution between Transposable Elements (TEs) and Recombination

Transposable elements (TEs)

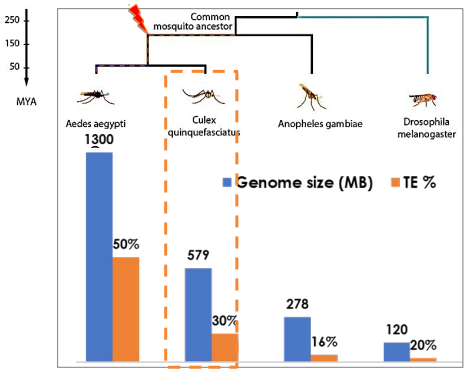
TEs are mobile DNA, mostly dispersed repeats, highly repetitive sequences and detected in almost all the organisms sequenced so far. TEs were classified in two classes (Class I and Class II), superfamilies and families based on their transposition mechanism and sequence features. More and more studies continue to support the role of such repeated elements in genome evolution.



[source: Wessler, PNAS 2006 adapted]

Enrichment of TEs in heterochromatic regions

Previous studies showed a strong and negative correlation between TE distribution and recombination overall. The TE distribution patterns across genomes generally show an enrichment in particular areas such as constitutive heterochromatic harboring low recombination and low gene density. The TE distribution is the consequence of both TE insertion bias and natural selection against deleterious TE insertions.



[Source Dudchenko, et al. 2017 adapted]

Recombination

Recombination consists on the exchange of DNA fragments within and between chromosomes. This evolutionary force guarantees the diversity of genetic material over generations. Recombination varies among species and along chromosomes. Such heterogeneity may impact the levels of diversity, the efficiency of selection, and by

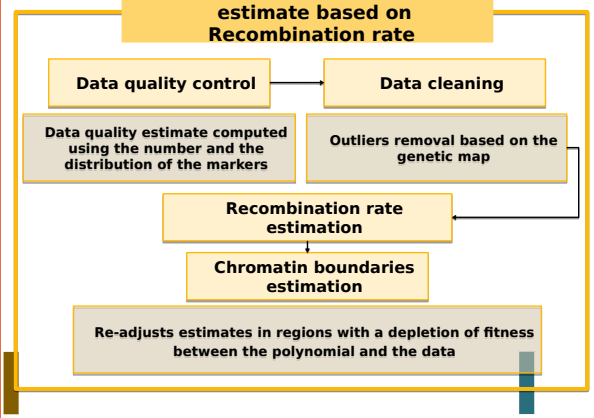
TE content increase since the divergence from the *An. gambiae* lineage

To understand how TEs are distributed along genomes and decipher their association with recombination, we investigated a genomic comparative study among the taxonomic group of disease-vector mosquitoes with *Drosophila melanogaster* as outgroup. The increase of the TE content since the divergence from the *An. gambiae* lineage suggests an increased level of TE activity and/or weaker force of selection against TE insertions in the two culicinae lineages. We thus first choose to focus on TEs and recombination in *Culex pipiens,* recently re-assembled (*CpipJ3*) [Dudchenko et al 2017].

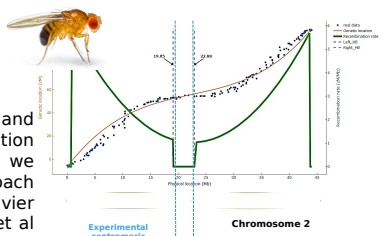
Automatic recombination rates and chromatin boundaries estimates

We started developing a statistical R package called **BRec** that generates broad-scale maps of recombination based on genetic and physical maps. Our package offers several functionalities to estimate automatically and in a more accurate way the recombination rates along entire chromosomes [*Mansour et al 2018, in prep*].

BRec: Chromatin Boundaries estimate based on Recombination rate



BRec validation with *Drosophila melanogaster*

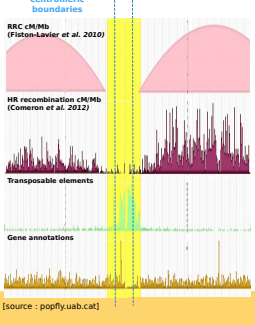


Using previous broad and fine-scale recombination rate estimates, we validated our approach (Pval<0.05) [Fiston-Lavier et al 2010; Comeron et al 2012].

Combining experimental and other genomic features, we also appreciate the precision of the estimations of the chromatin boundaries [flybase.org: Release 5; Fiston-Lavier et al 2009].

Coordinates of heterochromatin boundaries on each chromosome	Cytological results from [7] converted* to Release 5 (Mb)	Our results (Mb)
2	[19.67 - 22.74]	[19.05 - 22.88]
3	[20.55 - 24.62]	[21.17 - 23.39]
X	20.37	19.53

[* flybase.org/convert/coordinates]

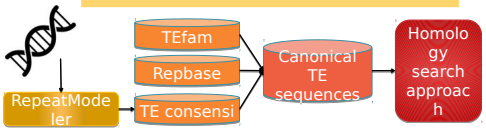


[source : popfly.uab.cat]

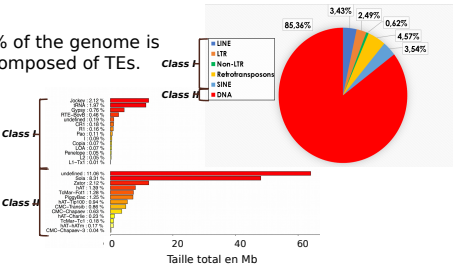
Cx. Pipiens : a genome enriched in DNA elements

We then re-annotated all TE insertions and analyze the TE distribution taking into account the chromatin boundaries defined in the last release of the *Cx. pipiens* genome (*CpipJ3*).

A genome enriched in MITEs



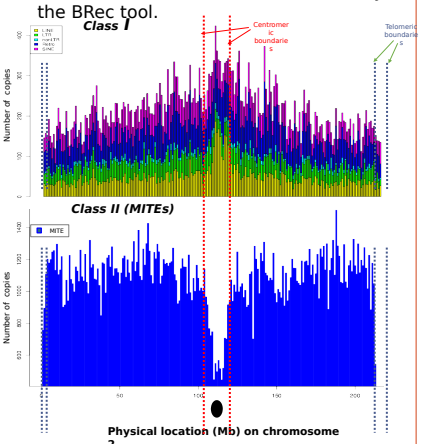
33% of the genome is composed of TEs.



Enrichment in DNA elements mostly due to some families:
Sola, Zator, TC1 and more specifically MITE (>10 000 copies).

Distribution bias for some TE families

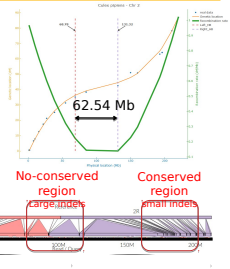
The TE distribution supports the centromeric boundaries estimated by the BRec tool.



While LINEs are enriched in centromeric regions, a paucity of MITEs is observed in heterochromatin.

Accumulation of active elements mainly into centromeres

Chromatin boundaries estimates suggest in *Cx. pipiens* an invasion of genetic novelties mainly centromeric regions



Cx. pipiens genomes

Conclusion

BRec, a new R package

We provide here an automated tool for the estimation of the chromatin boundaries based on the recombination rate.

A vast majority of DNA elements

Almost one-third of the genome is composed of DNA elements, mostly MITEs. Such non-autonomous and short elements do not appear as deleterious elements as they are mainly located in euchromatic regions.

TE invasion in centromeres

Previous studies revealed retrotransposons as the dominant TEs in mosquitoes [Arensburger et al 2010]. Our finding support this observation. We observed a strong insertion bias of LINEs in centromeric regions.

This highlights an increased level of TE activity in *Cx. pipiens* through specific TE families. If true, we may also expect to observe a similar pattern of TE activity in *Ae. Aegypti*. However, we cannot exclude a reduced intensity of selection against TE insertions. The estimations of the TE activity should help discriminate between these two hypotheses.

Acknowledgements

We thank the Algerian Government, the Ministry of Higher Education and Scientific Research and the Labex CeMEB for funding this thesis. We also thank Remy Costa for its help for plotting the conservation regions.

How to improve genome assembly using repetitive elements.

Quentin Delorme * [1], Annie Chateau† [2,3], Anna-Sophie Fiston-Lavier [4], Yasmine Mansour [5,6,7]

[1] Laboratoire dÍnformatique de Robotique et de Microélectronique de Montpellier (LIRMM) – Université de Montpellier : UMR5506, Centre National de la Recherche Scientifique : UMR5506 – 161 rue Ada - 34095 Montpellier, France

[2] Laboratoire d'Informatique de Robotique et de Microélectronique de Montpellier (LIRMM) – Université Montpellier II - Sciences et techniques, CNRS : UMR5506 – CC 477, 161 rue Ada, 34095 Montpellier Cedex 5, France

[3] Institut de Biologie Computationnelle (IBC) – CNRS : UMR5506, Université Montpellier II - Sciences et techniques – 95 rue de la Galéra, 34095 Montpellier, France

[4] Institut des Sciences de l'Evolution - Montpellier (ISEM) – CNRS : UMR5554, Institut de recherche pour le développement [IRD] : UMR226 – Place E. Bataillon CC 064 34095 Montpellier Cedex 05, France

[5] Institut des Sciences de lÉvolution [Montpellier] (ISEM) – Université de Montpellier, Institut de recherche pour le développement [IRD] : UR226, Centre National de la Recherche Scientifique : UMR5554 – Place E. Bataillon CC 064 34095 Montpellier Cedex 05, France

[6] Institut de Biologie Computationnelle (IBC) – Centre de Coopération Internationale en Recherche Agronomique pour le Développement, Institut National de la Recherche Agronomique, Institut National de Recherche en Informatique et en Automatique, Université de Montpellier, Centre National de la Recherche Scientifique – 95 rue de la Galéra, 34095 Montpellier, France

[7] Laboratoire dÍnformatique de Robotique et de Microélectronique de Montpellier (LIRMM) – Université de Montpellier : UMR5506, Centre National de la Recherche Scientifique : UMR5506 – CC 477, 161 rue Ada, 34095 Montpellier Cedex 5, France

Repetitive DNA sequences are abundant in almost all species: RRs (Repetitive Regions) may represent up to 90% of genome size [1]. Despite being a fundamental source of genomic diversity and novelty, RRs are responsible of assembly errors yielding bad quality of genome assemblies [2]. Even with advanced high-throughput sequencing technologies, genome assembly is facing a big challenge towards achieving its optimum quality. While reads assembly overcome this issue, often by collapsing or excluding repeats from contigs, scaffolding step ought to handle RRs.

The perspective of this work is to detect, classify and use misassemblies due to RRs to improve genome assemblies. Our hypothesis is that some RRs like Transposable Elements (TEs) are more disruptive elements in the face of genome

assembly process than others, due to their biology. We intend to test whether the assembly errors are more likely caused by long and young TE insertions [3]. We are currently working on Anopheles gambiae's reference genome. Anopheles gambiae is the principal vector of malaria, a disease that afflicts more than 500 million people and causes more than 1 million deaths each year. Improving assemblies may lead to a better understanding of his genome's dynamic and

*Speaker

†Corresponding author: annie.chateau@lirmm.fr

appearance of insecticide resistance. We intend to exploit sequence similarities between repeats family on a three-step process :

- A first step consists to investigate how information on TEs obtained independently of the assembly, could limit their disruptive effects. Using CENSOR [4], we are able to detect different types of RRs dans tag them on contigs.

- In a second step, we put together contigs clusters based on labeled RRs families. This step is meant to reduce possibilities of misjunction between contigs holding two different kind of RRs.

- In each cluster each combinaison of two contigs, leading to the formation of hypothetic scaffolds, is querying against the repeat database Repbase. Thus, scaffolds can be validate by matching with an existing repeat region, leading to the reconstruction of the original sequence.

The aim is to generate scaffold graph from those RRs informations. This graph could be different than scaffold graph based on paired-end reads informations. Here, the challenge will be to confront orientation informations from both graph and try to resolve hypothetic conflict. Algorithmic approach will be developped for evaluation of information relevance.

C. Biemont. A brief history of the status of transposable elements: from junk DNA to major players in evolution. Genetics, 186(4):1085{1093, Dec 2010.

H.Tang. Genome assembly, rearrangement, and repeats. Chemical Reviews, 107:3391{3406, 2007.

Rajiv C. McCoy, Ryan W. Taylor, Timothy A. Blauwkamp, Joanna L. Kelley, Michael Kertesz, Dmitry Pushkarev, Dmitri A. Petrov, and Anna-Sophie Fiston-Lavier. Illumina truseq synthetic long-reads empower de novo assembly and resolve complex, highly-repetitive transposable elements. PLOS ONE, 9(9):1{13, 09 2014.

J.Urka et al. Censor - a program for identication and elimination of repetitive elements from dna sequences. Computers and Chemistry, 20:119{122, 1996.

Keywords: Scaffolding, Transposable Elements, Repetitive Regions

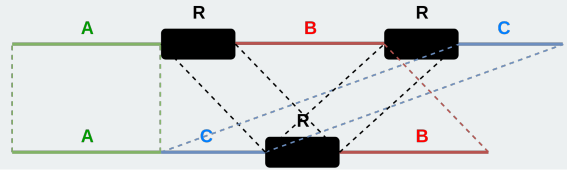
HOW TO IMPROVE GENOME SCAFFOLDING USING REPETITIVE REGIONS ?

Quentin DELORME[1,2,3], Yasmine MANSOUR[2,3], Anna-Sophie FISTON-LAVIER*[3], Annie CHATEAU*[2]

1: Master « Sciences et Numérique pour la Santé » , parcours « Bioinformatique, Connaissances, Données », Université de Montpellier
2: LIRMM – Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier, Montpellier
3: ISEM – Institut des Sciences de l'Évolution de Montpellier, Montpellier
*: These authors contributed equally to this work

Impact of Repetitive Regions (RRs) on quality of genome assemblies

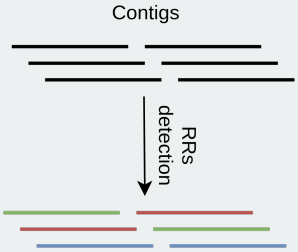
Repetitive DNA sequences[1] are abundant in almost all species: RRs (Repetitive Regions) may **represent up to 90 % of genome size**[1]. Despite being a fundamental source of genomic diversity and novelty, RRs are reponsible of assembly errors like misarrangments or sequence skipping, yielding **bad quality of genome assemblies** (for more details, see poster #91 in this session).



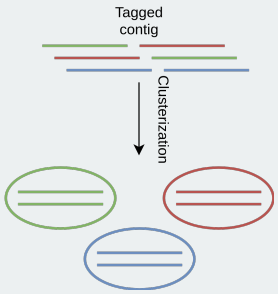
We are currently working on ***Anopheles Gambiae***'s reference genome, **which presents about 20 % of RRs**[2]. *Anopheles Gambiae* is the principal vector of malaria, a disease that afflicts more than 500 millions people and causes more than one million deaths each year. Improving assemblies may lead to a better understanding of its genome's dynamic and appearance of insecticide resistance.
The perspective of this work is to detect, classify and use misassemblies due to RRs to improve genome assemblies.

Repetitive Regions detection & caracterization:a three-step process

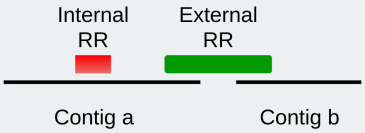
1/ Using **CENSOR**[3], we are able to **detect RRs on contigs.** Censor is based on RRs database **Repbase**[4] and identifies repeats by sequence homology. Each contig is characterized by repetitives region's name and position(s) on contig.



2/ We then **clusterize the contigs** based on their RR annotation**.**

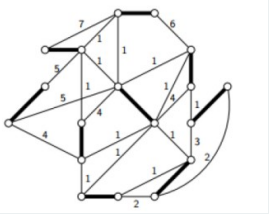


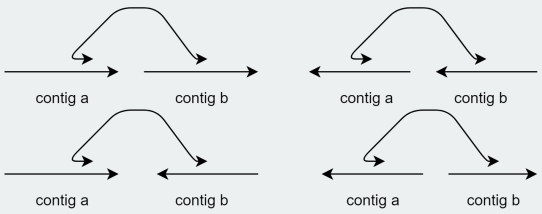
3/ **For each cluster**, **RR's position** on the contigs is evaluated as **internal or external** and contigs sharing a same external RR will be associated.



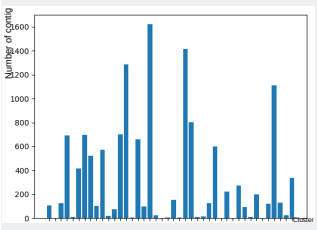
Using Repetitive Regions information to improve scaffolding graph

Our dataset from *Anopheles gambiae*'s genome is constituted by **43 000 contigs** among which **13 000 bear repeats**. **Alignment of paired-ends reads on contigs** leads to the generation of a **scaffold graph**. Here, **bold edges** represent **contigs**, **vertices** represent the **ends of the contig** and **thin edges** represent the **link between contigs**. The **score** represents the **number of paired reads supporting** the link. Impact of RR on this graph causes distorded support scores whose lead to reconstruction errors.

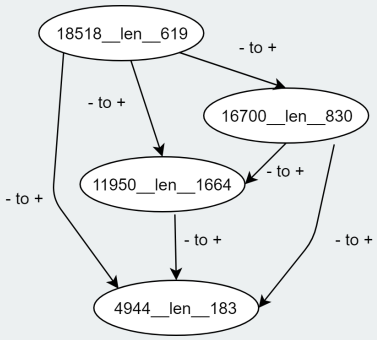




Our data process led to the **constitution of 43 clusters** of different sizes: **the more contigs are numerous, the more overlaps of external RRs can occur.**
Also, we try to infer **link between presence of RR and multiplicity of contig**. This information may lead to improvement of scaffolding quality. To this end, we evaluate multiplicity of contigs according to the presence or absence of RRs. We observe twice more overlapping external RRs on multiple contigs than on other contigs. This confirms that external RRs are great candidates to improve scaffolding.



Multiplicity / RRs type	Yes	No
Any	33.47 %	31.77 %
External	13.63 %	6.68 %



For each cluster, connected contigs are associated in a new sort of scaffold graph.
In our graph, **vertices are contigs** and **edges represent repeats overlapping contigs**. These edges are polarized according to contig's orientation.

Perspectives

The aim is **not to replace paired-end graph but to complete it:** we have to find a way to **conciliate paired-end and RRs informations**. The challenge will be to confront orientation information from both graphs and try to resolve hypothetic conflicts.

Bibliography

1 Tang,H., « Genome assembly, rearrangement and repeats », *Chemical Reviews* 107:3391-3406 (2007) ;
2 Holt,RA *et al.* « The genome sequence of the malaria mosquito Anopheles Gambiae », *Science* 298:129-149 (2002) ;
3 Pavlicek, A , Kohany, O. , Jurka, J. . « Repeat mining:basic tools for detection and analysis » *Analytic Tools for DNA, genes and genomes nuts and bolts* (2005);
4 Jurka, J *et al.* «Repbase Update, a database of eukaryotic repetitive elements », Cytogenetic and Genome Research 110:462-467 (2005)

This project was supported by the Labex CeMEB incorporated into the I-site MUSE

In-depth analysis of the impact of transposable elements on genome assembly quality

Rémy Costa *† [1], Yasmine Mansour‡ [2,3], Annie Chateau§ [3,4], Anna-Sophie Fiston-Lavier¶ [2]

[1] Institut des Sciences de lÉvolution de Montpellier (ISEM) – Université de Montpellier, Institut de recherche pour le développement [IRD] : UR226, Centre National de la Recherche Scientifique : UMR5554 – Place E. Bataillon CC 064 34095 Montpellier Cedex 05, France

[2] Institut des Sciences de l'Evolution - Montpellier (ISEM) – CNRS : UMR5554, Institut de recherche pour le développement [IRD] : UMR226 – Place E. Bataillon CC 064 34095 Montpellier Cedex 05, France

[3] Laboratoire d'Informatique de Robotique et de Microélectronique de Montpellier (LIRMM) – Université Montpellier II - Sciences et techniques, CNRS : UMR5506 – CC 477, 161 rue Ada, 34095 Montpellier Cedex 5, France

[4] Institut de Biologie Computationnelle (IBC) – CNRS : UMR5506, Université Montpellier II - Sciences et techniques – 95 rue de la Galéra, 34095 Montpellier, France

Genome assembly has become crucial for conducting genomic studies in various field as environment, health, genetics, evolution and many more. Recent studies highlighted the impact of assembly quality on result interpretations [1]. While efficiency of bioinformatic tools used for assembly is increasing, errors of sequence construction from contigous short reads persist. One of the known sources of errors is repeated elements.
The presence of repeated elements can induce (i) chimeric contigs due to collapsed repeats and (ii) assembly breaks. Among repeated elements, transposable elements (TEs) are ubiquitous sequences, *i.e.* detected in the vast majority of sequenced genomes, and make up for a large fraction of them (*e.g.* up to 90% for the maize genome) [2]. A variety of TEs can be identified. They are classified according to their transposition mecanisms and sequence properties [3].

The recently sequenced and assembled genome of *Ambystoma mexicanum* (Mexican axolotl) shows that up to 97% of contigs encompass TEs at their ends. Analysis of these TEs showed that they are recent (sharing a high sequence identity) and abundant (present in numerous copies). Such active TEs mainly correspond to a specific group : LTR retrotransposons [4]. Even if advanced sequencing technologies has improved assembly quality such as long read sequencing, no short read based approaches allow investigating in-depth analysis of disruptive TEs. We expect TE-rich genomes to be harder to assemble, and specific type of TEs to cause more errors than others. Recent and long TEs with a high copy number should induce more assembly biases. As TEs do not insert homogenously in the genome, we also expect regions

*Speaker
†Corresponding author: remy.costa@etu.umontpellier.fr
‡Corresponding author: yasmine.mansour@umontpellier.fr
§Corresponding author: annie.chateau@lirmm.fr
¶Corresponding author: anna-sophie.fiston-lavier@umontpellier.fr

enriched in TEs to be more challenging to assemble.

Here we aim to test our hypotheses by estimating the impact of TEs on assembly quality through identifying the most disruptive TE types and analyzing the impact of TE density on the assembly quality. For that, we will use an approach based on assembly simulation by controlling TE features in the *Drosophila melanogaster* genome. This genome harbors one of the highest quality genomic sequences and annotations. Our results should help improving the process of genome assembly by taking advantage of the TE information.

Mahul Chakraborty, Nicholas W. Vankuren, Roy Zhao, Xinwen Zhang, Shannon Kalsow, and J. J. Emerson. Hidden genetic variation shapes the structure of functional elements in Drosophila. Nature Genetics, 50(1) :20–25, 2018.

Dario Copetti and Rod A. Wing. The Dark Side of the Genome : Revealing the Native Transposable Element/Repeat Content of Eukaryotic Genomes. Molecular Plant, 9(12) :1664–1666, 2016.

Thomas Wicker, François Sabot, Aurélie Hua-Van, Jeffrey L. Bennetzen, Pierre Capy, Boulos Chalhoub, Andrew Flavell, Philippe Leroy, Michele Morgante, Olivier Panaud, Etienne Paux, Phillip SanMiguel, and Alan H. Schulman. A unified classification system for eukaryotic transposable elements. Nature Reviews Genetics, 8(12) :973–982, 2007.

Sergej Nowoshilow, Siegfried Schloissnig, Ji Feng Fei, Andreas Dahl, Andy W.C. Pang, Martin Pippel, Sylke Winkler, Alex R. Hastie, George Young, Juliana G. Roscito, Francisco Falcon, Dunja Knapp, Sean Powell, Alfredo Cruz, Han Cao, Bianca Habermann, Michael Hiller, Elly M. Tanaka, and Eugene W. Myers. The axolotl genome and the evolution of key tissue formation regulators. Nature, 554(7690) :50–55, 2018.

Keywords: transposable elements, genome assembly, high, throughput sequencing, NGS

IN-DEPTH ANALYSIS OF THE IMPACT OF TRANSPOSABLE ELEMENTS ON GENOME ASSEMBLY QUALITY

Rémy Costa[1,2], Yasmine Mansour[2,3], Annie Chateau[3,4] , Anna-Sophie Fiston-Lavier[1,2]

1. Master Sciences et Numérique pour la Santé, parcours Bioinformatique, Connaissances, Données, Université de Montpellier
2. Institut des Sciences de l'Evolution - Montpellier (ISEM)
3. Laboratoire d'Informatique de Robotique et de Microélectronique de Montpellier (LIRMM)
4. Institut de Biologie Computationnelle (IBC), Montpellier

Transposable Elements (TEs), source of assembly errors

Genome assembly has become crucial for conducting genomic studies in various field. Recent studies highlighted the impact of assembly quality on result interpretations (Chakraborty *et al.*, 2018). While efficiency of bioinformatic tools used for assembly is increasing, errors of sequence construction from contigous short reads persist. One of the known sources of errors is **repeated elements (cf Fig 1)**.

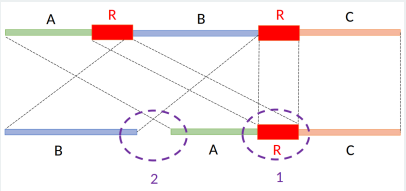


Figure 1. *Assembly errors due to repeated elements. The two copies of the same repeat are represented as red rectangles (R).*

The presence of repeated elements can induce :
(1) chimeric contigs due to collapsed repeats
(2) assembly breaks

Among repeated elements, transposable elements (TEs) are ubiquitous sequences, *i.e.* detected in almost all of genomes sequenced so far, and make up for a large fraction of them (Copetti and Rod, 2016). A variety of TEs can be identified. They are classified according to their transposition mecanisms and sequence properties in TE types (DNA, LINE, LTR, SINE; Wicker *et al.*, 2007). TEs do not insert homogenously in the genome.
Because of the biases they can induce, we expect TE-rich genomes and regions to be harder to assemble, and specific type of TEs to cause more errors than others. It is then important to determine the most disruptive TEs and how to assess their impact.

A large fraction of contigs flanked by TEs

The recently sequenced and assembled genome of ***Ambystoma mexicanum*** (Nowoshilow *et al.*, 2018), using an approach combining long-read sequencing (PacBio), optical mapping and a genome assembler (MARVEL), revealed a 32Gb genome with a high proportion of repetitive sequences (65.6% of the contig assembly, representing 18.6Gb). TEs represent the largest fraction of these repeated elements.

Fig 2. *Axolotl – (Malta National Aquarium)*

Analysis of TEs showed that they are recent (sharing a high sequence identity), abundant (present in numerous copies) and including elements of more than 10kb in length, corresponding to a specific group : Long Terminal Repeats retrotransposons, also called LTRs. Such long elements represent a challenge for assembly, as **97% of contigs encompass LTR at their ends.**

Even if advanced sequencing technologies has improved assembly quality such as long read sequencing, they remain expensive. No short read based approaches allow investigating in-depth analysis of disruptive TEs. Thus, we need to estimate and caracterise the impact of TEs to assess the quality of the assembly.

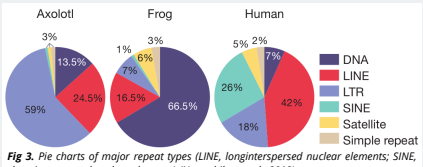


Fig 3. *Pie charts of major repeat types (LINE, longinterspersed nuclear elements; SINE, short interspersed nuclear elements) (Nowoshilow et al., 2018)*

Advanced sequencing technologies improved high-repeat density regions

We used ***Drosophila melanogaster*** genome as it harbors one of the highest quality genomic sequences and annotations to assess the evoution of the assembly quality. The last D. melanogaster assembly (Release 6) version is improved by physical mapping, cytogenetic map--ping, and sequence finishing.

Figure 4a. *Dot plot of Chromosome 2L – Release 6 versus itself*

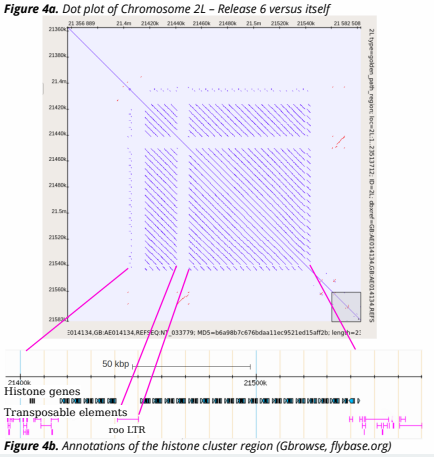


Figure 4b. *Annotations of the histone cluster region (Gbrowse, flybase.org)*

To illustrate the improvment of the genome assemblies through time, we selected one high-repeat density region : the **cluster of histone genes**.

This cluster located on the 2L chromosom arm of 21.5Mb is composed of 23 tandem units (Fig 4). This cluster is located in the centromeric region, known to be challenging to assemble.

We then compared the same region on older releases *versus* release 6 to illustrate this evolution. However, most of genome sequences and assemblies so far do not reach such high quality, reinforcing the necessity to estimate the impact of TEs.

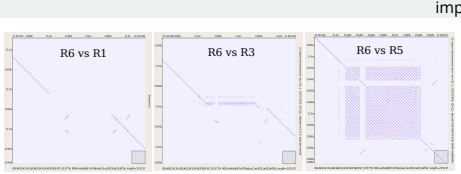


Figure 5a. *Dot plots comparing releases of D. melanogaster genome showing the impact of repeated elements and the evolution of the assembly quality (x-axis : R6 ; dot plot realised using the program Ugene).*

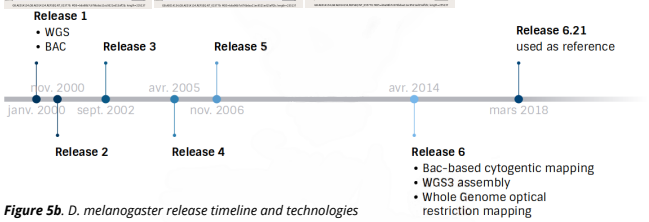


Figure 5b. *D. melanogaster release timeline and technologies used to assemble the genome.*

Estimating the impact of TEs on genome assembly

1) How to estimate the impact of TEs on genome assembly?

Different approaches can be used based on the data available : (i) without reference sequences, we can analyse the contig ends (see axolot study); (ii) with high quality reference sequences, we can launch a **genomic comparative study** focusing on the misalignment regions. Using Mummer (Kurtz *et al.* 2004), an alignment package dedicated to large DNA sequences, we identified the sequences breakpoints and aimed to estimate the number of TE sequences in the vincinity of these breakpoints on chromosome arm 2L.

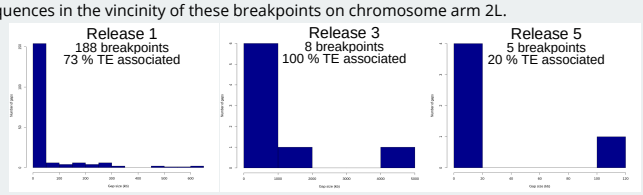


Figure 6. *Histograms of the number of gaps by size related to TEs, for Releases 1, 3 and 5 compared to Release 6.*

2) Can we characterize this association between the breakpoints and TEs ? What are the more disruptive TEs ?

We expected young and long TE sequences to be the most disruptive elements such as LTR elements. To test this hypothesis, we analyzed the TE sequences associated to the breakpoints (type, length, copy number, age). Our analyses support the disruptive effect of LTR elements. However, the sequencing technologies help reducing their impact.

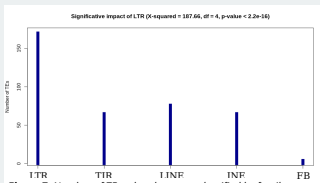


Figure 7. *Number of TEs related to gaps, classified by family*

Discussion / Perspectives

Althought the sequencing and assembly technics have been improved, our preliminary findings support a high impact of TEs on genome assembly. The status for most of the genomes sequenced so far is « draft », thus closer to the releases 1 and 3 than the release 6 of *D. melanogaster*. As TEs are ubiquitous, we may expect to identify the same impact for most of the genomes.

Some TEs are more disruptive than others. In several studies (like ours), LTR elements are often emphazed as disruptive (synonyme) genomic elements as they are still active elements. We also show that by combining short-reads, long-reads and optical mapping, it is possible to drastically reduce the effect of TE (data not show). However, such approach is costly and time-consuming.

To go further, we are elaborating an approach based on assembly simulation by controlling TE features. Analysis of the impact of TEs on genome assembly will allow to propose new approaches in order to improve genome assembly. TE informations can be infered to the scaffolding (*see poster #27*).

This project was supported by the Labex Cemeb (ERJ) incorporated into the I-site MUSE

How to involve repetitive regions in scaffolding improvement

Rémy Costa[1,4], Quentin Delorme[1,2], Yasmine Mansour[1,2,3], Anna-Sophie Fiston-Lavier[1,3] and Annie Chateau[1,2]

[1] Université de Montpellier

[2] Laboratoire d'Informatique, Robotique et Micro-électronique de Montpellier

[3] Institut des Sciences de l'Évolution de Montpellier

[4] Master Science et Numérique pour la Santé, parcours Bioinformatique, Connaissances, Données

Corresponding author: remy.costa@etu.umontpellier.fr

Abstract

Context and motivation. Repetitive regions (RR) in DNA sequences are present in almost all organisms and may represent over 80% of the genome size. Fundamental source of genetic plasticity and diversity, yet they are a source of complication when it comes to assemble genomes [1]. Assembly produces contigs of various sizes, sometimes really smaller than the original chromosome size. To reduce the fragmentation of chromosomes, the scaffolding process involves additional information, for instance pairing between reads, to infer how contigs are relatively organized [2]. Repetitive regions are disturbing both assembly and scaffolding processes, which are based on graphs. One way to untangle ambiguous parts of these graphs is to use long reads, produced by third-generation sequencing technologies. However, this is not always possible due to high cost and lower quality. Here we propose to use RR sequences themselves to enhance the scaffolding step.

Methodology. The scaffold graph is defined as follows: vertices represent contig extremities, while edges are of two kinds: (1) contig edges, linking both extremities of a contig, and (2) inter-contig edges relating the pairing-information. A weight function on the inter-contig edges indicates how many pairs are supporting this edge. Due to repeats, some of the inter-contigs edges are erroneous and have to be removed from the graph. In other cases, they are supported by RR. Our method is based on a pipeline progressively refining inter-contig edges through RR analysis, described as follows:

1. find the known RR sequences using a repeat database [3], map them on contigs, tag the contigs with this information, and cluster them according to these tags;
2. inside each cluster, determine inter-contig edges sharing coherent RR sequence parts;
3. modify the weight of the validated inter-contig edges;
4. delete edges incoherent with RR composition or length;
5. after scaffolding, use the RR canonical sequence to fill the gaps between contigs.

An additional knowledge about well-documented RRs (such as Transposable Elements) may help to improve Step 2, and answer the following question: do assembly errors come essentially from recent RRs ? Step 3 can be achieved in different ways, thus we propose to try several weight function perturbations. Step 4 is quite expeditious and may be smoothed by introducing a probabilistic measure to ponder the inter-contig weight instead of deleting it.

Validation. The benchmark is composed of organisms offering different repetition rates and sizes. To validate our approach, we use simulated data from model species, amongst them very high quality genomes such as *Drosophila melanogaster* and *Caenorhabditis elegans*. We will carefully examine the influence of each decision step, in the previous pipeline, on the final quality of the scaffolded genome. Also, an analysis will be driven on deleted edges to determine the relevance of this step and calibrate the probabilistic measure. Genome quality will be measured using the QUAST tool [4].

References

[1] Haixu Tang. Genome assembly, rearrangement, and repeats. *Chemical Reviews*, 107(8):3391–3406, 2007.

[2] Martin Hunt, Chris Newbold, Matthew Berriman, and Thomas D. Otto. A comprehensive evaluation of assembly scaffolding tools. *Genome Biology*, 15(3):R42, Mar 2014.

[3] Weidong Bao, Kenji K. Kojima, and Oleksiy Kohany. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob DNA*, 6:11, 2015.

[4] Alla Mikheenko, Andrey Prjibelski, Vladislav Saveliev, Dmitry Antipov, and Alexey Gurevich. Versatile genome assembly evaluation with quast-lg. *Bioinformatics*, 34(13):i142–i150, 2018.

HOW TO INVOLVE REPETITIVE REGIONS IN SCAFFOLDING IMPROVEMENT

Rémy COSTA[1,2,4], Quentin DELORME[1,2], Yasmine MANSOUR[1,2,3], Anna-Sophie FISTON-LAVIER[1,3] and Annie Chateau[1,2]

1. Université de Montpellier
2. Laboratoire d'Informatique de Robotique et de Microélectronique de Montpellier (LIRMM)
3. Institut des Sciences de l'Évolution - Montpellier (ISEM)
4. Master Sciences et Numérique pour la Santé, parcours Bioinformatique, Connaissances, Données ; Université de Montpellier

Impact of repetitive regions on genome reconstruction

Repetitive DNA sequences are present in almost all organisms, and repetitive regions (RR) may **represent up to 90 % of the genome size**. RRs are a source of genetic plasticity and diversity, but they are responsible of complications when it comes to genome reconstruction, like chimeric contigs due to collapsed repeats (1) or assembly breaks (2) [Fig. 1], **reducing the quality of the assembled genome**.

Assembly produces contigs of various sizes, sometimes really smaller than the original chromosome size. To reduce the fragmentation of chromosomes, the scaffolding process involves additional information, for instance pairing between reads to infer how contigs are relatively organized. RRs are disturbing both assembly and scaffolding processes, which are based on graphs. **The perspective of this work is to use RR sequences to enhance the scaffolding step.**

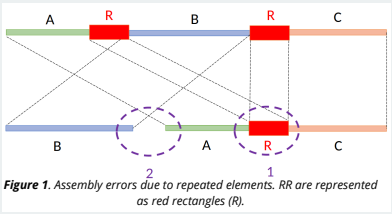
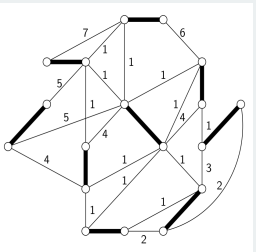


***Figure 1**. Assembly errors due to repeated elements. RR are represented as red rectangles (R).*

Validation protocol

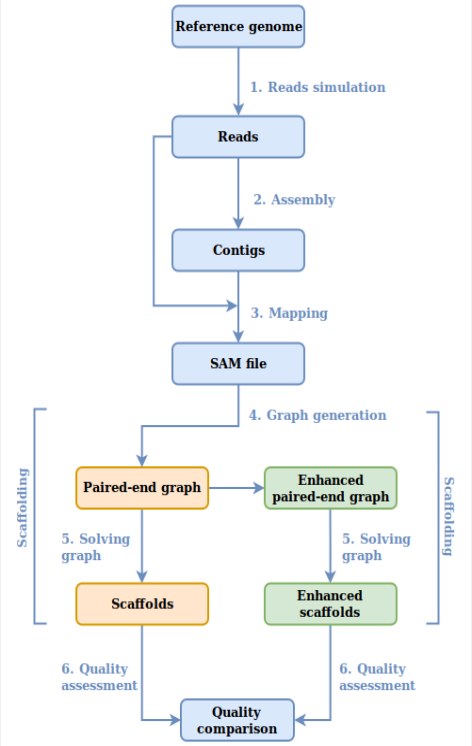
Paired-end Graph

Scaffolds are build using mapping of paired-end reads on contigs. In the scaffold graph below, vertices represent contig extremities. **Contig edges** (bold line) **link both ends of acontig**. **Inter-contig edges** (thin line) relate the **pairing-information**. A weight function on the inter-contig edges indicates the number of paired-end reads supporting this edge.



Graph enhancement

RRs may in some cases induce erroneous support scores in inter-contig edges leading to reconstruction errors and have to be eliminated from the graph.
Our method is based on a **pipeline refining inter-contig edges through RR analysis**, described as follows:
1. **identifiy RR sequences** on contigs using a repeat database, map them on contigs, tag the contigs with this information, and cluster them according to RR families;
2. **Determine** inter-contig edges sharing coherent RR sequence parts inside each cluster;
3. **Modify the weight** of the validated inter-contig edges;
4. **Delete edges** incoherent with RR composition or length;



1. Reads simulation

We used **ART** (Weichun *et al.*, 2011) to generate our paired-end reads with a **20X coverage**, simulating **Illumina's HighSeq2000** from high quality genome references: *Drosophila melanogaster* and *Caenorhabditis elegans*.

2. Assembly

To realise the assembly, we used **Minia** (Chikhi & Rizk, 2013) and **Spades** (Bankevich *et. al*, 2012) in order to compare the most effecicient tool.

3. Mapping

The mapping was realised with **BWA** (Li *et al.*, 2009) and **Minimap2** (Li *et al.*, 2018).

4. Generating graphs

We generated the paired-end graph with **Scaftools** (Chateau & Giroudeau, 2014).

5. Solving graphs

The graph solution was also generated with **Scaftools**.

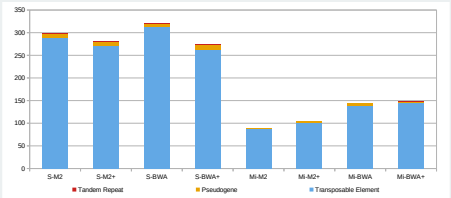
6. Quality assessment

Finally, the quality comparison of the scaffolds obtained in our pipeline was realised using **QUAST-LG** (Mikheenko *et al*, *2018*) with the reference genome.

Results

D. melanogaster	SPAdes				Minia			
	Minimap 2		BWA		Minimap 2		BWA	
		Enhanced		Enhanced		Enhanced		Enhanced
Genome fraction	83.586	83.147	83.564	83.163	82.691	82.357	82.749	82.42
NG50	138 662	129 502	141 803	133 722	120 493	115 298	115 249	114 878
Number of misassemblies	708	**552**	770	**567**	159	**164**	252	**261**

Results show a slight reduction of the covered genome fraction and the NG50, but an **improvement in the reduction of misassemblies up to 26 %** with SPAdes (and no improvement with minia). To analyse further these misassemblies, we aligned them on the reference genome to observe if RR were implicated.



RRs are implicated in 60 to 70 % of the misassemblies. Even if we found some tandem repeats and pseudogenes, the vast majority is composed of **transposable elements**.

Perspectives

We designed an efficient method to reduce the number of misassemblies due to RRs on the scaffolding. **Remaining misassemblies are also mainly due to RRs escaping the method**. The most disturbing type of RRs identified are young and active transposable elements (TEs).

We are currently working on solutions to adress this issue and furhter increase the quality of the reconstruction. Our first goal is to **confront the distance information of the paired-end reads between two contigs to the length of the RR detected between at their extremities.** If the information distance concurs, this method will allow us to inferhe sequence between two contigs with the consensus sequence of the RR.

We also wish to smoothen Step 4 of the pipeline by **introducing a probabilistic measure to ponder the inter-contig weight** instead of deleting it.

We'll analyze the deleted edges to determine the relevence of this step and calibrate the probabilistic measure.

Mansour *et al. BMC Bioinformatics* (2021) 22:396
https://doi.org/10.1186/s12859-021-04233-1

BMC Bioinformatics

RESEARCH Open Access

BREC: an R package/Shiny app for automatically identifying heterochromatin boundaries and estimating local recombination rates along chromosomes

Yasmine Mansour[1,2*], Annie Chateau[2] and Anna-Sophie Fiston-Lavier[1,3*]

From 19th International Conference on Bioinformatics 2020 (InCoB2020)
Virtual. 25-29 November 2020

*Correspondence:
yasmine.mansour@umontpellier.fr; anna-sophie.fiston-lavier@umontpellier.fr
[1] Genomics Department, Institute of Evolution Science of Montpellier (ISEM), Montpellier, France
Full list of author information is available at the end of the article

Abstract

Background: Meiotic recombination is a vital biological process playing an essential role in genome's structural and functional dynamics. Genomes exhibit highly various recombination profiles along chromosomes associated with several chromatin states. However, eu-heterochromatin boundaries are not available nor easily provided for non-model organisms, especially for newly sequenced ones. Hence, we miss accurate local recombination rates necessary to address evolutionary questions.

Results: Here, we propose an automated computational tool, based on the Marey maps method, allowing to identify heterochromatin boundaries along chromosomes and estimating local recombination rates. Our method, called **BREC** (heterochromatin **B**oundaries and **REC**ombination rate estimates) is non-genome-specific, running even on non-model genomes as long as genetic and physical maps are available. BREC is based on pure statistics and is data-driven, implying that good input data quality remains a strong requirement. Therefore, a data pre-processing module (data quality control and cleaning) is provided. Experiments show that BREC handles different markers' density and distribution issues.

Conclusions: BREC's heterochromatin boundaries have been validated with cytological equivalents experimentally generated on the fruit fly *Drosophila melanogaster* genome, for which BREC returns congruent corresponding values. Also, BREC's recombination rates have been compared with previously reported estimates. Based on the promising results, we believe our tool has the potential to help bring data science into the service of genome biology and evolution. We introduce BREC within an R-package and a Shiny web-based user-friendly application yielding a fast, easy-to-use, and broadly accessible resource. The BREC R-package is available at the GitHub repository https://github.com/GenomeStructureOrganization.

Keywords: Heterochromatin regions, Centromere position, Recombination rate, Non-genome-specific, Data quality control, Graphical user interface

Background

Meiotic recombination is a vital biological process that plays an essential role in investigating genome-wide structural and functional dynamics. Recombination events are observed in almost all eukaryotic genomes. Crossover, a one-point recombination event, is the exchange of DNA fragments between sister chromatids during meiosis. Recombination is a fundamental process that ensures genotypic and phenotypic diversity. Thereby, it is strongly related to various genomic features such as gene density, repetitive DNA, and DNA methylation [1–3].

Recombination rate varies not only between species but also within species and along chromosomes. Different heterochromatin regions exhibit different profiles of recombination events. Therefore, in order to understand how and why the recombination rate varies, it is vital to break down the chromosome structure into smaller blocks where several genomic features, besides recombination rate, are also known to exhibit different profiles. Chromatin boundaries allow to distinguish between two primary states of chromatin that can be defined as euchromatin, which is lightly compact with a high gene density, and on the contrary, heterochromatin, which is highly compact with a paucity in genes. The heterochromatin is represented in different chromosome regions: the centromere and the telomeres. Euchromatin and heterochromatin regions exhibit different behaviors in terms of genomic features and dynamics related to their biologic function, such as the cell division process that ensures the organism viability. Consequently, easily distinguishing chromatin states is necessary for conducting further studies in various research fields and to be able to address questions related to cellular processes such as meiosis, gene expression, epigenetics, DNA methylation, natural selection and evolution, genome architecture and organization, among others [4–6]. In particular, the profound understanding of centromeres, their complete and precise structure, organization, and evolution is currently a hot research area. These repeat-rich heterochromatin regions are currently still either poorly or not assembled at all across eukaryote genomes. Despite the enormous advances offered by the Next Generation Sequencing (NGS) technologies, centromeres are still considered enigmas, mostly because they prevent genome assembly algorithms from reaching their optimal performance to achieve more complete whole genome sequences [7]. Besides, the highly diverse mechanisms of heterochromatin positioning [8] and repositioning [9] remain a complicated obstacle in the face of fully understanding genome organization. Thus, generating high resolution genetic, physical, and recombination maps and locating heterochromatin regions is increasingly attractive to the community across an extensive range of taxa [10–16].

Numerous methods for estimating recombination rates exist. Genomic inference methods, covering population-based, pedigree-based and gamete-based approaches, have been included in the latest review by [17]. Among the listed methods, population genetic-based methods [18] provide accurate fine-scale estimates. Nevertheless, these methods are costly, time-consuming, require substantial expertise, and most of all, do not apply to all kinds of organisms. Moreover, the sperm-typing method [19], which is also extremely accurate, providing high-density recombination maps, is male-specific

and is applicable only on limited genome regions. On the other hand, a purely statistical approach, the Marey Maps [20], could avoid some of the above issues based on other available genomic data: the genetic and physical distances of genomic markers.

The Marey maps approach consists of correlating the physical map with the genetic map representing respectively physical and genetic distances for a set of genetic markers on the same chromosome. Despite the efficiency of this approach and mostly the availability of physical and genetic maps, generating recombination maps rapidly and for any organism is still challenging. Hence, the increasing need for an automatic, portable, and easy-to-use solution.

Some Marey map-based tools already exist, two of which are primarily used. The MareyMap Online [21, 22] applies to multiple species, yet, it does not allow an accurate estimate of recombination rates on specific regions like the chromosome extremities. Second, the *Drosophila melanogaster* Recombination Rate Calculator (RRC) [23] solves the previous issue by adjusting recombination rate estimates on such chromosome regions, but as indicated by its name, it is *D. melanogaster*-specific. With the emerging NGS technologies, accessing whole chromosome sequences has become possible on a wide range of species. Therefore, we may expect an exponential increase in the markers number, requiring more adapted tools to handle such new scopes of data efficiently.

Here, we propose a new Marey map-based method as an automated computational solution that aims to, firstly, identify heterochromatin boundaries (HCB) along chromosomes, secondly, estimate local recombination rates, and lastly, adjust recombination rates on chromosome along the chromosomal regions marked by the identified boundaries. Our proposed method, called **BREC** (heterochromatin **B**oundaries and **REC**ombination rate estimates), is provided within an R-package and a Shiny web-based graphical user interface. BREC takes as input the same genomic data, genetic and physical distances, as in previous tools. It follows a workflow (see Fig. 1) that, first, tests the data quality and offers a cleaning option, then estimates local recombination rates and identify HCB. Finally, BREC re-adjusts recombination rate estimates along heterochromatin regions, the centromere and telomere(s), in order to keep the estimates as authentic as possible to the biological process [24]. Identifying the boundaries delimiting euchromatin and heterochromatin allows investigating recombination rate variations along the whole genome, helping to compare recombination patterns within and between species. Furthermore, such functionality is fundamental for identifying the position of the centromeric and telomeric regions. Indeed, the position of the centromere along the chromosome has an influence on the chromatin environment, and recent studies are interested in investigating how genome architecture may change with centromere organization [7].

Our results have been validated with cytological equivalents, experimentally generated on the fruit fly *D. melanogaster* genome [4, 25, 26]. Moreover, since BREC is non-genome-specific, it could efficiently be run on other model as well as non-model organisms for which both genetic and physical maps are available. Even though it is still an ongoing study, BREC has also been tested with different species, and the results are reported.

This paper is organized as follows: the set of our results, based on both simulated and real data, are reported in "Results" section. They are then discussed in

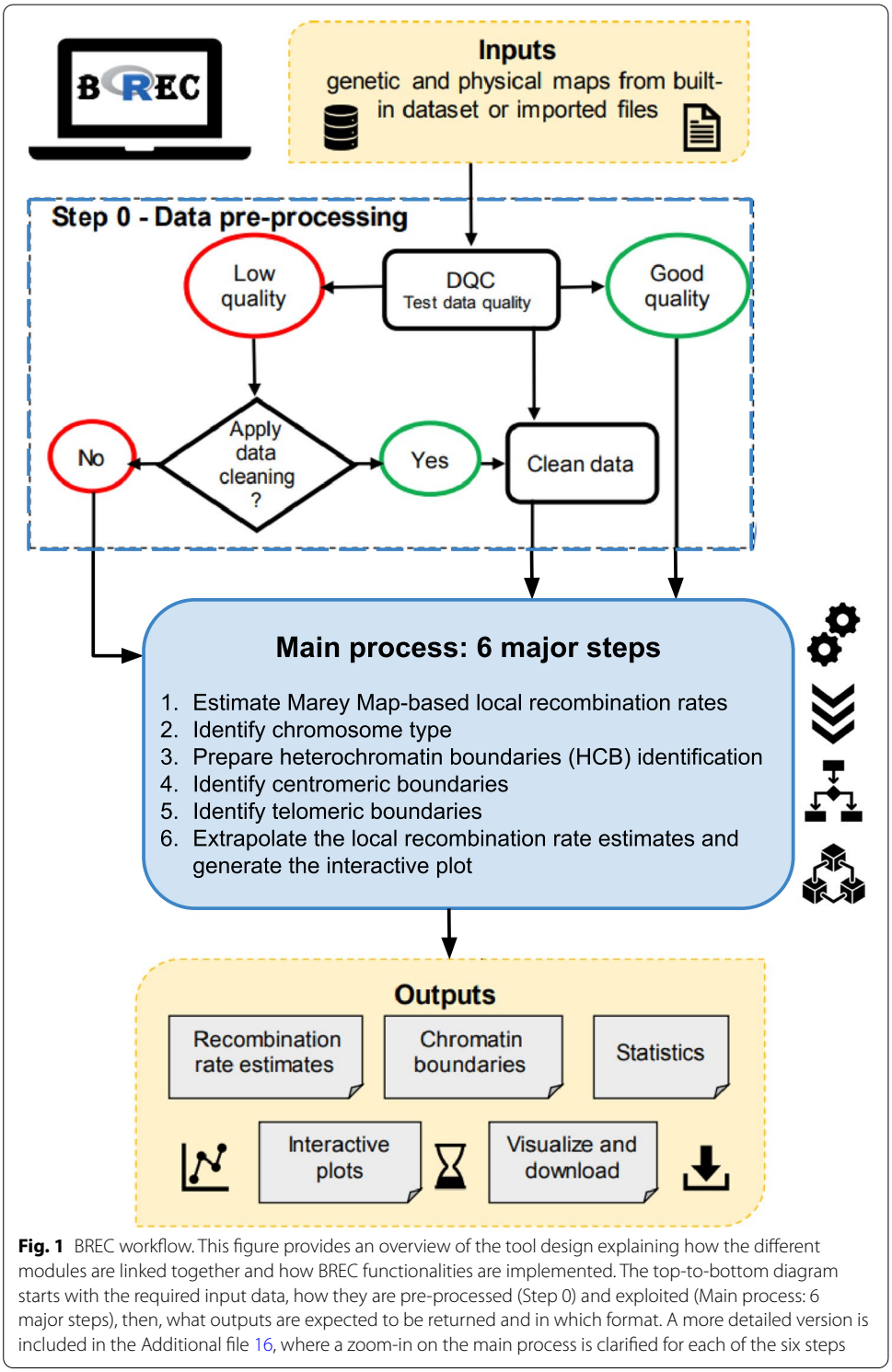


Fig. 1 BREC workflow. This figure provides an overview of the tool design explaining how the different modules are linked together and how BREC functionalities are implemented. The top-to-bottom diagram starts with the required input data, how they are pre-processed (Step 0) and exploited (Main process: 6 major steps), then, what outputs are expected to be returned and in which format. A more detailed version is included in the Additional file 16, where a zoom-in on the main process is clarified for each of the six steps

"Discussion" section. Concluding remarks with some perspectives are outlined in "Conclusions" section. The full set of BREC modules, detailed within a step-by-step workflow, as well as further details on the data involved, and how the methods were calibrated and validated, are presented in "Methods" section. Additional files: 1, 3, 4,

5, 7, 8, 9, 10, 11, 12, 13, 14, 16, 18, 19 consist of Figures S1–S15, and Additional files: 2, 6, 15, 17, 20, 21 include Tables S1-S6).

Results

In this section, we present the results obtained through the following validation process. First, we automatically re-identified HCB with an approximate resolution to the reference equivalents. Second, we tested the robustness of BREC methods according to input data quality, using the well-studied *D. melanogaster* genome data, for which recombination rate and HCB have already been accurately provided [4, 23, 25, 27] (Additional file 1). Besides, we extended the robustness test to a completely different genome, the domesticated tomato *S. lycopersicum* [28] to better interpret the study results. Even if the Loess span value does not impact the HCB identification, but only the resulting recombination rate estimates, the span values used in this study are: 15% for *D. melanogaster* (for comparison purpose) and 25% for the rest of the experiments. Our analysis shows that BREC is applicable to data from various organisms, as long as the data quality is good enough. BREC is data-driven, thus, the outputs strongly depend on the markers density, distribution, and chromosome type identified (automatically, or with the user's a priori knowledge).

Approximate, yet congruent HCB

Fruit fly genome D.melanogaster

Our approach for identifying HCB has been primarily validated with cytological data experimentally generated on the *D. melanogaster* Release 5 genome [4, 25, 26, 29]. For all five chromosomal arms (X, 2L, 2R, 3L, 3R). This genome presents a mean density of 5.39 markers/Mb and a mean physical map length of 22.92Mb. We obtained congruent HCB with a good overlap and shift, distance between the physical position of the reference and BREC, from 20Kb to 4.58Mb (see "Data and implementation" section). We did not observe a difference in terms of mean shift for the telomeric and centromeric BREC identification ($\chi^2 = 0.10$, df $= 1$, $p - value = 0.75$)(See Table 1 and Additional file 2). We observe a lower resolution for the chromosomal arms 3L and 3R (see Additional file 3). This suggests that those two chromosomal arms' data might not present as good quality as the rest of the genome. Interestingly, the local markers density for these two chromosomal arms shows a high variation, unlike the other chromosomal arms. For instance, the 2L for which BREC returns accurate results, shows a lower variation (see Additional file 4). Without these two arms, the max shift for both centromeric and telomeric BREC boundaries is smaller than 1.54Mb, with a mean shift decreasing from 1.43 to 0.71 Mb.

This first analysis suggests that BREC methods return accurate results on this genome. However, the boundaries identification process appears very sensitive to the markers' local density and distribution along a chromosome (see Additional file 3). Therefore, we conducted further experiments on a different dataset, the tomato genome (see Additional file 5).

Tomato genome S. lycopersicum

Results of experimenting BREC behaviour on all 12 chromosomes of *S. lycopersicum* genome [28] are shown as values in Additional file 6 and as plots in Additional file 7. This genome presents a mean density of 2.64 markers/Mb and a mean physical map length of 62.71Mb. We observe a variation in the shift value representing the difference on the physical map between reference HCB and their equivalents returned by BREC. Unlike the *D. melanogaster* genome, which is of a smaller size, with five telocentric chromosomes (chromosomal arms) and a strongly different markers distribution, the tomato genome exhibits a completely different study case. It is a plant genome, with approximately 8-fold bigger genome size. It is organized as twelve atelocentric chromosomes of a mean size of 60Mb, except for chromosomes 2 and 6, which are more likely to be rather considered telocentric based on their markers distribution. Also, we observe a long plateau of markers along the centromeric region with lower density than the rest of the chromosomes. Something which highly differs from *D. melanogaster* data. We believe all these differences between both genomes give a good validation and evaluation for BREC behavior towards various data quality scenarios. Furthermore, since BREC is a data-driven tool, these experiments help analyze data-related limitations that BREC could face while resolving differently. From another point of view, BREC results on the tomato genome highlight the fact that markers distribution along heterochromatin regions, in particular, strongly impacts the identification of eu-heterochromatin boundaries, even when the density is of 2 markers/Mb or more.

Consistency despite the low data quality

We aim in this part to study to what extent BREC results are depending on the data quality.

BREC handles low markers density

We started by assessing the markers' density on the BREC estimates. We generated simulated datasets with decreasing fractions of markers for each chromosomal arm (from 100% to 30%). For that, we randomly selected a fraction of markers, 30 times, and computed the mean shift between BREC and the reference telomeric and centromeric boundaries. We have noted that BREC's resolution decreases drastically with the fraction and therefore with the marker density (see Additional file 8). However, BREC results appeared stable until 70% of the data for all the chromosomal arms, more specifically for the telomeric boundary detection. Only for the centromeric boundary of the chromosomal arm 3R, we observed the opposite pattern: BREC returns more accurate telomeric boundary estimates when the markers' number decreases. This supports the low quality of the data around the 3R centromere.

This simulation process allowed to set a minimum density threshold representing the minimum value for data density in order to guarantee accurate results for BREC estimates at 5 markers/Mb (fraction of around 70% of the data) on average in *D. melanogaster*. This analysis also supports the fact that because the markers' density alone can not explain the BREC resolution, BREC may also be sensitive to the marker distribution.

Additional file 4 clearly shows that markers' density varies within and between the five chromosomal arms with a mean of 4 to 8 markers/Mb. The variance is induced by the extreme values of local density, such as 0 or 24 markers/Mb on the chromosomal arm X. Still, the overall density is around 5 markers/Mb for the whole genome.

BREC handles heterogeneous distribution

Along chromosomes, genetic markers are not homogeneously distributed. Therefore, to assess the impact of the distribution of markers on BREC results, we designed different *data scenarios* regarding a reference data distribution (see "Simulated data for quality control testing" section). We choose as reference the chromosomal arms 2L and 2R of *D. melanogaster* as we have obtained the most accurate results with their data. After the concatenation of the two arms, we ended up with a metacentric simulated chromosome as a starting simulation scenario (total physical length of 44Mb). While this length was kept unchanged, markers local density and distribution were modified (see "Simulated data for quality control testing" section and Additional file 9).

One particular yet typical case is the centromeric gap. Throughout our analysis, we consider that a chromosome presents a centromeric gap if its data exhibit a lack of genetic markers on a relatively large region on the physical map. Centromeric regions usually are less accessible to sequence due to their highly compact chromatin state. Consequently, these regions are also hard to assemble, and that is why many genomes have chromosomes presenting a centromeric gap. It is essential to know that a centromeric gap is not always precisely located in the middle of a chromosome. Instead, its physical location depends on the chromosome type (see more details in Additional file 10).

We also assess the veracity of BREC on datasets with variable distributions using simulated data with and without a centromeric gap (see Additional file 9).

For all six simulation datasets, BREC results overlap the reference boundaries. Thus BREC correctly handles the presence of a centromeric gap (see Additional file 9: (a)(c)(e)). BREC remains robust to a non-uniform distribution of markers, under the condition that regions flanking the boundaries are greater than 2 markers/Mb (see Additional file 11). In the case of a non-uniform distribution, BREC resolution is higher when the local density is stronger around heterochromatin regions (see Additional file 9: (c)(d)(e)(f)). This suggests that low density on euchromatin regions far from the boundaries is not especially a problem either.

Accurate local recombination rate estimates

After identifying the HCB, BREC provides optimized local estimates of recombination rate along the chromosome by taking into account the absence of recombination in heterochromatin regions. Recombination rates are reset to zero across the centromeric and telomeric regions regardless of the regression model. To closely compare the third degree polynomial with Loess, using different span values, we experimented with this aspect on *D. melanogaster* chromosomal arms and reported the results in Additional file 12.

To assess the veracity of the recombination rates along the whole genome, we compared BREC results with previous recombination rate estimates (see Fig. 2; [4, 25]).

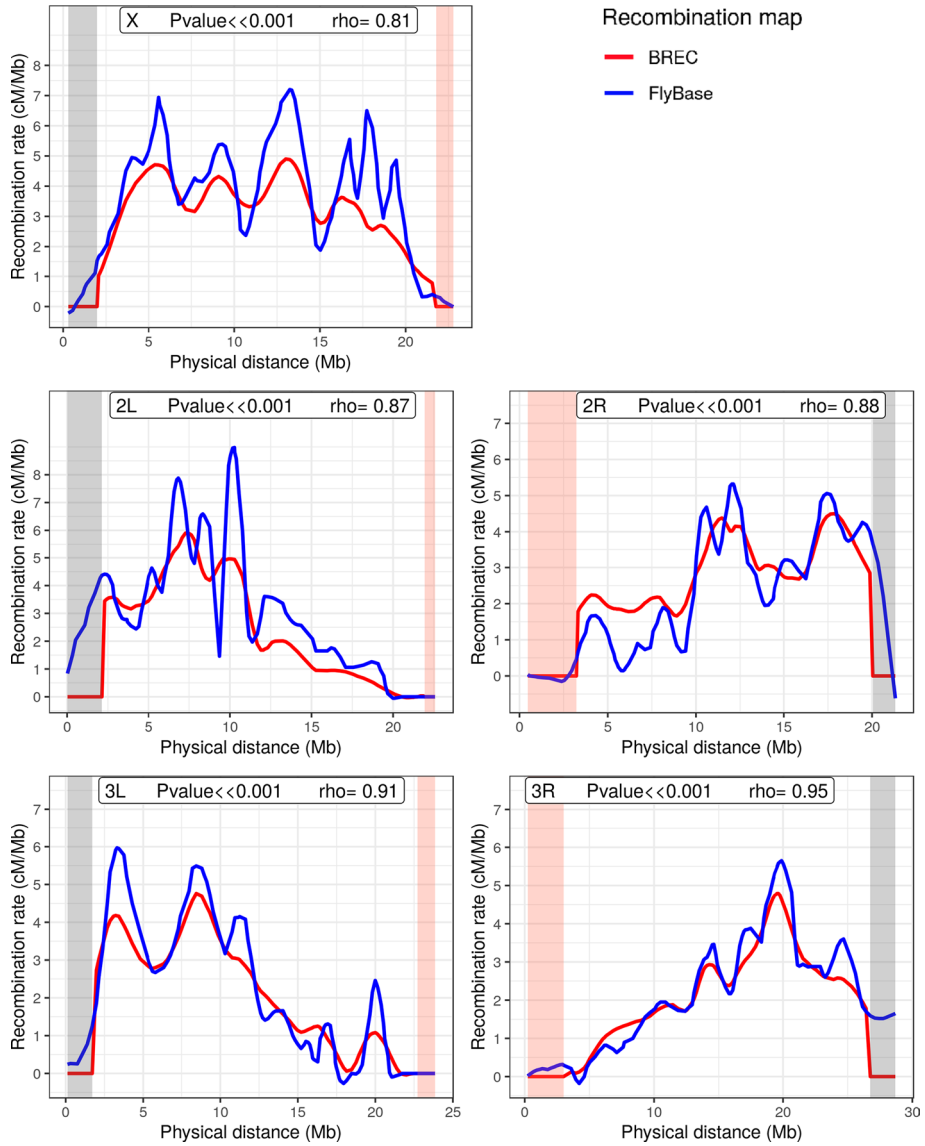


Fig. 2 Comparison of BREC versus FlyBase recombination rate recombination rates along the five chromosomal arms (X, 2L, 2R, 3L, 3R) of D. melanogaster Release 5. Both recombination maps are obtained using the same regression model: Loess with span 15%. The HCB defined by BREC are represented in red and the reference data are in blue. Heterochromatin regions identified by BREC are highlighted in yellow

BREC recombination rate estimates are significantly strongly correlated with reference data (Spearman's: $P \ll 0.001$) while the reference estimates fail in telomeric regions.

BREC is non-genome-specific

NGS, High Throughput Sequencing (HTS) technologies, and numerous further computational advances are increasingly providing genetic and physical maps with more and more accessible markers along the centromeric regions. Such progress in the availability of data of poorly accessible genomic regions is a huge opportunity to shift our knowledge of heterochromatin DNA sequences and their dynamics, as in the case of Transposable Elements (TEs), for example. Therefore, BREC is not identifying centromeric

Fig. 3 Screenshots of BREC web application - Run BREC web page **a** and **b** show the inputs interface. **c** It shows the output of running BREC on the specified inputs, represented with an interactive web-based plot as a result

gaps as centromeric regions as it might seem. Instead, it is targeting centromeric as well as telomeric boundaries identification regardless of the presence or absence of markers neither of their density or distribution variations across such complicated genomic regions (see Additional file 13). Given that BREC is non-genome-specific, applying HCB identification on various genomes has allowed to widen the experimental design and to test more thoroughly how BREC responds to different *data scenarios*. Despite the several challenges due to data quality issues and following a data-driven approach, BREC is a non-genome-specific tool that aims to help to tackle biological questions.

Easy, fast and accessible tool via an R-package and a Shiny app

BREC is an R-package entirely developed with the R programming language. The current version of the package and documentation are available on the GitHub repository: https://github.com/GenomeStructureOrganization.

In addition to the interactive visual results provided by BREC, the package comes with a web-based Graphical User Interface (GUI) build using the shiny and shinydashboard

Table 1 BREC HCB compared to reference boundaries from the reference genome of ***D. melanogaster***

Chromosomal arm	Centromeric (Mb)			Telomeric (Mb)		
	Boundaries		Shift	Boundaries		Shift
	Reference	BREC		Reference	BREC	
X	20.67	20.10	0.56	2.46	0.92	1.54
2L	19.95	20.33	0.38	0.70	0.68	0.02
2R	6.09	5.01	1.08	20.02	20.71	0.69
3L	18.41	20.30	1.90	0.36	2.26	1.91*
3R	8.35	3.77	4.58*	27.25	25.64	1.61
Min. shift	0.38			0.02		
Max. shift	4.58			1.91		
Mean shift	1.70			1.15		
Median shift	1.08			1.54		

The shift is the absolute value of the distance between the BREC and the reference physical heterochromatin boundary. The first five rows represent all chromosomal arms. Grouped columns present reference, BREC and shift values for the centromeric boundaries (Columns 2–4), and for the telomeric boundaries (Columns 4–6). Here the boundary values correspond to the internal HCB. The external boundaries are represented by the physical positions of the first and the last markers of the chromosomes. All values are expressed in Megabase (Mb). The asterisk indicates the largest shift value reported on centromeric and telomeric boundaries separately (see corresponding Additional file 3). The last four rows represent general statistics on the shift value. From top to bottom, they are minimum, maximum, mean, and median respectively. See details on the shift metrics in "Validation metrics" section

libraries. The intuitive GUI makes it a lot easier to use BREC without struggling with the command line (see screenshots in Fig. 3d and Additional file 14).

As for the speed aspect, BREC is quite fast when executing the main functions. We reported the running time for *D. melanogaster R5* and *S. lycopersicum* in Additional files 2 and 15, respectively (plotting excluded). Nevertheless, when running BREC via the Shiny application, and due to the interactive plots displayed, it takes longer because of the plotly rendering. Still, it depends on the size of the genetic and physical maps used, as well as the markers density, as slightly appears in the same tables. The results presented from other species (see Additional file 13) highlight better this dependence.

Discussion

The main two results of BREC are the eu-heterochromatin boundaries and the local recombination rate estimates (see Fig. 2 and Additional file 3).

The HCB algorithm, which identifies the location of centromeric and telomeric regions on the physical map, relies on the regression model obtained from the correlation between the physical distance and the genetic distance of each marker. Then, the goodness-of-fit measure, the R-squared, is used to obtain a curve upon which the transition between euchromatin and heterochromatin is detectable.

On the other hand, the recombination rate algorithm, which estimates local recombination rates, returns the first derivative of the previous regression model as the recombination rates, then resets the derivative values to zero along the heterochromatin regions identified (see Additional file 16).

We validated BREC methods with a reference dataset known to be of high quality: *D. melanogaster*. While two distinct approaches were respectively implemented for the

detection of telomeric and centromeric regions, our results show a similar high resolution (see Table 1 and Additional file 3). Then we analysed BREC's robustness using simulations of a progressive data degradation (see Additional files 8 and 11). Even if BREC is sensitive to the markers' distribution and thus to the local markers' density, it can correctly handle a low global markers' density. For the *D. melanogaster* genome, a density of 5 markers/Mb seems to be sufficient to detect the HCB accurately.

We also validated BREC using the domesticated tomato *S. lycopersicum* dataset (see Additional files 6 and 7). At first glance, one might ask: why validating with this species when the results do not seem really congruent? In fact, we have decided to investigate this genome as it provides a more insightful understanding of the data-driven aspect of BREC and how data quality strongly impacts the heterochromatin identification algorithm. Variations in the local density of markers in this genome are particularly associated with the relatively large plateaued centromeric region representing more than 50% of the chromosome's length. Such *data scenario* is quite different from what we previously reported on the *D. melanogaster* chromosomal arms. This is partially the reason for which we chose this genome for testing BREC limits.

While analyzing the experiments more closely, we found that BREC processes some of the chromosomes as presenting a centromeric gap, while that is not actually the case. Thus, we forced the HCB algorithm to automatically apply the *with-no-centromeric-gap-algorithm*, then, we were inspired to implement this option into the GUI in order to give the users the ability to take advantage of their *a priori* knowledge and by consequence to use BREC more efficiently. Meanwhile, we are considering how to make BREC completely automated regarding this point for an updated version later on. Besides, the reference heterochromatin results we used for the BREC validation are rather an approximate than an exact indicator. The physical positions used as reference correspond to the first and last markers tagged as "heterochromatin" on the spreadsheet file published by the Tomato Genome Consortium authors in [28]. However, we hesitated before validating BREC results with these approximate reference values due to the redundant existence of markers tagged as "euchromatin" directly before or after these reference positions. Unfortunately, we were unable to validate telomeric regions since the reference values were not available. As a result, we are convinced that BREC is approximating well enough in the face of all the disrupting factors mentioned above.

On the other hand, this method's ambition is to escape species-dependence, which means it is conceived to apply to a various range of genomes. To test that, we also launched BREC on genomic data from different species (the house mouse's chromosome 4, roundworm's chromosome 3, and the chromosome 1 of zebrafish). Experiments on these whole genomes showed that BREC works as expected and identifies chromosome types in 95% of cases (see Additional file 13).

One can assume, with the exponential increase of genomic resources associated with the revolution of the sequencing technologies, that more fine-scale genetic maps will be available. Therefore, BREC has quite the potential to widen the horizon of deployment of data science in the service of genome biology and evolution. It will be crucial to develop a dedicated database to store all this data.

BREC package and design offer numerous advantageous functionalities (see Additional file 17) compared to similar existing tools [22, 23]. Thus, we believe our new computational solution will allow a large set of scientific questions, such as the ones raised by the authors of [5, 30], to be addressed more confidently, considering model as well as non-model organisms, and with various perspectives.

Conclusions

We designed a user-friendly tool called BREC that analyses genomes on the chromosome scale, from the recombination point-of-view. BREC is a rapid and reliable method designed to determine euchromatin-heterochromatin boundaries on chromosomal arms or whole chromosomes (resp. telocentric or metacentric). BREC also uses its heterochromatin boundary results to improve the recombination rate estimates along the chromosomes.

Currently, the Shiny app is being deployed on the https://shinyapps.io server, in order to provide an install-free experience to the users. In addition, the "whole genome" version of BREC is a work in progress. It will allow to run BREC on all the chromosomes of a genome of interest at once. This version might also present the identified heterochromatin regions on chromosome ideograms. As short-term perspectives for this work, we may consider extending the robustness tests to additional datasets with high quality and mandatory information (*e.g.* boundaries identified with the cytological method, high quality maps). Retrieving such datasets seems to become less and less complicated. We may also improve the identification of boundaries with a more refined analysis around them, using an iterative multiscale algorithm for instance. Finally, we will be happy to consider the users' feedback and improve our tool's ergonomy and usability. As mid-term perspectives, we underline that BREC could integrate other algorithms aiming to provide further analysis options such as the comparison of heterochromatin regions between closely related species. Also, we are aware that it would be interesting to compare BREC results with more existing methods. Thus, we plan to properly do so in the near future.

Methods

New approach: BREC

BREC is designed following the workflow represented in Fig. 1. To ensure that the broadest range of species could be analyzed by our tool, we designed a pipeline that adapts behavior with respect to input data. Each step of the workflow relies mostly on statistical analysis, adaptive algorithms, and decision proposals led by empirical observation.

The workflow starts with a pre-processing module (called "Step 0") aiming to prepare the data prior to the analysis. Then, it follows six main steps: (1) estimate Marey Map-based local recombination rates, (2) identify chromosome type, (3) prepare the HCB identification, (4) identify the centromeric boundaries, (5) identify the telomeric boundaries, and (6) extrapolate the local recombination rate map and generate an interactive plot containing all BREC outputs (see Fig. 1). Each step is detailed hereafter and summarised in Additional file 16.

Step 0 - Apply data pre-processing

Since we have noticed that BREC estimates are sensitive to the quality of input data, we propose a pre-processing step to assess data quality and suggest an optional data cleaning for outliers. As such, we could ensure proper functioning during further steps.

Data quality control (DQC) The quality of input data is tested regarding two criteria: (1) the density of markers and (2) the homogeneity of their distribution on the physical map along a given chromosome. First, the mean density, defined as the number of markers per physical map length, is computed. This value is compared with the minimum required threshold of 2 markers/Mb. Based on the displayed results, the user gets to decide if data cleaning is required or not. The threshold of 2 markers/Mb is selected based on a simulation process that allowed to test BREC results while decreasing markers density until the observed HCB estimates seemed to be no longer exploitable (see "Simulated data for quality control testing" section). Second, the distribution of input data is tested via a comparison with a simulated uniform distribution of identical markers density and physical map length. This comparison is applied using Pearson's χ^2 test [31], which allows examining how close the observed distribution (input data) is to the expected one (simulated data).

Data cleaning The cleaning step aims to reduce the disruptive impact of noisy data, such as outliers, in order to provide a more accurate recombination rate and heterochromatin boundary results. If the input data fails to pass the Data Quality Control (DQC) test, the user has the option to apply or not a cleaning process. This process consists of identifying the extreme outliers and eliminating them upon the user's confirmation. Outliers are detected using the distribution statistics of the genetic map (see Additional file 18). More precisely, inter-marker distances (separating each two consecutive points) are computed along the genetic map. Using a boxplot, distribution statistics (quartiles, mean, median) are applied on these inter-marker distances to identify outliers, which are chosen as the 5% of the data points with a greater genetic distance than the maximum extreme value, and should be discarded. Thus, the cleaning targets markers for which the genetic distance is quite larger than most of the rest. After the first cleaning iteration, DQC is applied again to assess the new density and distribution. The user can also choose to bypass the cleaning step, but BREC's behavior is no longer guaranteed in such cases.

Step 1 - Estimate Marey Map-based local recombination rates

Once the data are cleaned, the recombination rate can be estimated based on the Marey map [20] approach by: (1) correlating genetic and physical maps, (2) generating two regression models -third degree polynomial and Loess- that better fits these data, (3) computing the prime derivative for both models which will represent preliminary recombination maps for the chromosome. The primary purpose of interpolation here is to provide local recombination rate estimates for any given physical position, instead of only the ones corresponding to available markers.

At this point, both recombination maps are used to identify the chromosome type as well as the approximate position of centromeric and telomeric regions. Nevertheless, as a final output, BREC will return only the Loess-based adjusted map for recombination rates since it provides finer local estimates than the polynomial-based map.

Step 2 - Identify chromosome type

BREC provides a function to identify the type of a given chromosome according to the position of its centromere. This function is based on the physical position of the smallest value of recombination rate estimates, which primarily indicates where the centromeric region is more likely to be located. Our experimentation allowed to come up with the following scheme (see Additional file 10). Two main types are identified: telocentric and atelocentric [32]. Atelocentric type could be either metacentric (centromere located approximately in the center with almost two equal arms) or not metacentric (centromere located between the center and one of the telomeres). The latter includes the two most known subtypes, submetacentric and acrocentric (recently considered types rather than subtypes). It is tricky for BREC to distinguish between submetacentric and acrocentric chromosomes correctly. Their centromeres' position varies slightly, and capturing this variation (based on the smallest value of recombination rate on both maps -polynomial and Loess-) could not be achieved yet. Therefore, we chose to provide this result only if the implemented process allowed to identify the subtype automatically. Otherwise, the user gets the statistics on the chromosome's data and is invited to decide according to further *a priori* knowledge. The two subtypes (metacentric and not metacentric) are distinguished following intuitive reasoning inspired by their definition found in the literature. First, BREC identifies whether the chromosome is an arm (telocentric) or not (atelocentric). Then, it tests if the physical position of the smallest value of the estimated recombination rate is located between 40% to 60% interval. In this case, the subtype is displayed as *metacentric*. Otherwise, it is displayed as *not metacentric*. The recombination rate is estimated using the Loess model ("LOcal regrESSion") [33, 34].

Step 3 - Prepare the HCB identification

The HCB identification is a purely statistical approach relying on the coefficient of determination R^2, which measures how good the generated regression model fits the input data [35]. We chose this approach because the Marey map usually exhibits a lower quality of markers (density and distribution) on the heterochromatin regions. Thus, we aim to capture this transition from high to low quality regions (or vice versa) as it reflects the transition from euchromatin to heterochromatin regions (or vice versa). The coefficient R^2 is defined as the cumulative sum of squares of differences between the interpolation and observed data. R^2 values are accumulated along the chromosome. In order to eliminate the biased effect of accumulation, R^2 is computed twice: $R^2 - forward$ starts the accumulation from the beginning of the chromosome to provide the left centromeric and left telomeric boundaries. In contrast, $R^2 - backwards$ starts from the end of the chromosome, providing the right centromeric and right telomeric boundaries. These R^2 values were calculated using the `rsq` package in R. To compute R^2 cumulative vectors, `rsq` function is applied on the polynomial regression model. In fact, there is no such function for non-linear regression models like the Loess because, in such models, high R^2 does not always indicate a good fit. A sliding window is defined and applied on the R^2 vectors to precisely analyze their variations (see details in the next step). In the case of a telocentric chromosome, the position of the centromere is then deduced as the left or the right side of the arm, while in the case of an atelocentric chromosome, the existence of a centromeric gap is investigated.

Step 4 - Identify centromeric boundaries

Since the centromeric region is known to present reduced recombination rates, the starting point for detecting its boundaries is the physical position corresponding to the smallest polynomial-based recombination rate value. A sliding window is then applied to expand the starting point into a region based on R^2 variations in two opposite directions. The sliding window's size is automatically computed for each chromosome as the largest value of ranges between each two consecutive positions on the physical map (indicated as i and $i+1$ in Eq. 1). After making sure the sliding window includes at least two data points, the mean of local growth rates inside the current window is computed and tested compared to zero. If it is positive (resp. negative) on the forward (resp. backward) R^2 curve, the value corresponding to the window's ending edge is returned as the left (resp. right) boundary. Else, the window moves by a step value equal to its size.

$$sliding\_window\_size(chromosome) = max\{|physPos_{i+1} - physPos_i| : 1 \leq i \leq n-1\} \quad (1)$$

There are some cases where chromosome data present a centromeric gap. Such a lack of data produces biased centromeric boundaries. To overcome this issue, chromosomes with a centromeric gap are handled with a slightly different approach. After comparing the mean of local growth rates regarding to zero, accumulated slopes of all data points within the sliding window are computed, adding one more point at a time. If the mean of accumulated slopes keeps the same variation direction as the mean of growth rates, the centromeric boundary is set as the window's ending edge. Else, the window slides by the same step value as before (equal to its size). The difference between the two chromosome types is that only one sliding window is used for the telocentric case, its starting point is the centromeric side, and it moves away from it. As for the atelocentric case, two sliding windows are used (one on each R^2 curve), their starting point is the same, and they move in opposite directions to expand the centromere into a region.

Step 5 - Identify telomeric boundaries

Since telomeres are considered heterochromatin regions as well, they also tend to exhibit low fitness between the regression model and the data points. More specifically, the accumulated R^2 curve tends to present a significant depletion around telomeres. Therefore, a telomeric boundary is defined here as the physical position of the most significant depletion corresponding to the smallest value of the R^2 curve. As such, in the telocentric case, only one R^2 curve is used. It gives one boundary of the telomeric region (the other boundary is defined by the beginning of the left telomere or the end of the right telomere). Whilst in the atelocentric case, where the are two telomeres, the depletion on $R^2 - forward$ detects the end of the left telomeric region, and the depletion on $R^2 - backwards$ detects the beginning of the right telomeric region. The other two boundaries (the beginning of the left telomere and the end of the right telomere) are defined to be, respectively, the same values of the two markers with the smallest and the largest physical position available within the input data of the chromosome of interest.

Step 6 - Extrapolate the local recombination rate estimates and generate interactive plot

The extrapolation of recombination rate estimates at the identified centromeric and telomeric regions automatically performs an adjustment by resetting the initial biased values to zero along these heterochromatin ranges. Finally, all of the above BREC outputs are combined to generate one interactive plot to display for visualization and download (see details in "Easy, fast and accessible tool via an R-package and a Shiny app" section).

It is important to emphasize that throughout the whole main process module, only Step 1 " Estimating Marey map-based local recombination rates " comes from previous methods ([20, 21]). Otherwise, each of the steps 2-6 are fully developed (designed and implemented) within BREC and represent a new contribution, in addition to step zero " Data pre-processing ", as mentioned above.

Data and implementation

Validation data

The only input dataset to provide for BREC is genetic and physical maps for one or several chromosomes. A simple CSV file with at least two columns for both maps is valid. If the dataset is for more than one chromosome or the whole genome, a third column, with the chromosome identifier, is required.

Our results have been validated using Release 5 of the fruit fly *D. melanogaster* [36, 37] genome as well as the domesticated tomato *Solanum lycopersicum* genome (version SL3.0).

We also tested BREC using other datasets of different species: house mouse (*Mus musculus castaneus*, MGI) chromosome 4 [38], roundworm (*Caenorhabditis elegans*, ws170) chromosome 3 [39], zebrafish (*Danio rerio*, Zv6) chromosome 1 [40], respectively (see Additional file 13), as samples from the multi-genome dataset included within BREC (see further details on the full built-in dataset in "Description of main components of the Shiny app" section).

Fruit fly genome D.melanogaster Physical and genetic maps are available for download from the FlyBase website (http://flybase.org/; Release 5) [26]. This genome is represented here with five chromosomal arms: 2L, 2R, 3L, 3R, and X (see Additional file 2), for a total of 618 markers, 114.59Mb of physical map and 249.5cM of genetic map. This dataset is manually curated and is already clean from outliers. Therefore, the cleaning step offered within BREC was skipped.

Tomato genome S. lycopersicum Domesticated tomato with 12 chromosomes has a genome size of approximately 900Mb. Based on the latest physical and genetic maps reported by the Tomato Genome Consortium [28], we present both maps content (markers number, markers density, physical map length, and genetic map length) for each chromosome in Additional file 15. For a total of 1957 markers, 752.47Mb of physical map and 1434.49cM of genetic map along the whole genome.

Simulated data for quality control testing

We call *data scenarios*, the layout in which the data markers are arranged along the physical map. For experimentally testing the limits of BREC, various *data scenarios* have

been specifically designed based on *D. melanogaster* chromosomal arms (see Additional file 9).

In an attempt to investigate how the markers' density varies within and between the five chromosomal arms of *D. melanogaster* Release 5 genome, the density has been analyzed in two ways: locally (with 1Mb-bins) and globally (on the whole chromosome). Additional file 4 shows the results of this investigation, where each little box indicates how many markers are present within the corresponding region of size 1Mb on the physical map. The mean value represents the global density. It is also shown in Additional file 2 where the values are slightly different. This is due to computing the markers' density in two different ways with respect to the analysis. Additional file 2, presenting the genomic features of the validation dataset, shows markers density in Column 3, which is simply the result of the division of markers number (in column 2) by the physical map length (in Column 4). For example, in the case of chromosomal arm X, this gives $165/21.22 = 7.78 markers/Mb$. On the other hand, Additional file 4, aimed for analyzing the variation of local markers density, displays the mean of of all 1Mb-bins densities, which is calculated as the sum of local densities divided by the number of bins, and this gives $165/22 = 7.5 markers/Mb$.

The exact same analysis has been conducted on the tomato genome *S. lycopersicum* where the only difference lies in using 5-Mb instead of 1-Mb bins, due to the larger size of its chromosomes (see Additional file 5).

Validation metrics

The measure we used to evaluate the resolution of BREC's HCB is called *shift* hereafter. It is defined as the difference between the observed heterochromatin boundary (*observed_HCB*) and the expected one (*expected_HCB*) in terms of physical distance (in Mb)(see Equation 2).

$$shift = |observed\_HCB - expected\_HCB| \tag{2}$$

The *shift* value is computed for each heterochromatin boundary independently. Therefore, we observe only two boundaries on a telocentric chromosome (one centromeric and one telomeric). In comparison, we observe four boundaries in the case of an atelocentric chromosome (two centromeric giving the centromeric region and two telomeric giving each of the two telomeric regions).

The *shift* measure was introduced not only to validate BREC's results with the reference equivalents but also to empirically calibrate the DQC module, where we are mostly interested in the variation of its value as per variations of the quality of input data.

Implementation and Analysis

The entire BREC project was developed using the R programming language (version 3.6.3/2020-02-29) and the RStudio environment (version 1.2.5033).

The graphical user interface is build using the shiny and shinydashboard packages. The web-based interactive plots are generated by the plotly package. Data simulations, result analysis, reproducible reports, and data visualizations are implemented using a large set of packages such as tidyverse, dplyr, R markdown, Sweave and knitr among

others. The complete list of software resources used is available on the online version of the BREC package accessible at https://github.com/GenomeStructureOrganization.

From inside an R environment, the BREC package can be downloaded and installed using the command in the code chunk in Additional file 19. In case of installation issues, further documentation is available online on the ReadMe page of the GitHub repository. If all runs correctly, the BREC shiny application will be launched on your default internet browser (see Shiny interface screenshots in Additional file 14).

All BREC experiments have been carried out using a personal computer with the following specs:

- Processor: Intel® Core™ i7-7820HQ CPU @ 2.90GHz x 8
- Memory: 32Mo
- Hard disc: 512Go SSD
- Graphics: NV117 / Mesa Intel® HD Graphics 630 (KBL GT2)
- Operating system: 64-bit Ubuntu 20.04 LTS

Description of main components of the Shiny app

Build-in dataset

Users can either run BREC on a dataset of 44 genomes, mainly imported from [41], enriched with two mosquito genomes from [42] and updated with *D. melanogaster* Release 6 from FlyBase [26] (see Additional files 20 and 21), already available within the package, or, load new genomes data according to their own interest.

User-specific genomic data should be provided as inputs within at least a 3-column CSV file format, including for each marker: chromosome identifier, genetic distance, and physical distance, respectively. On the other hand, outputs from BREC running results are represented via interactive plots.

GUI input options

The BREC shiny interface provides the user with a set of options to select as parameters for a given dataset (see Fig. 3a). These options are mainly necessary in case the user works on his/her own dataset and this way the appropriate parameters would be available to choose from. First, a tab to specify the running mode (one chromosome). Then, a radio button group to choose the dataset source (existing within BREC or importing new dataset). For the existing datasets case, there is a drop-down scrolling list to select one of the available genomes (over 40 options), a second one for the corresponding physical map unit (Mb or pb) and a third one for the chromosome ID (available based on the dataset and not the genome biologically speaking). While for the import new dataset case, three more objects are added (see Fig. 3b); a fileInput to select csv data file, a textInput to enter the genome name (optional), and a drop-down scrolling list to select the data separator (comma , semicolon or tab character -set as the default-). As for the Loess regression model, the span parameter is required. It represents the percentage of how many markers to include in the local smoothing process. There is a numericInput object set by default at value 15% with an indication about the range of the span values allowed (min = 5%, max = 100%, step = 5%). The user should keep in mind that the span value actually goes from zero to one, yet, in a matter of simplification, BREC handles the conversion on its own. Thus, for example,

a value of zero basically means that no markers are used for the local smoothing process by Loess, and so, it will induce a running error. Lastly, there is a checkbox to apply data cleaning if checked. Otherwise, the cleaning step will be skipped. This options could save the user some running time if s/he already have a priori knowledge that a specific genome's dataset has already been manually curated). The user is then all set to hit the Run button. BREC will start processing the chromosome of interest by identifying its type (telocentric or atelocentric). Since this step is quite difficult to automatically get the correct result, the user might be invited to interfere via a popup alert asking for a chromosome type confirmation (see Fig. 3b). As shown in Additional file 14a, all available genomes could be accessed from the left-hand panel (in dark grey) and specifically on the tab " Genomic data " where two pages are available: " Download data files " which provides a data table corresponding to the selected genome on a scrolling list along with download buttons, and " Dataset details " displaying a more global overview of the whole build-in aata repository (see Additional file 14b). To give a glance at the GUI outputs, Fig. 3c shows BREC results displayed within an interactive plot where the user will have the an interesting experience by hovering over the different plot lines and points, visualising markers labels, zooming in and out, saving a snapshot as a PNG image file, and many more available options thanks to the plotly package.

Abbreviations

BREC: heterochromatin **B**oundaries and **REC**ombination rate estimates; HCB: **H**etero**C**hromatin **B**oundaries; *D. melanogaster*: D*rosophila melanogaster* (Fruitfly).

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-021-04233-1.

Additional file 1. BREC workflow steps applied on chromosomal arm 2L of *D. melanogaster* Release 5.

Additional file 2. Genomic features and BREC running time for the *D. melanogaster* Release 5 genome.

Additional file 3. Plots representing results of BREC and reference HCB on the *D. melanogaster* genome.

Additional file 4. Variations of markers local density per 1-Mb bins along *D. melanogaster* Release 5 chromosomal arms.

Additional file 5. Variations of markers local density per 5-Mb bins along the tomato genome *S. lycopersicum* 12 chromosomes.

Additional file 6. Results of BREC and reference HCB on the genome of *S. lycopersicum*.

Additional file 7. Plots representing results of BREC and reference HCB on the *S. lycopersicum* genome.

Additional file 8. The impact of decreasing markers density on the resolution of BREC's HCB expressed by the shift value..

Additional file 9. Distribution simulations.

Additional file 10. A schematic description of the chromosome type identification process implemented within BREC.

Additional file 11. Low density simulations.

Additional file 12. Comparison of regression models for recombination rate estimates along the five chromosomes (X, 2L, 2R, 3L, 3R) of *D. melanogaster* Release 5.

Additional file 13. BREC results on different species: from top to bottom are *M. musculus* (house mouse) chromosome 4, *C. elegans* (roundworm) chromosome 3, *D. rereo*(zebrafish) chromosome 1, respectively.

Additional file 14. Screenshots of BREC web application - Genomic data web pages.

Additional file 15. Genomic features and BREC running time for *S. lycopersicum*.

Additional file 16. BREC workflow.

Additional file 17. Comparing BREC with similar widely used tools.

Additional file 18. The data cleaning process implemented within BREC.

Additional file 19. Download, install and launch BREC.

Additional file 20. BREC's built-in dataset of genomic data.

Addtional file 21. This is a spreadsheet (.xlsx file). It includes accessible links to the genetic and physical maps for the 44 genomes mentioned in Additional file 20. Adapted from the Additional file named Table S1, published by [41].

Acknowledgements
Our thanks go to the members of the Institute of Sciences and Evolution of Montpellier (ISEM) as well as the Laboratory of Computer Science, Robotics and Microelectronics of Montpellier (LIRMM). In particular, team colleagues at the " Phylogeny and Molecular Evolution " and the " Methods and Algorithms in Bioinformatics ", respectively, for their insightful discussions.

About this supplement
This article has been published as part of BMC Bioinformatics Volume 22 Supplement 6, 2021: 19th International Conference on Bioinformatics 2020 (InCoB2020). The full contents of the supplement are available online at https://bmcbioinformatics.biomedcentral.com/articles/supplements/volume-22-supplement-6.

Authors' contributions
YM implemented and tested the methods, the R package, and the Shiny app. AC, ASFL and YM designed the methods, the experiments, and the analysis. AC, ASFL and YM wrote, edited, and revised the paper. All authors have read and approved the final manuscript.

Funding
This work has been supported by the Algerian Ministry of Higher Education and Scientific Research as part of the Algerian Excellence Scholarship Program AVERROES, with a full funding of YM's PhD project. As part of the Junior Research Group (ERJ 2018) grant, the hardware used to conduct this work was funded with the support of LabEx CeMEB, an ANR " Investissements d'avenir " program (ANR-10-LABX-04-01). The publication costs are funded by the CNRS as part of its name of the call for projects interdisciplinary program. Funding bodies had no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Availability of data and materials
The source code of the BREC R-package, including the Shiny app, is freely available at the GitHub repository https://github.com/GenomeStructureOrganization.

Declarations

Ethics approval and consent to participate
Not applicable

Consent to publish
Not applicable

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Genomics Department, Institute of Evolution Science of Montpellier (ISEM), Montpellier, France. [2]Informatics Department, Laboratory of Computer Science, Robotics and Microelectronics of Montpellier (LIRMM), Montpellier, France. [3]Institut Universitaire de France (IUF), Paris, France.

Received: 27 April 2021 Accepted: 4 June 2021
Published online: 06 August 2021

References

1. Coop G, Przeworski M. An evolutionary view of human recombination. Nat Rev Genet. 2007;8(1):23–34. https://doi.org/10.1038/nrg1947.
2. Duret L, Galtier N. Biased gene conversion and the evolution of mammalian genomic landscapes. Annu Rev Genom Hum Genet. 2009;10(1):285–311. https://doi.org/10.1146/annurev-genom-082908-150001.
3. Auton A, McVean G. Estimating recombination rates from genetic variation in humans. Methods Mol Biol. 2012;856:217–37. https://doi.org/10.1007/978-1-61779-585-5_9.
4. Chan AH, Jenkins PA, Song YS. Genome-wide fine-scale recombination rate variation in Drosophila melanogaster. PLoS Genet. 2012;8(12):1003090. https://doi.org/10.1371/journal.pgen.1003090.
5. Stapley J, Feulner PGD, Johnston SE, Santure AW, Smadja CM. Variation in recombination frequency and distribution across eukaryotes: patterns and processes. Philos Trans R Soc B Biol Sci. 2017;372(1736):20160455. https://doi.org/10.1098/rstb.2016.0455.
6. Morata J, Tormo M, Alexiou KG, Vives C, Ramos-Onsins SE, Garcia-Mas J, Casacuberta JM. The evolutionary consequences of transposon-related pericentromer expansion in melon. Genome Biol Evolut. 2018;10(6):1584–95. https://doi.org/10.1093/gbe/evy115.

7. Muller H, Gil J, Drinnenberg IA. The impact of centromeres on spatial genome architecture. Trends Genet. 2019;35(8):565–78. https://doi.org/10.1016/j.tig.2019.05.003.
8. Vanrobays E, Thomas M, Tatout C. Heterochromatin positioning and nuclear architecture. In: Annual plant reviews online, pp. 157–190. Wiley, Chichester (2017). https://doi.org/10.1002/9781119312994.apr0502.
9. Lu M, He X. Centromere repositioning causes inversion of meiosis and generates a reproductive barrier. Proc Natl Acad Sci. 2019;116(43):21580–91. https://doi.org/10.1073/pnas.1911745116.
10. Schueler MG, Higgins AW, Rudd MK, Gustashaw K, Willard HF. Genomic and genetic definition of a functional human centromere. Science. 2001;294(5540):109–15. https://doi.org/10.1126/science.1065042.
11. Weinstock GM, Robinson GE, Gibbs RA, Worley KC, Evans JD,Maleszka R, Robertson HM, Weaver DB, Beye M, Bork P, Elsik CG, Hartfelder K, Hunt GJ,Zdobnov EM, Amdam GV, Bitondi MMG, Collins AM, Cristino AS, Lattorff HMG, Lobo CH,Moritz RFA, Nunes FMF, Page RE, Simões ZLP, Wheeler D, Carninci P, Fukuda S,Hayashizaki Y, Kai C, Kawai J, Sakazume N, Sasaki D, Tagami M, Albert S, Baggerman G,Beggs KT, Bloch G, Cazzamali G, Cohen M, Drapeau MD, Eisenhardt D, Emore C, Ewing MA,Fahrbach SE, Forêt S, Grimmelikhuijzen CJP, Hauser F, Hummon, AB, Huybrechts J,Jones AK, Kadowaki T, Kaplan N, Kucharski R, Leboulle G, Linial M, Littleton JT,Mercer AR, Richmond TA, Rodriguez-Zas SL, Rubin EB, Sattelle DB, Schlipalius D,Schoofs L, Shemesh Y, Sweedler JV, Velarde R, Verleyen P, Vierstraete E, Williamson MR,Ament SA, Brown SJ, Corona M, Dearden PK, Dunn WA, Elekonich MM, Fujiyuki T, Gattermeier I,Gempe T, Hasselmann M, Kadowaki T, Kage E, Kamikouchi A, Kubo T, Kucharski R,Kunieda T, Lorenzen M, Milshina NV, Morioka M, Ohashi K, Overbeek R, Ross CA,Schioett M, Shippy T, Takeuchi H, Toth AL, Willis JH, Wilson MJ, Gordon KHJ,Letunic I, Hackett K, Peterson J, Felsenfeld A, Guyer M, Solignac M, Agarwala R,Cornuet JM, Monnerot M, Mougel F, Reese JT, Schlipalius D, Vautrin D, Gillespie JJ,Cannone JJ, Gutell RR, Johnston JS, Eisen MB, Iyer VN, Iyer V, Kosarev P, Mackey AJ,Solovyev V, Souvorov A, Aronstein KA, Bilikova K, Chen YP, Clark AG, Decanini LI,Gelbart WM, Hetru C, Hultmark D, Imler JL, Jiang H, Kanost M, Kimura K, Lazzaro BP,Lopez DL, Simuth J, Thompson GJ, Zou Z, De Jong P, Sodergren E, Csurös M,Milosavljevic A, Osoegawa K, Richards S, Shu CL, Duret L, Elhaik E, Graur D,Anzola JM, Campbell KS, Childs KL, Collinge D, Crosby MA, Dickens CM, Grametes LS,Grozinger CM, Jones PL, Jorda M, Ling X, Matthews BB, Miller J, Mizzen C,Peinado MA, Reid JG, Russo SM, Schroeder AJ, St Pierre SE, Wang Y, Zhou P, Jiang H,Kitts P, Ruef B, Venkatraman A, Zhang L, Aquino-Perez G, Whitfield CW,Behura SK, Berlocher SH, Sheppard WS, Smith DR, Suarez AV, Tsutsui ND, Wei X,Wheeler D, Havlak P, Li B, Liu Y, Jovilet A, Lee S, Nazareth LV, Pu LL, Thorn R,Stolc V, Newman T, Samanta M, Tongprasit WA, Claudianos C, Berenbaum MR, Biswas S,De Graaf DC, Feyereisen R, Johnson RM, Oakeshott JG, Ranson H, Schuler MA, Muzny D,Chacko J, Davis C, Dinh H, Gill R, Hernandez J, Hines S, Hume J, Jackson LR,Kovar C, Lewis L, Miner G, Morgan M, Nguyen N, Okwuonu G, Paul H, Santibanez J,Savery G, Svatek A, Villasana D, Wright R.Insights into social insects from the genome of the honeybee Apismellifera. Nature 2006;**443**(7114), 931–949.https://doi.org/10.1038/nature05260. arXiv:NIHMS150003.
12. Silva-Junior OB, Grattapaglia D. Genome-wide patterns of recombination, linkage disequilibrium and nucleotide diversity from pooled resequencing and single nucleotide polymorphism genotyping unlock the evolutionary history of Eucalyptus grandis. New Phytol. 2015;208(3):830–45. https://doi.org/10.1111/nph.13505.
13. Robert L, Nussbaum MDFF, McInnes RR, Willard HF. Thompson and Thompson genetics in medicine. Saunders W.B. Elsevier Health Sciences (2015).
14. Shen C, Li X, Zhang R, Lin Z. Genome-wide recombination rate variation in a recombination map of cotton. PLoS ONE. 2017;12(11):0188682. https://doi.org/10.1371/journal.pone.0188682.
15. Gui S, Peng J, Wang X, Wu Z, Cao R, Salse J, Zhang H, Zhu Z, Xia Q, Quan Z, Shu L, Ke W, Ding Y. Improving Nelumbo nucifera genome assemblies using high-resolution genetic maps and BioNano genome mapping reveals ancient chromosome rearrangements. Plant J. 2018;94(4):721–34. https://doi.org/10.1111/tpj.13894.
16. Rowan BA, Heavens D, Feuerborn TR, Tock AJ, Henderson IR, Weigel D. An ultra high-density arabidopsis thaliana crossover map that refines the influences of structural variation and epigenetic features. Genetics. 2019;213(3):302406–2019. https://doi.org/10.1534/genetics.119.302406.
17. Peñalba JV, Wolf JBW. From molecules to populations: appreciating and estimating recombination rate variation. Nat Res. 2020. https://doi.org/10.1038/s41576-020-0240-1.
18. Stumpf MPH, McVean GAT. Estimating recombination rates from population-genetic data. Nat Rev Genet. 2003;4(12):959–68. https://doi.org/10.1038/nrg1227.
19. Jeffreys AJ. High resolution analysis of haplotype diversity and meiotic crossover in the human TAP2 recombination hotspot. Hum Mol Genet. 2000;9(5):725–33. https://doi.org/10.1093/hmg/9.5.725.
20. Chakravarti A. A graphical representation of genetic and physical maps: the Marey map. Genomics. 1991;11(1):219–22. https://doi.org/10.1016/0888-7543(91)90123-V.
21. Rezvoy C, Charif D, Guéguen L, Marais GAB. MareyMap: an R-based tool with graphical interface for estimating recombination rates. Bioinformatics. 2007;23(16):2188–9. https://doi.org/10.1093/bioinformatics/btm315.
22. Siberchicot A, Bessy A, Guéguen L, Marais GAB. MareyMap online: a user-friendly web application and database service for estimating recombination rates using physical and genetic maps. Genome Biol Evol. 2017;9(10):2506–9.
23. Fiston-Lavier AS, Singh ND, Lipatov M, Petrov DA. Drosophila melanogaster recombination rate calculator. Gene. 2010;463(1–2):18–20. https://doi.org/10.1016/j.gene.2010.04.015.
24. Termolino P, Cremona G, Consiglio MF, Conicella C. Insights into epigenetic landscape of recombination-free regions. Chromosoma. 2016;125(2):301–8. https://doi.org/10.1007/s00412-016-0574-9.
25. Langley CH, Stevens K, Cardeno C, Lee YCG, Schrider DR, Pool JE, Langley SA, Suarez C, Corbett-Detig RB, Kolaczkowski B, Fang S, Nista PM, Holloway AK, Kern AD, Dewey CN, Song YS, Hahn MW, Begun DJ. Genomic variation in natural populations of Drosophila melanogaster. Genetics. 2012;192(2):533–98. https://doi.org/10.1534/genetics.112.142018.
26. ...Thurmond J, Goodman JL, Strelets VB, Attrill H, Gramates LS, Marygold SJ, Matthews BB, Millburn G, Antonazzo G, Trovisco V, Kaufman TC, Calvi BR, Perrimon N, Gelbart SR, Agapite J, Broll K, Crosby L, Dos Santos G, Emmert D, Falls K, Jenkins V, Sutherland C, Tabone C, Zhou P, Zytkovicz M, Brown N, Garapati P, Holmes A, Larkin A, Pilgrim C, Urbano P, Czoch B, Cripps R, Baker P. FlyBase 2.0: the next generation. Nucl Acids Res. 2019;47(D1):759–65. https://doi.org/10.1093/nar/gky1003.

27. Comeron JM, Ratnappan R, Bailin S. The many landscapes of recombination in *Drosophila melanogaster*. PLoS Genet. 2012;8(10):1002905. https://doi.org/10.1371/journal.pgen.1002905.
28. Sato S, et al. The tomato genome sequence provides insights into fleshy fruit evolution. Nature. 2012;485(7400):635–41. https://doi.org/10.1038/nature11119.
29. Riddle NC, Minoda A, Kharchenko PV, Alekseyenko AA, Schwartz YB, Tolstorukov MY, Gorchakov AA, Jaffe JD, Kennedy C, Linder-Basso D, Peach SE, Shanower G, Zheng H, Kuroda MI, Pirrotta V, Park PJ, Elgin SCR, Karpen GH. Plasticity in patterns of histone modifications and chromosomal proteins in *Drosophila heterochromatin*. Genome Res. 2011;21(2):147–63. https://doi.org/10.1101/gr.110098.110.
30. Lenormand T, Engelstädter J, Johnston SE, Wijnker E, Haag CR. Evolutionary mysteries in meiosis. R Soc Lond. 2016. https://doi.org/10.1098/rstb.2016.0001.
31. Agresti A. An introduction to categorical data analysis, pp. 1–356. Wiley, Hoboken (2007). https://doi.org/10.1002/0470114754.
32. Levan A, Fredga K, Sandberg AA. Nomenclature for centromeric position on chromosomes. Hereditas. 1964;52(2):201–20. https://doi.org/10.1111/j.1601-5223.1964.tb01953.x.
33. Cleveland WS, Devlin SJ. Locally weighted regression: an approach to regression analysis by local fitting. J Am Stat Assoc. 1988;83(403):596–610. https://doi.org/10.1080/01621459.1988.10478639.
34. Cleveland WS, Loader C. Smoothing by local regression: principles and methods. In: Härdle, W., Schimek, M.G. (eds.) Statistical theory and computational aspects of smoothing, pp. 10–49. Physica-Verlag HD, Heidelberg (1996). https://doi.org/10.1007/978-3-642-48425-4_2.
35. Zhang D. A coefficient of determination for generalized linear models. Am Stat. 2017;71(4):310–6. https://doi.org/10.1080/00031305.2016.1256839.
36. Hoskins RA, Carlson JW, Kennedy C, Acevedo D, Evans-Holm M, Frise E, Wan KH, Park S, Mendez-Lago M, Rossi F, Villasante A, Dimitri P, Karpen GH, Celniker SE. Sequence finishing and mapping of Drosophila melanogaster heterochromatin. Science. 2007;316(5831):1625–8. https://doi.org/10.1126/science.1139816.
37. Hoskins RA, Carlson JW, Wan KH, Park S, Mendez I, Galle SE, Booth BW, Pfeiffer BD, George RA, Svirskas R, Krzywinski M, Schein J, Accardo MC, Damia E, Messina G, Méndez-Lago M, De Pablos B, Demakova OV, Andreyeva EN, Boldyreva LV, Marra M, Carvalho AB, Dimitri P, Villasante A, Zhimulev IF, Rubin GM, Karpen GH, Celniker SE. The release 6 reference sequence of the *Drosophila melanogaster* genome. Genome Res. 2015;25(3):445–58. https://doi.org/10.1101/gr.185579.114.
38. Cox A, Ackert-Bicknell CL, Dumont BL, Ding Y, Bell JT, Brockmann GA, Wergedal JE, Bult C, Paigen B, Flint J, Tsaih S-W, Churchill GA, Broman KW. A new standard genetic map for the laboratory mouse. Genetics. 2009;182(4):1335–44. https://doi.org/10.1534/genetics.109.105486.
39. Hillier LDW, Marth GT, Quinlan AR, Dooling D, Fewell G, Barnett D, Fox P, Glasscock JI, Hickenbotham M, Huang W, Magrini VJ, Richt RJ, Sander SN, Stewart DA, Stromberg M, Tsung EF, Wylie T, Schedl T, Wilson RK, Mardis ER. Whole-genome sequencing and variant discovery in *C. elegans*. Nat Methods. **5**(2), 183–188 (2008). https://doi.org/10.1038/nmeth.1179.
40. Freeman JL, Adeniyi A, Banerjee R, Dallaire S, Maguire SF, Chi J, Ng B, Zepeda C, Scott CE, Humphray S, Rogers J, Zhou Y, Zon LI, Carter NP, Yang F, Lee C. Definition of the zebrafish genome using flow cytometry and cytogenetic mapping. BMC Genom. 2007;8(1):195. https://doi.org/10.1186/1471-2164-8-195.
41. Corbett-Detig RB, Hartl DL, Sackton TB. Natural selection constrains neutral diversity across a wide range of species. PLoS Biol. 2015;13(4):1002112. https://doi.org/10.1371/journal.pbio.1002112.
42. Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, Shamim MS, Machol I, Lander ES, Aiden AP, Aiden EL. De novo assembly of the aedes aegypti genome using hi-c yields chromosome-length scaffolds. Science. 2017;356(6333):92–5. https://doi.org/10.1126/science.aal3327.

Publisher's Note

Journal of Bioinformatics and Computational Biology

Involving Repetitive Regions in Scaffolding Improvement

Quentin Delorme*
LIRMM, Univ Montpellier, CNRS, Montpellier, France

Rémy Costa†
IGH-UMR9002, Univ Montpellier, CNRS, Montpellier, France

Yasmine Mansour‡
LIRMM-ISEM, Univ Montpellier, Montpellier, France

Anna-Sophie Fiston-Lavier§
ISEM, Univ Montpellier, CNRS, Montpellier, France

Annie Chateau¶
LIRMM, Univ Montpellier, CNRS, Montpellier, France

Received (Day Month Year)
Revised (Day Month Year)
Accepted (Day Month Year)

In this paper we interrogate the influence of Repeated Elements during the assembly process. We analyse the link between presence and nature of TE and misassembly events in genome assemblies. We propose to improve assemblies by taking into account the presence of repeated elements, including TEs, during the scaffolding step. We analyse the results and relate the misassemblies to TEs before and after correction.

Keywords: Assembly; Scaffolding; Repeated elements; Transposable elements;

1. Introduction

Motivation. Repeated genomic regions are usually defined as parts of the genomes which are enriched in repeated elements. Repeated elements are sequences appearing in several copies in genomes [1]. We use to classify them in three main categories: *Segmental duplications*, which are low-copy number elements encompassing several

*quentin.delorme@lirmm.fr
†remy.costa@live.fr
‡yasmine.mansour@umontpellier.fr
§anna-sophie.fiston-lavier@umontpellier.fr
¶chateau@lirmm.fr

genomics elements such as genes or other repeats; *Tandem repeats*, which are high-copy number elements present consecutively; *Transposable Elements* (TEs), which are high-copy number elements dispersed along the genomes.

Repeated Regions (RRs) are detected using specific tools, and there is a large variety of them [2], concerning particular taxonomic groups (Insects, Bacteria, etc.). A generic tool broadly used to find them is called RepBase [3]. Through its tool Censor [3] [4], it is possible to detect RRs coming from a variety of organisms and their localisation in a set of given sequences. Matching sequences are listed, together with the identifier of the RRs. Several families of repeats, and especially TEs, are mentioned in the following. These families are those usually classified in RepBase (LINE, SINE, LTR, etc.).

Repetitive regions in DNA sequences are present in almost all organisms and may represent over 80% of the genome size [5]. Fundamental source of genetic plasticity and diversity, yet, they are a source of complication when it comes to assembling genomes. Amongst repeated elements, we pay particular attention to TEs, which are variously present in the genomes we considered. They may be particularly able to bring exploitable information in assembly, due to their diversity and evolutionary pace.

Genomes are usually obtained by sequencing, which produces a set of *reads* whose length and quality depend on the sequencing technology [6]. Those reads are then assembled using dozens of possible tools, the most recent proposing hybrid strategies using both short and long reads [7]. Genome assembly has become crucial for conducting genomic studies in various fields as environment, health, genetics, evolution and many more. Recent studies has highlighted the impact of assembly quality on result interpretation [8]. Even with advanced high-throughput sequencing technologies, genome assembly is facing a big challenge towards achieving it's optimum quality. Indeed, most of the genomes in databases are fragmented in huge sets of *contigs*, short for contiguous DNA sequences. Such fragmentation is observed even for well-studied genomes, unless they have been sequenced again, with long read technologies for instance. To reduce this fragmentation and improve existing available genomes, the *scaffolding* step exploits additional information on original data (e.g. pairing information), to infer the order and the orientation of the contigs along the target genome, using a set of possibly inconsistent pairing information. Formally, it is possible to extract from these information a set of relationships between the contigs, that may be inconsistent. The *scaffold graph* is defined as follows: vertices represent contig extremities, while edges are of two kinds: (1) contig edges, linking both extremities of a contig, and (2) inter-contig edges relating the pairing-information. A weight function on the inter-contig edges indicates how many pairs are supporting this edge (see Figure 1). Due to repeats, some of the inter-contigs edges are erroneous and have to be removed from the graph. In other cases, they are supported by RRs. Interesting surveys on recent scaffolding methods are available in [9] and [10].

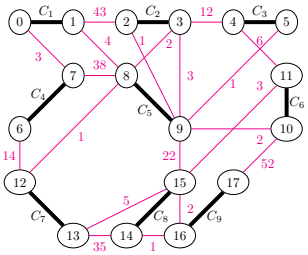


Fig. 1: A scaffold graph with nine contigs (bold edges) and twenty inter-contig edges. Vertices are contig extremities. For instance, contig C_1 is figured by vertices labelled by 0 and 1, the (0,1) direction corresponds to the forward reading of the contig in the assembly file, and the (1,0) direction corresponds to the reverse direction. Inter-contig edges are labelled by the number of pairs of reads connecting one contig extremity to another.

How can RRs induce assembly errors. RRs are disturbing *de novo* assembly from short reads sequencing data, during both contig production and scaffolding. Contig production methods in a short reads context are in majority based on De Bruijn graphs (DBG), and exploit the overlaps between reads, corresponding to paths in those graphs. During the traversal of the DBG, k-mers from repetitive regions are collapsed, yielding ambiguously branching paths in the DBG. Facing this problem, the methods decide either to cut potentially ambiguous paths, or to propose longer paths which may be erroneous ([11]). The presence of RRs disrupts assembly process by inducing : (1) chimeric contigs due to collapsed RRs or (2) assembly breaks. Figure 2 illustrates both cases.

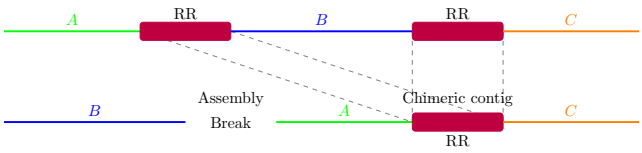


Fig. 2: **Misassembly scenario due to repeated regions.** Above, the original genome with three unique regions A, B, C and two copies of the same RR. The assembly process produces a set of contigs (below) with collapsed RR copies, yielding to an assembly break (isolated region B), and a chimeric contig (A-RR-C).

While the efficiency of bioinformatic tools used for assembly is increasing, errors of sequence construction from contiguous short reads persist. One way to untangle ambiguous parts of these graphs is to use long reads, produced by third-generation sequencing technologies, for instance like in [12]. However, this is not always possible due to high cost and lower quality. Recent state-of-the-art on the error correction tools targeting Illumina short reads shows that it is possible to enhance De Bruijn Graph [13]. especially when the correction targets reads near highly repetitive DNA regions [14].

The scaffolding step is also touched by the repetitive region issue. RRs locations

on contigs, especially when they are near the extremities, can lead to ambiguities at the scaffolding step. Indeed, most scaffolders use a graph structure establishing relationships between contigs sharing a piece of information. This information may come from a set of long reads (if available), or pairs of short reads, one read mapping on the first contig, and the mate mapping on the other contig. Typically, in this latter case, when the reads come from a RR, they may map ambiguously, and a choice has to be made during the processing of the graph. Here we propose, instead of just suffer from their presence, to use RR sequences to enhance scaffolding.

Contribution. The main question we address here is: How to improve the quality of genome assembly using RRs ? A secondary question is raised about the *type* of repeats which are the most involved in misassemblies. We focus here on the improvement of genomes produced using a *de novo* approach using short reads (improvement of existing assemblies in databases), with a relatively well-defined repeat landscape (repeats documented in the Repbase database). We propose a method based on a pipeline progressively refining inter-contig edges through RR analysis. The paper is organised as follows: in Section 2 we describe the method and the data used for validation, whereas results are presented in Section 3, and discussed in Section 4.

2. Materials and methods

2.1. *Method description*

We implemented a snakemake[15] pipeline summarized on Figure 3. The first four steps aim to produce datasets composed of both a reference genome and a contig set which can be compared to the reference. Further steps are separated in two paths: first path correspond to a classical scaffolding with paired-end reads information leading to generation of paired-ends scaffolding graph (PE graph), whereas the second path includes repeated regions analysis. The original part of our work lies in this second path, which we describe in details in Paragraph 2.1.2.

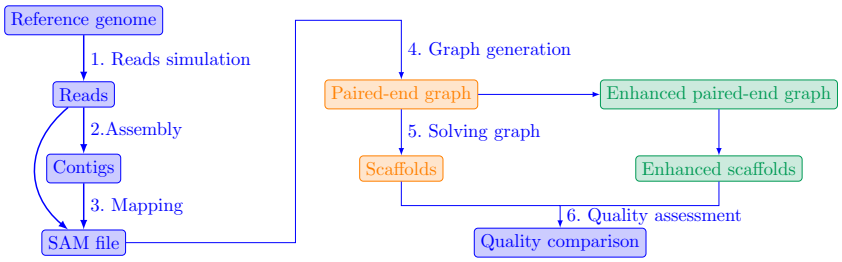


Fig. 3: Overview of the pipeline.

2.1.1. *Data production*

Simulation. We validated our approach on simulated data. The first step was to generate *paired-end reads* as basic data for the assembly and then the scaffolding. To simulate short reads, we chose the ART[16] software (version 2.5.8;), which produces *reads* close to the technologies commonly used, and because of its simplicity of use, while allowing a large choice of options.

Assembly. We chose to build the contigs with Spades (version 3.13.0 ; `http://cab.spbu.ru/software/spades/` ; [17]), which is one the mostly used assembly tools and proposed an iterative DBG approach, and Minia (version 3.2.1 ; `https://github.com/GATB/minia` ; [18]), which is very light in terms of memory consumption, thanks to its use of Bloom filters. We therefore obtain two separate contig files from different assembly programs which will each be used in all the following steps of the *pipeline* so that we can compare their qualities at the end.

Mapping. The next step is to map the *paired-end reads* to the contigs obtained in the previous step. The contigs were mapped on the reference sequences using Minimap2 (version 2.17 ; `https://github.com/lh3/minimap2` ; [19]) and BWA MEM (version 0.7.17-r1188 ; `https://github.com/lh3/bwa` ; [20]). Both mapping tools are also famous for their interesting performances and reliability. The initial protocol used BWA [20], an alignment tool using "reverse search" (*backward search*) with the Burrows-Wheeler transform. We chose to use BWA MEM, improvement of BWA, because the latter did not take into account the information in *paired-end reads*. We decided to compare it with Minimap2 for the speed of execution of the latter.

Graph generation. Generating *paired-end* scaffold graphs is done with the Scaftools tool [21], from the mapping of paired-end reads to the contigs. The graphs generated in each of the four cases (both assembly tools and both mapping tools) will then be passed into our graph improvement tool.

2.1.2. *Repeated Regions analysis*

Repeated Region detection. The consensus sequences of the repeated regions were obtained from the Repbase Update (RU) database. RU contains more than 38,000 sequences of different families or subfamilies. The RRs present within the contigs were then detected by aligning the RU consensus sequences using BLAST (megablast default parameters). We therefore obtain an alignment file used to label the contigs.

Clustering contigs according to repetitions family. Two contigs carrying repetitions of different families can be linked within the PE graph. This link is due to the similarities between these RRs but is not coherent with the biological reality. It is therefore necessary to separate the contigs according to the repetitions they carry in order to limit such incoherent links and instead, favor them in case of contigs carrying the same RR. The classification and clustering of repetitions can be done at different levels/scales: clusters that are too small would be less informative,

while clusters that are too large would make the further processing heavier/more complicated. We performed the clustering at the subfamily level.

Building the RR graph. At this stage, each contig is defined by the following values: its name, its length (ℓ), the name of the repetition family carried, the identifier of the original repetition (*repid*), the start bound (*start*) and the end bound (*end*) of the RR on the contig. If one of the bounds is equal to 1 or ℓ, the RR is considered *external*, otherwise it is qualified as *internal*. Within each cluster, the position of the RRs on each contig is evaluated and then exploited in order to join the contigs carrying the same RR. The purpose of these junctions is to orient the contigs according to the RR information they carry. These information allow, for each cluster, to generate a graph in Graphviz format[22]. The set of all these graphs is called the *RRs graph*. The processus leading to the RR graph is described on Figure 4.

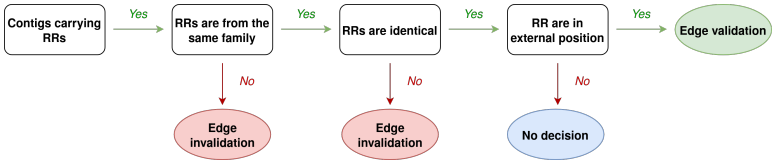


Fig. 4: RRs detection and characterization.

Using RR graphs to correct a PE graph. We use the edges from the RR graph to apply corrections to the PE graph. We recall that in both graphs, vertices are contig extremities and edges are links between these extremities. It is obvious that the corrections applied concern only the edges implied as for repetitions. However, we can assume that the edges not affected by RR are less likely to cause problems because they are not impacted by them. These corrections can be of several types:

- **Edges in common in the PE graph and the RR graph**. We are a priori assured of the validity of an edge if it is present within both graphs. In this case, we add an additional weight to the weight of the PE edge, to strengthen this edge in the final scaffolding. This weight is relative to the size of the cluster from which the RR comes from, with an additional weight of one per hundred elements in the cluster.
- **PE edges between contigs carrying RRs from different families**. In this case, the PE edges are removed from the PE graph, since the similarity yielding this edge has been invalidated by the RR sequences.
- **PE edge with only one contig carrying RRs**. In this case, the validation process depends on the way the RR is mapped on the contig. The invalidation is performed only when the RR should be present on both contigs (see Figure 5).

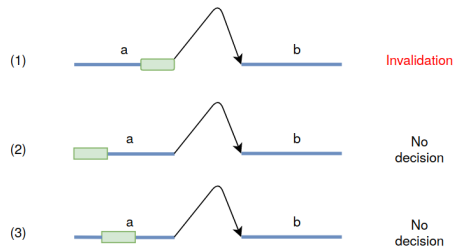


Fig. 5: PE Edge validation for case with only one contig carrying RR. Validation depends on position of RR within the contig.

2.1.3. *After the RR analysis: solving and quality analysis*

Solving the graphs. The resolution of the graphs obtained is also carried out with Scaftools, for the graphs of *paired-end* as well as for the improved graphs. By solving the graph, we mean extracting from the scaffold graph a set of paths of maximum total weight, corresponding to the scaffolds. Knowing that they cause incoherent alignments, the repeated regions will induce a bias in the scores of intercontigs edges, which will result in poor resolution of the graph. From each original reference genome, we obtain at the end of the *pipeline*, 8 different genomes.

Quality assessment. Each assembly was validated with QUAST-LG (version 5.0.2 ; `http://cab.spbu.ru/software/quast-lg./` ; [23],[24]). We expected to get a reduction of *misassemblies* in the tests performed with RRs-corrected PE graphs (PE+RR graph).

2.2. *Data*

We decided to take as reference genomes *Drosophila melanogaster* for the very high quality of its sequenced genome as well as the knowledge of its repeated regions[25], and *Caenorhabditis elegans* for its small genome, containing little repetitions, and also for its sequencing quality. We simulated sequencing data using *D. melanogaster* and *C. elegans* with the following common specifications:

- Simulated technology: Illumina HiSeq 2000
- Coverage: 20X
- Reads size: 100bp
- Insert size 300bp
- Standard deviation: 10%

D.melanogaster	SPAdes				minia			
	minimap2		BWA-MEM		minimap2		BWA-MEM	
	PE only	PE + RR	PE only	PE + RR	PE only	PE + RR	PE only	PE + RR
Scaffolds	1894	2019	1861	2032	2212	2158	2307	2249
Unaligned scaffolds	8	8	8	**6**	103	**66**	140	**109**
Coverage (%)	83.586	83.147	83.564	83.163	82.691	82.357	82.749	82.42
NG50	138 662	129 502	141 803	133 722	120 493	115 298	115 249	114 878
Nb of misassemblies	708	**552**	770	**567**	159	164	252	261
Improvement rate %	**22.03**		**26.36**		-3.14		-3.57	

Table 1: Result on the *D. melanogaster* dataset. Bold figures shows the improvement achieved by the method. The improvement rate on last row is calculated using the number of misassemblies (100×(PE only - (PE+RR))/PE only).

3. Results

3.1. *Effect on the assembly quality*

For each dataset, the eight scaffold sets produced by the pipeline are compared to the reference using QUAST. We selected the following criteria to analyse the efficiency of the approach: number of contigs (in this case, number of final scaffolds), number of unaligned contigs (scaffolds), percentage of the genome covered by the scaffolds, NG50 (corresponding to the scaffold size such that 50% of the known or estimated genome size are supposed to be of the NG50 length or longer), and the number of misassemblies.

D. melanogaster.

Table 1 shows the results for the eight genomes produced on the *Drosophila Melanogaster* dataset.

For *D. melanogaster*, the results show a slight reduction in genome and NG50 coverage (length for which the collection of contigs of this length cover at least half of the reference genome) but an improvement in the number of *misassemblies* up to 26% for SPAdes (but no improvement with minia). SPAdes provides fewer contigs than minia, and produces far fewer unaligned contigs. It also provides greater genome coverage. Our hypothesis to explain this difference between both assembly tools is that minia, due to its decision process to cut nodes with a large in or out-degree in the DBG, may isolate more drastically repeated region as contigs, thus RRs could not help connecting them to other contigs. Difference between the use of minimap2 vs. the use of BWA-MEM in the mapping does not appear to be significant.

C. elegans Table 2 shows the results for the eight genomes produced on the *Caenorhabditis elegans* dataset.

Results are not so positive for the *C. elegans* genome, when applied on the whole genome. Misassemblies are more numerous with the application of the method, contrary to the expectations. Improvement rate are negative, but small. Again, results are better for SPAdes than for minia.

On the contrary, when the method is applied separately on each chromosome,

C. elegans	SPAdes				minia			
	minimap2		BWA-MEM		minimap2		BWA-MEM	
	PE only	PE + RR	PE only	PE + RR	PE only	PE + RR	PE only	PE + RR
Scaffolds	4266	4597	4244	4541	5236	5286	5230	5301
Unaligned scaffolds	0	0	0	0	1	0	3	4
Coverage (%)	89.686	89.26	89.673	89.325	87.804	87.501	87.825	87.549
NG50	33 576	29 337	33 157	29 884	24 884	24 437	24 864	24 275
Nb of misassemblies	1770	1783	1893	1921	981	1009	1258	1305
Improvement rate %	-0.73		-1.48		-2.85		-3.89	

Table 2: Result on the *C. elegans* dataset (whole genome). The improvement rate on last row is calculated using the number of misassemblies (100×(PE only - (PE+RR))/PE only).

C. elegans	SPAdes				minia			
	minimap2		BWA-MEM		minimap2		BWA-MEM	
	PE only	PE + RR	PE only	PE + RR	PE only	PE + RR	PE only	PE + RR
Chr. 1	292	**276**	307	**298**	163	173	225	230
Chr. 2	282	**230**	283	**251**	124	**119**	151	151
Chr. 3	225	**213**	230	**220**	119	120	161	170
Chr. 4	378	**345**	389	**367**	210	**194**	272	**253**
Chr. 5	358	**344**	388	**370**	186	192	243	**238**
Chr. X	233	**193**	248	**212**	125	**119**	153	**148**
Improvement rate (min/mean/max)%	3.91 / **9.84** / 18.44 p-value=0.88%		2.93 / **7.25** / 14.52 p-value=0.56%		-6.13 / **1.05** / 7.62 p-value=68.07%		-5.60 / **0.75** / 6.99 p-value=55.92%	

Table 3: Number of misassemblies on the *C. elegans* dataset, chromosome per chromosome. Results are significantly better with SPAdes, and equivalent with minia (see p-values).

results are far better, as shown on Table 3 (only the number of misassemblies are reported here, for a better readibility), for SPAdes.

3.2. *RR within the misassemblies*

To analyze the *misassemblies* detected by QUAST, we mapped them on the reference genome. We crossed this mapping with a GFF file of *D. melanogaster* genome, with RRs (tandem repetitions, pseudo-genes and transposable elements), and detected the RRs present at the ends of the *missasemblies*. We have observed that RRs are involved in 60% to 70% of the remaining *misassemblies*. Even if we detect some tandem repetitions and pseudo-genes, the vast majority is composed of transposable elements. We can therefore deduce that transposable elements are the most disturbing for the reconstruction of genomes, because of their numerous specificities (size, activity, age). We performed this analysis on the genome of *D. melanogaster* using the latest available version of its sequenced genome (*release* 6.26), which lists all of the annotated regions known to date. Results of this analysis on the eight scaffoldings, for the 2R chromosomal arm, are presented on Figure 6. Results are very similar on other chromosomal arms and chromosomes.

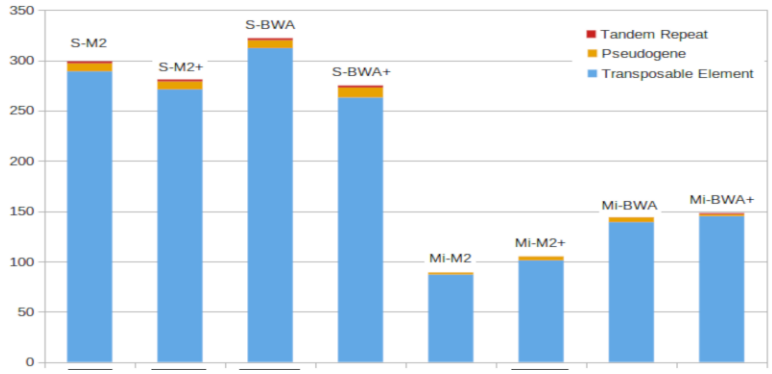


Fig. 6: Analysis of the number of repeats on the extremities of misassemblies along the 2R chromosomal arm of *D. melanogaster*. For the assembly: "S" stands for *Spades* and "Mi" for *Minia*. For the mapping: "M2" stands for *Minimap2*. For scaffolding graphs, the "+" sign indicates an enhanced graph (PE+RR).

To complete this analysis and find out if one type of TE is particularly involved in the assembly disturbance, we also considered an "historical" approach, and had a look at the the first release of the *Drosophila melanogaster* genome. This previous release is more fragmented, and the gaps are essentially due to repeat-rich regions [25] which necessitate non-trivial techniques to be partially desintricated. We mapped the drosophila known TEs on the gaps constated when aligning release 1 against release 6 and examined each categories. Result is shown on Figure 7, revealing that essentially LTR are responsible for these misassemblies.

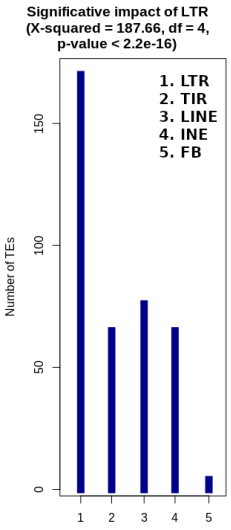


Fig. 7: Number of TEs related to gaps, classified by type for *D. melanogaster R1 vs. R6*

4. Discussion and conclusion

Improving the quality of sequence reconstructions is necessary for a better understanding of the evolution of genomes and their dynamics. Repeated regions present challenges for genome assembly and scaffolding. We presented a pipeline based on scaffold graph enhancement when combining classical paired-end reads information with repeated regions information.

This pipeline shows promising results when used with the SPAdes assembly tool. Probably due to the fact that we based our analysis on reference genomes, which are well-assembled but escape repeat-rich regions, the result may not appear spectacular, however it opens a win-

dow on assembly improvement. We also showed that repeated regions are involved on the misassemblies, and that they are essentially transposable elements, which is not surprising but allows us to concentrate on these particular repeats. Amongst those transposable elements, LTR were responsible for the vast majority of gaps constated on the *Drosophila melanogaster* previous releases.

A lot of pending questions remain however. First, it would be interesting to exploit other options when using the pipeline. For the moment, the re-weighting of the consistent edges is quite arbitrary, and depends on the size of the clusters. It would be interesting to study the robustness of this criteria, with respect to the clustering scale for instance, as well as it possible improvement using distance information. Indeed, distance between contigs may be estimated using the pairing information together with the insert size between mate fragments in the short reads sequencing. This estimation is not really precise, but may help refining the consistency in ambiguous cases, when compared to the length of the detected RRs. In the presented version, the removal of intercontig edges is a binary decision process: we decide to keep or to remove edges. This process could be done with more subtility by introducing a continuous measure on the edges reliability, which would influence the weight of the edge positively ("keep the edge" case) or negatively ("remove the edge" case). For instance we could try to quantify how we can come across these RRs randomly, and consequently to establish probability of decision. Of course, another natural perspective of our work is to extend it to a larger variety of genomes and assembly tools.

References

1. E. Lerat. Identifying repeats and transposable elements in sequenced genomes: How to find your way through the dense forest of programs, jun 2010.
2. P. Goerner-Potvin and G. Bourque. Computational tools to unmask transposable elements. *Nature Reviews Genetics*, page 1, sep 2018.
3. W. Bao, K. K. Kojima, and O. Kohany. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA*, 6(1):11, jun 2015.
4. J. Jurka, P. Klonowski, V. Dagman, and P. Pelton. Censor - A program for identification and elimination of repetitive elements from DNA sequences. *Computers and Chemistry*, 20(1):119–121, 1996.
5. T. Wicker, H. Gundlach, M. Spannagl, C. Uauy, P. Borrill, R. H. Ramírez-González, R. De Oliveira, K. F. X. Mayer, E. Paux, and F. Choulet. Impact of transposable elements on genome structure and evolution in bread wheat. *Genome Biol*, 19(1):103, 08 2018.
6. E. R. Mardis. DNA sequencing technologies: 2006-2016, feb 2017.
7. J. R. Miller, P. Zhou, J. Mudge, J. Gurtowski, H. Lee, T. Ramaraj, B. P. Walenz, J. Liu, R. M. Stupar, R. Denny, L. Song, N. Singh, L. G. Maron, S. R. McCouch, W. R. McCombie, M. C. Schatz, P. Tiffin, N. D. Young, and K. A.T. Silverstein. Hybrid assembly with long and short reads improves discovery of gene family expansions. *BMC Genomics*, 18(1):541, jul 2017.
8. M. Chakraborty, N. W. Vankuren, R. Zhao, X. Zhang, S. Kalsow, and J. J. Emerson. Hidden genetic variation shapes the structure of functional elements in Drosophila. *Nature Genetics*, 50(1):20–25, jan 2018.

9. I. Mandric, J. Lindsay, I. I. Măndoiu, and A. Zelikovsky. Scaffolding algorithms. In I. Măndoiu and A. Zelikovsky, editors, *Computational Methods for Next Generation Sequencing Data Analysis*, chapter 5, pages 107–132. Wiley, 2016.
10. E. S. Rice and R. E. Green. New Approaches for Genome Assembly and Scaffolding. *Annu Rev Anim Biosci*, 7:17–40, 02 2019.
11. T. Treangen and S. Salzberg. Repetitive dna and next-generation sequencing: Computational challenges and solutions. *Nature reviews. Genetics*, 13:36–46, 11 2011.
12. M. Qin, S. Wu, A. Li, F. Zhao, H. Feng, L. Ding, and J. Ruan. LRScaf: improving draft genomes using long noisy reads. *BMC Genomics*, 20(1):955, Dec 2019.
13. M. Heydari, G. Miclotte, Y. Van De Peer, and J. Fostier. Illumina error correction near highly repetitive DNA regions improves de novo genome assembly. *BMC Bioinformatics*, 20(1):298, jun 2019.
14. M. Heydari, G. Miclotte, P. Demeester, Y. Van de Peer, and J. Fostier. Evaluation of the impact of Illumina error correction tools on de novo genome assembly. *BMC Bioinformatics*, 18(1):374, aug 2017.
15. F. Mölder, K. P. Jablonski, B. Letcher, M. B. Hall, C. H. Tomkins-Tinch, V. Sochat, J. Forster, S. Lee, S. O. Twardziok, A. Kanitz, A. Wilm, M. Holtgrewe, S. Rahmann, S. Nahnsen, and J. Köster. Sustainable data analysis with Snakemake. *F1000Res*, 10:33, 2021.
16. W. Huang, L. Li, J. R. Myers, and G. T. Marth. ART: a next-generation sequencing read simulator. *Bioinformatics*, 28(4):593–594, Feb 2012.
17. A. Bankevich, S. Nurk, D. Antipov, A. A. Gurevich, M. Dvorkin, A. S. Kulikov, V. M. Lesin, S. I. Nikolenko, S. Pham, A. D. Prjibelski, A. V. Pyshkin, A. V. Sirotkin, N. Vyahhi, G. Tesler, M. A. Alekseyev, and P. A. Pevzner. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol*, 19(5):455–477, May 2012.
18. R. Chikhi and G. Rizk. Space-efficient and exact de Bruijn graph representation based on a Bloom filter. *Algorithms Mol Biol*, 8(1):22, Sep 2013.
19. H. Li. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18):3094–3100, 09 2018.
20. H. Li and R. Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760, Jul 2009.
21. M. Weller, A. Chateau, and R. Giroudeau. Exact approaches for scaffolding. *BMC Bioinformatics*, 16 Suppl 14:S2, 2015.
22. E. R. Gansner and S. C. North. An open graph visualization system and its applications to software engineering. *SOFTWARE - PRACTICE AND EXPERIENCE*, 30(11):1203–1233, 2000.
23. A. Gurevich, V. Saveliev, N. Vyahhi, and G. Tesler. QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, 29(8):1072–1075, Apr 2013.
24. A. Mikheenko, A. Prjibelski, V. Saveliev, D. Antipov, and A. Gurevich. Versatile genome assembly evaluation with QUAST-LG. *Bioinformatics*, 34(13):i142–i150, 07 2018.
25. R. A. Hoskins, J. W. Carlson, K. H. Wan, S. Park, I. Mendez, S. E. Galle, B. W. Booth, B. D. Pfeiffer, R. A. George, R. Svirskas, M. Krzywinski, J. Schein, M. C. Accardo, E. Damia, G. Messina, M. Méndez-Lago, B. de Pablos, O. V. Demakova, E. N. Andreyeva, L. V. Boldyreva, M. Marra, A. B. Carvalho, P. Dimitri, A. Villasante, I. F. Zhimulev, G. M. Rubin, G. H. Karpen, and S. E. Celniker. The Release 6 reference sequence of the Drosophila melanogaster genome. *Genome Res*, 25(3):445–458, Mar 2015.

Appendix B

Grants awarded and peripheral scientific activities

B.1 The journey of my PhD towards bioinformatics

Figure B.1 presents a timeline with the main milestones: the challenging start of this thesis project

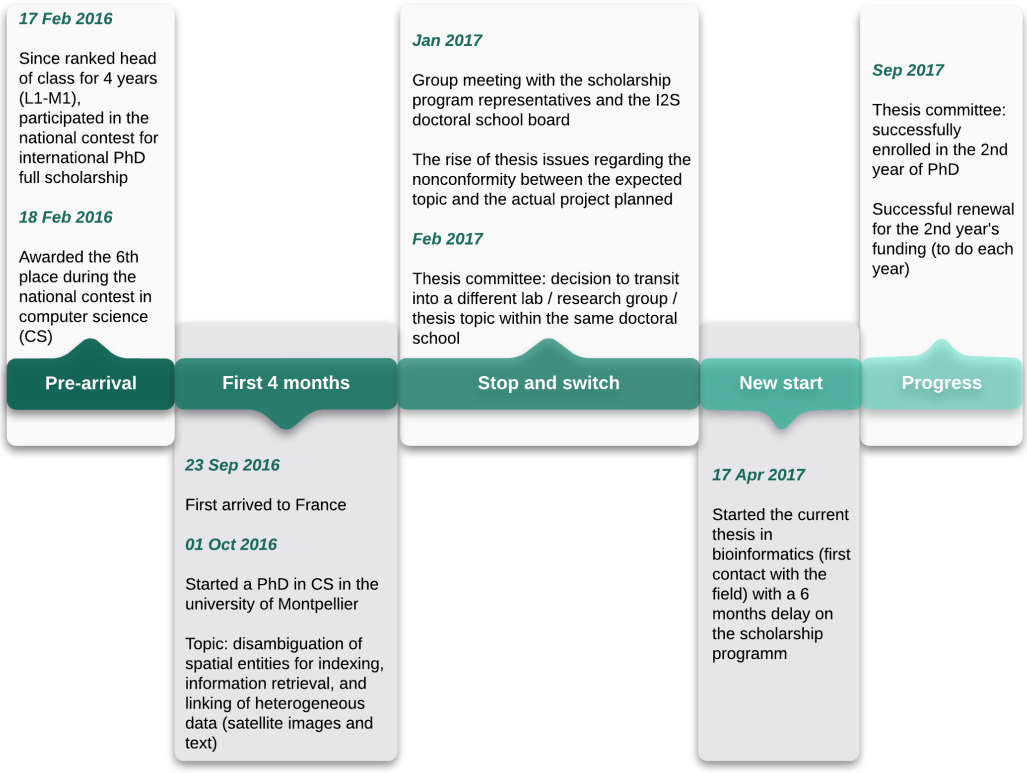


FIGURE B.1: Timeline from the scholarship contest to the steady progress in bioinformatics.

B.2 Scholarships, fellowships and merit grants awarded

ISCB International Society for Computational Biology and Bioinformatics https://www.iscb.org/, fellowship to cover registration fees for ECCB European Conference on Computational Biology, virtual, Barcelona, 2020 https://www.bsc.es/news/events/virtual-19th-european-conference-computational-biology-eccb2020

EBI European Bioinformatics Institute https://www.ebi.ac.uk/, selection and acceptance for a workshop on analysing NGS data and presenting a poster, last-minute cancel due to visa delay, 2018

ERJ (Équipe de Recherche Junior) a grant to build a Junior Research Team funded by the Labex CeMEB https://www.labex-cemeb.org/fr/formation/equipes-de-recherche-juni 2017

ISCD Institut des Sciences du Calcul et des Données https://iscd.sorbonne-universite.fr/academics/schools/ summer school "Scientific trends at the interfaces" at Station de biologie marine in Roscoff, 2017

Figure B.2 shows a timeline for the various scientific opportunities enriching this thesis project, especially in terms of the grants awarded, and highlighting the before and during Covid-19 pandemic years.

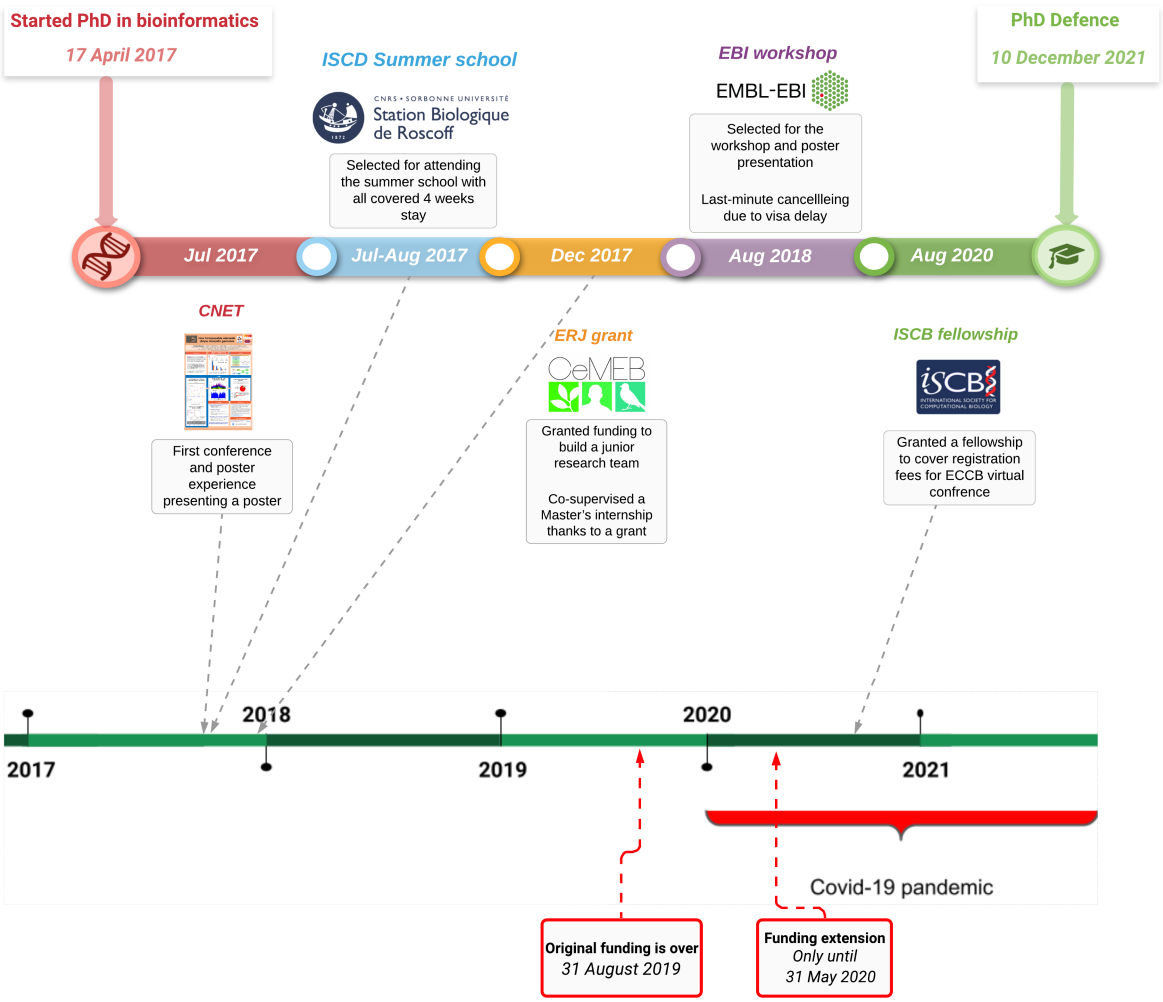


FIGURE B.2: Timeline for the grants awarded.

B.3 Doctoral training > 200 hours

- Scientific courses
- In-person courses
- Virtual live courses
- Online courses
- Additional academia/PhD life courses

Comparative Genomics - lectures - online 9h, (16-19/09/2020), Swiss Institute of Bioinformatics (SIB - Streamed from Lausanne)

Cérémonie solennelle 2019 de rentrée des doctorants 6h, (31/01/2019), Collège Doctoral de l'Université de Montpellier

English course 15h, (Mars-June, 2018), LIRMM

FLE - Français Langue Étrangère 30h, (January-April, 2018), Collège Doctoral de l'université de Montpellier

ISCD Summer School : Bioinformatics and Visual Data Analysis Merit scholarship 120h, (17/07– 11/08/2017), Marine station CNRS / UPMC, Roscoff

GATB programming day : The Genome Analysis Toolbox with de-Bruijn graph 7h, (16/06/2017), IBC Montpellier

Public speech interactive pedagogy level 2 7h, (11/05/2017), Collège Doctoral de l'université de Montpellier

Dealing with scientific literature: efficient reading and good note taking habits 25h, (09/01/2017), Collège Doctoral de l'université de Montpellier

B.4 Co-supervision experience

Co-supervised a Master's internship student thanks to a grant from Labex CeMEB (6K€)

- ERJ project : 1st edition offering funding to build a junior research team, 2018;
- based on a selected research project;
- as a support for the PhD student's research (hardware, conference fees, etc.);
- gratification of the intern Rémy COSTA (M1 BCD);
- 4 months: April-August 2018.

Co-supervised a group of 4 undergraduate students in computer science: TER 2019

B.5 Organizing and chairing experience

SDD ISEM 2020: Semaine Des Doctorants (PhD students week)

Appendix C

Case study

C.1 First identification of chromatin regions in *Cx. pipiens* genome

The *Cx. pipiens CpipJ3* genome is also composed of three chromosome-length scaffolds of 75, 213, and 195Mb length, with 16, 19, and 23 markers, respectively. By consequence, the current genetic maps provide a very low density of 0.14 markers/Mb (see Figure C.1-right). With such low marker density, BREC fails to return accurate chromatin boundaries (see Figure C.1-left).

We show in Table C.1 a summary of both boundaries identification and examine at the same time the quality of the recombination rate estimate. It appears that for *Cx. pipiens*, the RR estimate is equal to zero on Chromosomes 1 and 3, whereas the estimate for Chromosome 2 is also mostly equal to zero. We explain this by the sensitivity of the regression model (loess) to the paucity of data (here the low density of markers). Thus, to examine the TE content and distribution, we choose to come back to a simpler model, the polynomial regression model, which is exploited in next section.

Raw data (no cleaning)		BREC HCB (Mb)				Plot Interpretation	
		centro left	centro right	telo left	telo right	HCB	RR
Aedes Aegypti	*AaegL4* (Dudchenko et al., 2017)						
	chr1	136.51	189.50	58.15	268.12	correct	correct
	chr2	254.97	272.73	48.29	504.68	correct	correct
	chr3	197.97	198.39	23.80	411.59	correct	correct
Culex pipiens	*CpipJ3* (Dudchenko et al., 2017)						
	chr1	74.59	100.06	115.95	47.54	partially correct	RR=0
	chr2	38.56	219.99	7.75	196.05	partially correct	partially correct
	chr3	90.24	158.28	64.22	34.43	partially correct	RR = 0

TABLE C.1: Summary table of resulting boundaries estimates with default regression model, on *Ae. aegypti* and *Cx. pipiens* genomes.

C.2 Analysis of the distribution of TEs along the *Cx. pipiens* genome

To analyse the TE distribution in genomes, we have first to identify and qualify repeats along the sequence. This task present an intrinsic complexity, due at the

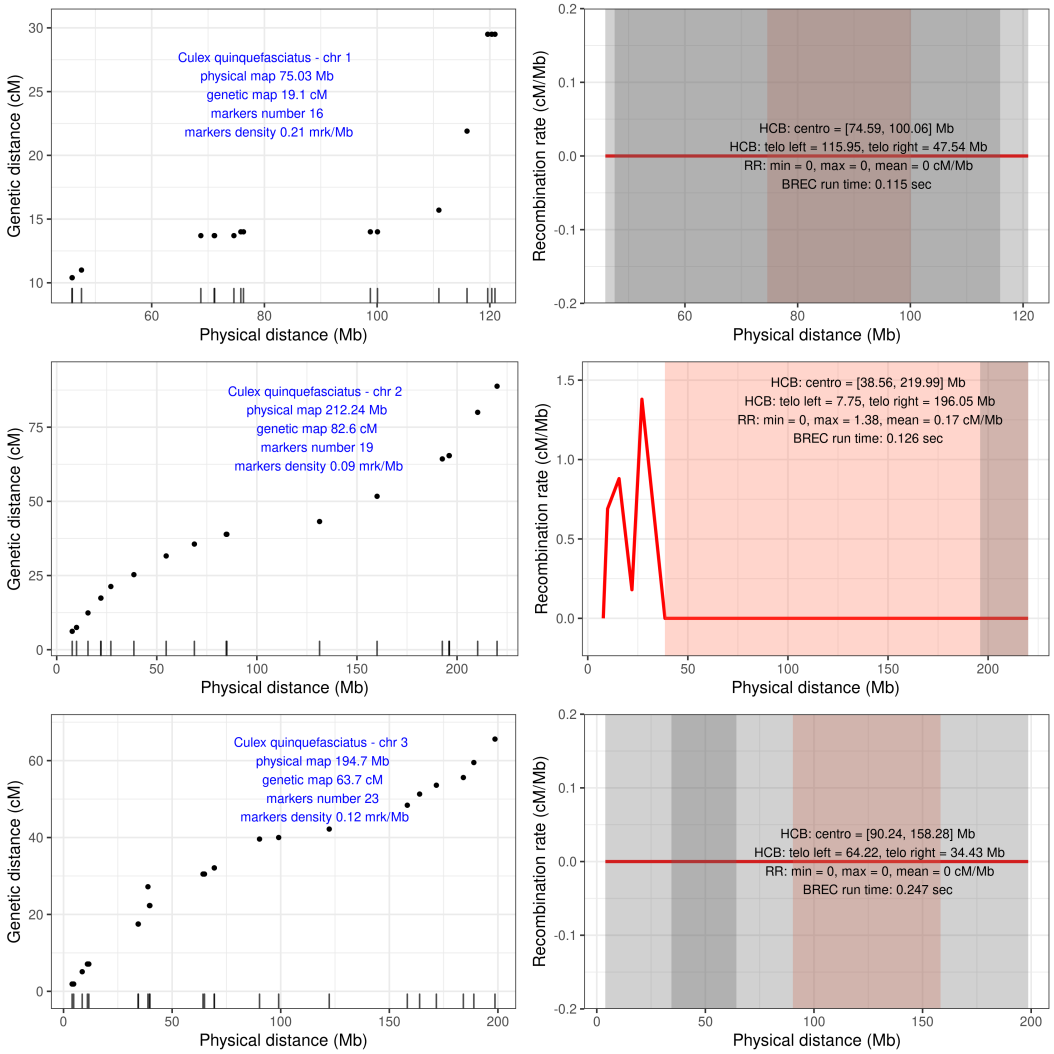


FIGURE C.1: Genomic features (right) and BREC results (left) for the *Cx. pipiens CpipJ3* genome.

same time to the diverse structural nature of these repeats, and how TE insertions might be either correctly or ambiguously annotated on a reference genome for the purpose of identification.

To perform the annotation task, we choose the tools which are mostly used in the repeat community: RepeatMasker and RepeatModeler[1]. RepeatModeler lists all repetitive regions in a genome, then build clusters corresponding to RR families and produce a consensus sequence for each cluster. Its output consists in a multifasta file containing the annotated consensus sequences. RepeatMasker uses this file to annotate the genome with relaxed parameters in order to find even more divergent elements. Though, these tools are not specific TEs which are on our focus. Repbase update (Bao, Kojima, and Kohany, 2015) also provides information on the nature of repetitive sequences, which allows to identify TEs among other repeats. A database specific to TEs is also available: TEfam, which is part of Dfam database [2].

C.2.1 *Cx. pipiens*: a genome enriched in DNA elements

We annotated individual TE insertions in *Cx. pipiens*. We built a *Culex* specific TE library, a set of canonical sequences representative of TE families in this genome. We then annotated them combining results from homology-based (TEfam database: `https://tefam.biochem.vt.edu/`) and signature-based approaches. We reported a high diversity with TE families from the three main types of TEs (DNA, LTR, non-LTR) (see Figure C.2).

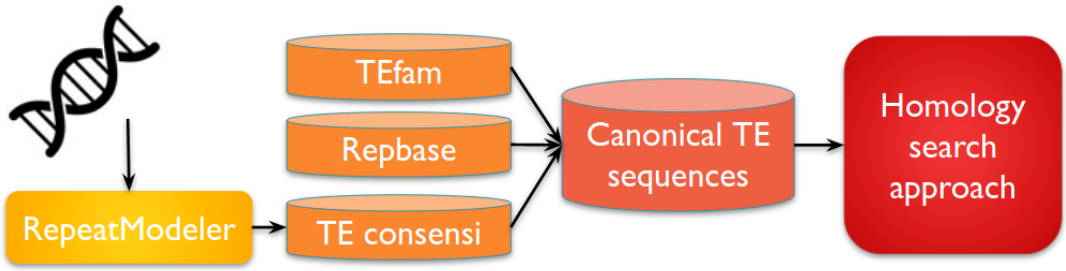


FIGURE C.2: *Culex* TE database construction pipeline.

Our results suggest that *Cx. pipiens* genome is enriched in DNA TEs: More than two-third of the genome is composed of DNA TEs, and among them, around 75% of the DNA elements are MITEs. Figure C.3 shows the distribution of TEs by class.

C.2.2 Centromeres are enriched in one type of TEs: LINE elements

Our analysis show a non-homogeneous distribution of TEs along the *Cx. pipiens* chromosomes, with an enrichment of TEs in the heterochromatin (see Figure C.4).

We re-launched BREC with a polynomial regression model, and we ended up with more accurate heterochromatic boundaries (red dotted lines for centromeric boundaries and green dotted lines for telomeric boundaries). Interestingly, we observe that TE distribution varies depending on the TE type (Class I *Vs.* Class II). Class

[1] `https://www.repeatmasker.org/`
[2] `https://www.dfam.org/home`

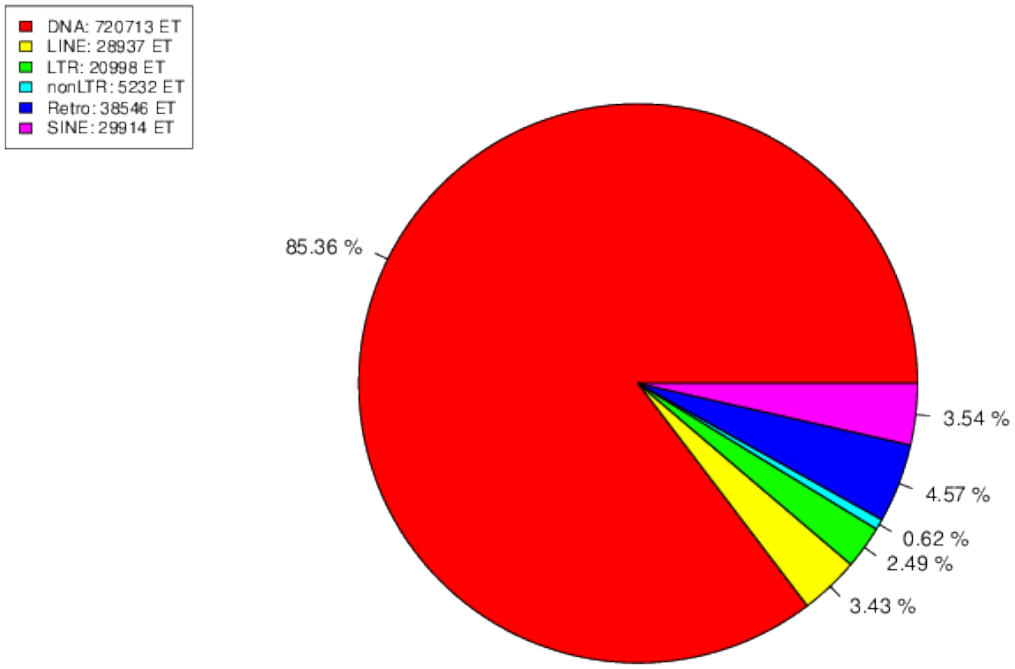


FIGURE C.3: Distribution of transposable elements by class in the *Cx. pipiens* genome.

I elements appear to be enriched in centromeric regions, while Class II elements (specifically MITEs) are enriched in euchromatic regions. This difference reinforce the quality of BREC's eu-heterochromatic boundaries estimates (see Figure C.4).

If we focus only on the more abundant TE families: LINEs (Class I) and MITEs (Class II) elements, we also observe a drastic TE difference in distribution (see Figures C.5 and C.6).

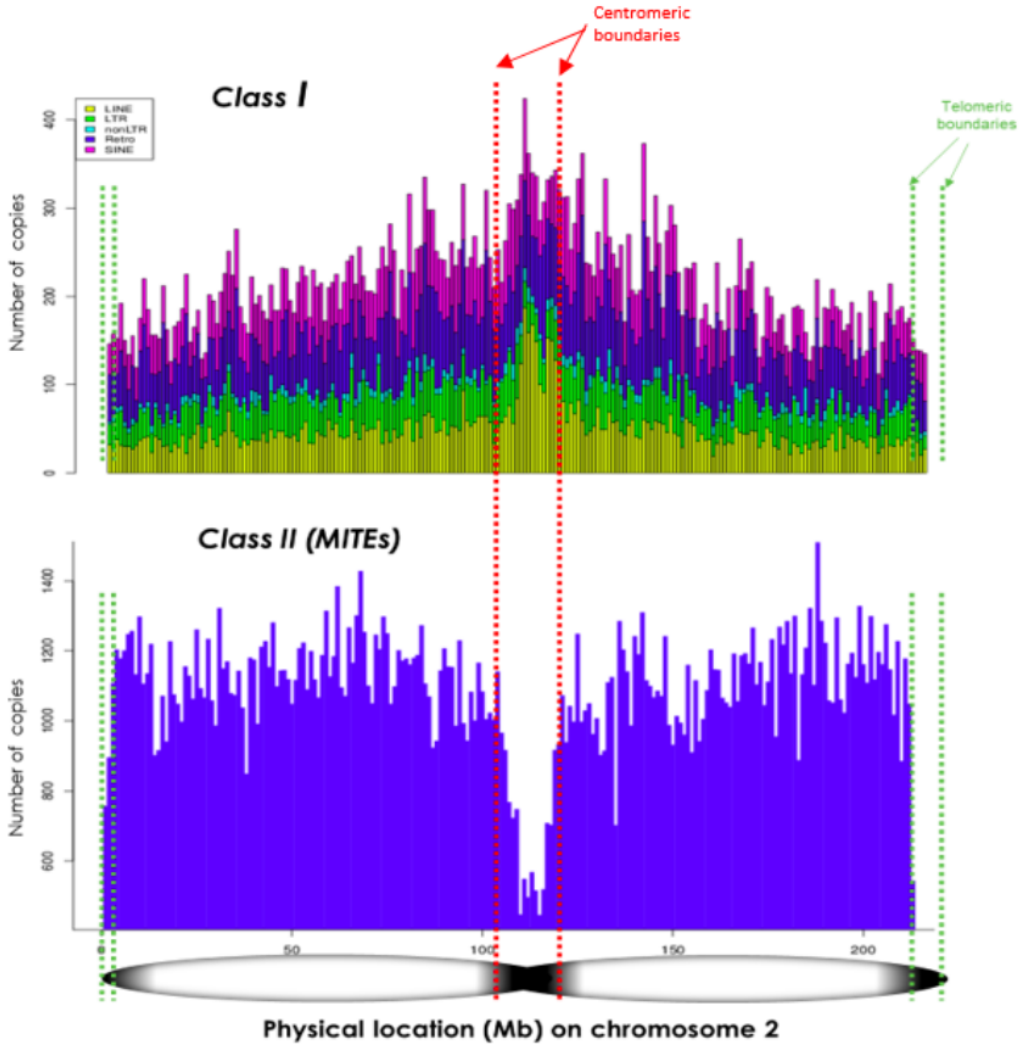


FIGURE C.4: Correlation of TEs distribution and heterochromatin boundaries in *CpipJ3* - chromosome 2.

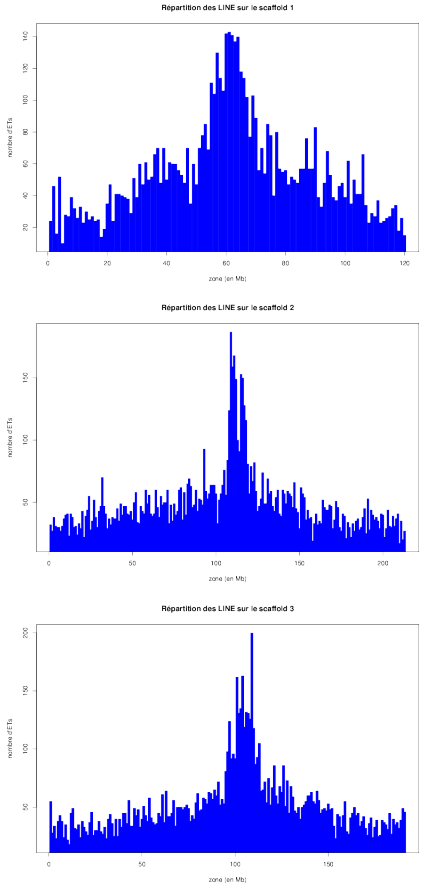


FIGURE C.5: LINEs distribution along *Cx. pipiens* chromosomes.

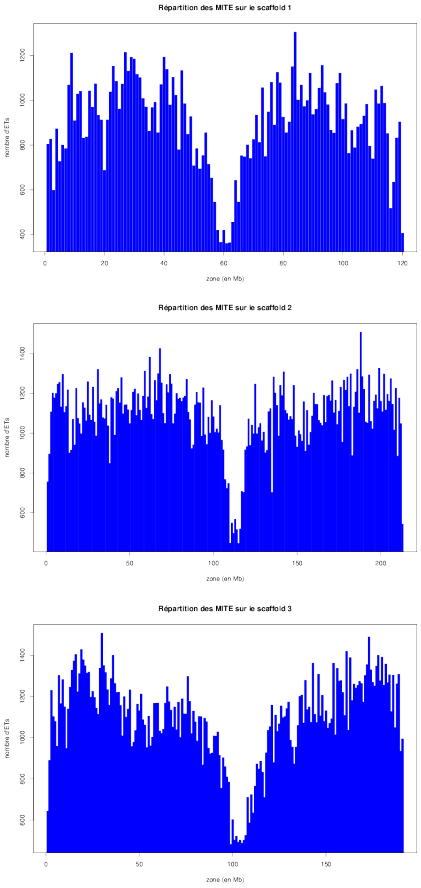


FIGURE C.6: MITEs distribution along *Cx. pipiens* chromosomes.

The TE distribution is the consequence of both TE insertion bias and natural selection against deleterious TE insertions. As both TE families are known to be active in this genome, our results may suggest a TE insertion bias. However, we can exclude that long TE elements insertions like LINEs (around 9kb length) are more deleterious than MITEs (around 100bp length) in gene-rich regions. By consequence, such elements might be rapidly removed by purifying selection.

Bibliography

Agresti, Alan (2007). *An Introduction to Categorical Data Analysis*. Hoboken, NJ, USA: John Wiley & Sons, Inc., pp. 1–356. ISBN: 9780470114759. DOI: 10.1002/0470114754.

Anderson, Sarah N. et al. (2019). "Transposable elements contribute to dynamic genome content in maize". In: *The Plant Journal* 100.5, pp. 1052–1065. ISSN: 0960-7412. DOI: 10.1111/tpj.14489.

Antipov, D. et al. (2016). "hybridSPAdes: an algorithm for hybrid assembly of short and long reads". In: *Bioinformatics* 32.7, pp. 1009–1015.

Auton, Adam and Gil McVean (2012). "Estimating recombination rates from genetic variation in humans". In: *Methods in Molecular Biology* 856, pp. 217–237. DOI: 10.1007/978-1-61779-585-5_9.

Bankevich, Anton et al. (2012). "SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing". In: *Journal of Computational Biology* 19.5, pp. 455–477. DOI: 10.1089/cmb.2012.0021.

Bansal, Vikas and Christina Boucher (2019). "Sequencing Technologies and Analyses: Where Have We Been and Where Are We Going?" In: *iScience* 18, pp. 37–41. ISSN: 25890042. DOI: 10.1016/j.isci.2019.06.035.

Bao, Weidong, Kenji K. Kojima, and Oleksiy Kohany (2015). "Repbase Update, a database of repetitive elements in eukaryotic genomes". In: *Mobile DNA* 6.1, p. 11. DOI: 10.1186/s13100-015-0041-9.

Bernardi, G. (2021). "The "Genomic Code": DNA Pervasively Moulds Chromatin Structures Leaving no Room for "Junk"". In: *Life (Basel)* 11.4.

Biémont, Christian (2010). "A Brief History of the Status of Transposable Elements: From Junk DNA to Major Players in Evolution". In: *Genetics*. DOI: 10.1534/genetics.110.124180.

Bleidorn, Christoph (2017). *Phylogenomics*. Cham: Springer International Publishing, pp. 1–222. ISBN: 978-3-319-54062-7. DOI: 10.1007/978-3-319-54064-1.

Bourque, Guillaume et al. (2018). "Ten things you should know about transposable elements". In: *Genome Biology* 19.1, pp. 1–12. ISSN: 1474-760X. DOI: 10.1186/s13059-018-1577-z.

Butler, J. et al. (2008). "ALLPATHS: de novo assembly of whole-genome shotgun microreads". In: *Genome Res* 18.5, pp. 810–820.

Chakraborty, Mahul et al. (2018). "Hidden genetic variation shapes the structure of functional elements in Drosophila". In: *Nature Genetics* 50.1, pp. 20–25. DOI: 10.1038/s41588-017-0010-y.

Chakravarti, Aravinda (1991). "A graphical representation of genetic and physical maps: The Marey map". In: *Genomics* 11.1, pp. 219–222. DOI: 10.1016/0888-7543(91)90123-V.

Chan, Andrew H., Paul A. Jenkins, and Yun S. Song (2012). "Genome-Wide Fine-Scale Recombination Rate Variation in Drosophila melanogaster". In: *PLoS Genetics* 8.12. Ed. by Gil McVean, e1003090. DOI: 10.1371/journal.pgen.1003090.

Chen, D. et al. (2020). "Human L1 Transposition Dynamics Unraveled with Functional Data Analysis". In: *Mol Biol Evol* 37.12, pp. 3576–3600.

Chikhi, Rayan and Guillaume Rizk (2013). "Space-efficient and exact de Bruijn graph representation based on a Bloom filter". In: *Algorithms for Molecular Biology* 8.1, p. 22. DOI: 10.1186/1748-7188-8-22.

Cleveland, William S. and Susan J. Devlin (1988). "Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting". In: *Journal of the American Statistical Association* 83.403, pp. 596–610. DOI: 10.1080/01621459.1988.10478639.

Cleveland, William S. and Clive Loader (1996). "Smoothing by Local Regression: Principles and Methods". In: *Statistical Theory and Computational Aspects of Smoothing*. Ed. by Wolfgang Härdle and Michael G. Schimek. Heidelberg: Physica-Verlag HD, pp. 10–49. ISBN: 978-3-642-48425-4. DOI: 10.1007/978-3-642-48425-4_2.

Comeron, Josep M., Ramesh Ratnappan, and Samuel Bailin (2012). "The Many Landscapes of Recombination in Drosophila melanogaster". In: *PLoS Genetics* 8.10. Ed. by Dmitri A. Petrov, e1002905. DOI: 10.1371/journal.pgen.1002905.

Compeau, Phillip E C, Pavel A. Pevzner, and Glenn Tesler (2011). "How to apply de Bruijn graphs to genome assembly". In: *Nature Biotechnology* 29.11, pp. 987–991. ISSN: 1087-0156. DOI: 10.1038/nbt.2023.

Consortium, International Human Genome Sequencing (2001). "Initial sequencing and analysis of the human genome". In: *Nature* 409.6822, pp. 860–921. ISSN: 0028-0836. DOI: 10.1038/35057062.

Coop, Graham and Molly Przeworski (2007). "An evolutionary view of human recombination". In: *Nature Reviews Genetics* 8.1, pp. 23–34. ISSN: 1471-0056. DOI: 10.1038/nrg1947.

Corbett-Detig, Russell B., Daniel L. Hartl, and Timothy B. Sackton (2015). "Natural Selection Constrains Neutral Diversity across A Wide Range of Species". In: *PLoS Biology* 13.4. Ed. by Nick H. Barton, e1002112. DOI: 10.1371/journal.pbio.1002112.

Cox, Allison et al. (2009). "A New Standard Genetic Map for the Laboratory Mouse". In: *Genetics* 182.4, pp. 1335–1344. ISSN: 0016-6731. DOI: 10.1534/genetics.109.105486.

Dudchenko, Olga et al. (2017). "De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds". In: *Science* 356.6333, pp. 92–95. ISSN: 0036-8075. DOI: 10.1126/science.aal3327. eprint: http://science.sciencemag.org/content/356/6333/92.full.pdf.

Duret, Laurent and Nicolas Galtier (2009). "Biased Gene Conversion and the Evolution of Mammalian Genomic Landscapes". In: *Annual Review of Genomics and Human Genetics* 10.1, pp. 285–311. ISSN: 1527-8204. DOI: 10.1146/annurev-genom-082908-150001.

Elsik, Christine G. et al. (2014). "Finding the missing honey bee genes: lessons learned from a genome upgrade". In: *BMC Genomics* 15.1, p. 86. ISSN: 1471-2164. DOI: 10.1186/1471-2164-15-86.

Fiston-Lavier, Anna Sophie et al. (2010). "Drosophila melanogaster recombination rate calculator". In: *Gene* 463.1-2, pp. 18–20. DOI: 10.1016/j.gene.2010.04.015.

Foster, Woodbridge A. and Edward D. Walker (2019). "Mosquitoes (Culicidae)". In: Elsevier, pp. 261–325. DOI: 10.1016/B978-0-12-814043-7.00015-7.

Freeman, Jennifer L. et al. (2007). "Definition of the zebrafish genome using flow cytometry and cytogenetic mapping". In: *BMC Genomics* 8.1, p. 195. DOI: 10.1186/1471-2164-8-195.

Freire, B., S. Ladra, and J. R. Parama (2021). "Memory-Efficient Assembly using Flye". In: *IEEE/ACM Trans Comput Biol Bioinform* PP.

Gansner, E. R. and S. C. North (2000). "An open graph visualization system and its applications to software engineering". In: *SOFTWARE - PRACTICE AND EXPERIENCE* 30.11, pp. 1203–1233.

Ghurye, Jay and Mihai Pop (2019). "Modern technologies and algorithms for scaffolding assembled genomes." In: *PLoS computational biology* 15.6. Ed. by Nicola Segata, e1006994. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1006994.

Giani, Alice Maria et al. (2020). *Long walk to genomics: History and current approaches to genome sequencing and assembly*. DOI: 10.1016/j.csbj.2019.11.002.

Gui, Songtao et al. (2018). "Improving Nelumbo nucifera genome assemblies using high-resolution genetic maps and BioNano genome mapping reveals ancient chromosome rearrangements". In: *Plant Journal* 94.4, pp. 721–734. DOI: 10.1111/tpj.13894.

Gurevich, Alexey et al. (2013). "QUAST: Quality assessment tool for genome assemblies". In: *Bioinformatics* 29.8, pp. 1072–1075. DOI: 10.1093/bioinformatics/btt086.

Heydari, Mahdi et al. (2017). "Evaluation of the impact of Illumina error correction tools on de novo genome assembly". In: *BMC Bioinformatics* 18.1, p. 374. DOI: 10.1186/s12859-017-1784-8.

Heydari, Mahdi et al. (2019). "Illumina error correction near highly repetitive DNA regions improves de novo genome assembly". In: *BMC Bioinformatics* 20.1, p. 298. DOI: 10.1186/s12859-019-2906-2.

Hillier, La Deana W. et al. (2008). "Whole-genome sequencing and variant discovery in C. elegans". In: *Nature Methods* 5.2, pp. 183–188. DOI: 10.1038/nmeth.1179.

Holt, Robert A. et al. (2002). "The Genome Sequence of the Malaria Mosquito Anopheles gambiae". In: *Science* 298.5591, pp. 129–149. ISSN: 0036-8075. DOI: 10.1126/science.1076181.

Hoskins, Roger A et al. (2007). "Sequence finishing and mapping of Drosophila melanogaster heterochromatin". In: *Science* 316.5831, pp. 1625–1628. DOI: 10.1126/science.1139816.

Hoskins, Roger A et al. (2015). "The Release 6 reference sequence of the Drosophila melanogaster genome". In: *Genome Research* 25.3, pp. 445–458. DOI: 10.1101/gr.185579.114.

Huang, Cheng Ran Lisa, Kathleen H. Burns, and Jef D. Boeke (2012). "Active Transposition in Genomes". In: *Annual Review of Genetics* 46.1, pp. 651–675. ISSN: 0066-4197. DOI: 10.1146/annurev-genet-110711-155616.

Jackman, S. D. et al. (2017). "ABySS 2.0: resource-efficient assembly of large genomes using a Bloom filter". In: *Genome Res* 27.5, pp. 768–777.

Jeffreys, A. J. (2000). "High resolution analysis of haplotype diversity and meiotic crossover in the human TAP2 recombination hotspot". In: *Human Molecular Genetics* 9.5, pp. 725–733. DOI: 10.1093/hmg/9.5.725.

Juneja, Punita et al. (2014). "Assembly of the Genome of the Disease Vector Aedes aegypti onto a Genetic Linkage Map Allows Mapping of Genes Affecting Disease Transmission". In: *PLoS Neglected Tropical Diseases* 8.1, p. 8. ISSN: 19352727. DOI: 10.1371/journal.pntd.0002652.

Jung, Hyungtaek et al. (2020). "Twelve quick steps for genome assembly and annotation in the classroom". In: *PLOS Computational Biology* 16.11. Ed. by Francis Ouellette, e1008325. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1008325.

Katz, Daniel S., Morane Gruenpeter, and Tom Honeyman (2021). "Taking a fresh look at FAIR for research software". In: *Patterns* 2.3. DOI: 10.1016/j.patter.2021.100222.

Kent, Tyler V, Jasmina Uzunović, and Stephen I Wright (2017). "Coevolution between transposable elements and recombination". In: *Philosophical Transactions of the Royal Society B* 372.1736, p. 20160458. ISSN: 0962-8436. DOI: 10.1098/RSTB.2016.0458.

Langley, Charles H. et al. (2012). "Genomic variation in natural populations of Drosophila melanogaster". In: *Genetics* 192.2, pp. 533–598. DOI: 10.1534/genetics.112.142018.

Lenormand, Thomas et al. (2016). *Evolutionary mysteries in meiosis*. DOI: 10.1098/rstb.2016.0001.

Levan, Albert, Karl Fredga, and Avery A. Sandberg (1964). "Nomenclature for centromeric position on chromosomes". In: *Hereditas* 52.2, pp. 201–220. DOI: 10.1111/j.1601-5223.1964.tb01953.x.

Li, Heng (2018). "Minimap2: Pairwise alignment for nucleotide sequences". In: *Bioinformatics* 34.18, pp. 3094–3100. DOI: 10.1093/bioinformatics/bty191. arXiv: 1708.01492.

Li, Heng and Richard Durbin (2009). "Fast and accurate short read alignment with Burrows-Wheeler transform". In: *Bioinformatics* 25.14, pp. 1754–1760. DOI: 10.1093/bioinformatics/btp324.

Li, Z. et al. (2012). "Comparison of the two major classes of assembly algorithms: overlap-layout-consensus and de-bruijn-graph". In: *Briefings in Functional Genomics* 11.1, pp. 25–37. ISSN: 2041-2649. DOI: 10.1093/bfgp/elr035.

Logsdon, Glennis A., Mitchell R. Vollger, and Evan E. Eichler (2020). "Long-read human genome sequencing and its applications". In: *Nature Reviews Genetics* 21.10, pp. 597–614. ISSN: 1471-0056. DOI: 10.1038/s41576-020-0236-x.

Lu, Min and Xiangwei He (2019). "Centromere repositioning causes inversion of meiosis and generates a reproductive barrier". In: *Proceedings of the National Academy of Sciences* 116.43, pp. 21580–21591. ISSN: 0027-8424. DOI: 10.1073/pnas.1911745116.

Luo, Junwei et al. (2021). "A comprehensive review of scaffolding methods in genome assembly". In: *Briefings in Bioinformatics*. ISSN: 1467-5463. DOI: 10.1093/bib/bbab033.

Mandric, I. et al. (2016). "Scaffolding Algorithms". In: *Computational Methods for Next Generation Sequencing Data Analysis*. Ed. by I. Măndoiu and A. Zelikovsky. Wiley. Chap. 5, pp. 107–132.

Mansour, Y., A. Chateau, and AS. Fiston-Lavier (2021). "BREC: an R package/Shiny app for automatically identifying heterochromatin boundaries and estimating local recombination rates along chromosomes". In: *BMC Bioinformatics* 22.S6, p. 396. ISSN: 1471-2105. DOI: 10.1186/s12859-021-04233-1.

Matthews, Benjamin J. et al. (2018). "Improved reference genome of Aedes aegypti informs arbovirus vector control". In: *Nature* 563.7732, pp. 501–507. ISSN: 14764687. DOI: 10.1038/s41586-018-0692-z.

McClintock, B. (1950). "The origin and behavior of mutable loci in maize." In: *Proceedings of the National Academy of Sciences of the United States of America* 36.6, pp. 344–355. ISSN: 00278424. DOI: 10.1073/pnas.36.6.344.

McCullers, Tabitha J. and Mindy Steiniger (2017). "Transposable elements in Drosophila". In: *Mobile Genetic Elements* 7.3, pp. 1–18. ISSN: 2159-256X. DOI: 10.1080/2159256X.2017.1318201.

McGinnis, Scott and Thomas L Madden (2004). "BLAST: At the core of a powerful and diverse set of sequence analysis tools". In: *Nucleic Acids Research* 32.WEB SERVER ISS. W20–5. ISSN: 03051048. DOI: 10.1093/nar/gkh435.

Melo, Elverson and Gabriel Wallau (2020). "Mosquito genomes are frequently invaded by transposable elements through horizontal transfer". In: *PLoS Genetics* 16. DOI: 10.1371/journal.pgen.1008946.

Mikheenko, Alla et al. (2018). "Versatile genome assembly evaluation with QUAST-LG". In: *Bioinformatics*. Vol. 34. 13, pp. i142–i150. DOI: 10.1093/bioinformatics/bty266.

Morata, Jordi et al. (2018). "The evolutionary consequences of transposon-related pericentromer expansion in melon". In: *Genome Biology and Evolution* 10.6. Ed. by Richard Cordaux, pp. 1584–1595. ISSN: 17596653. DOI: 10.1093/gbe/evy115.

Mukherjee, S. et al. (2019). "Whole genome sequence and de novo assembly revealed genomic architecture of Indian Mithun (Bos frontalis)". In: *BMC Genomics* 20.1, p. 617.

Muller, Héloïse, José Gil, and Ines Anna Drinnenberg (2019). "The Impact of Centromeres on Spatial Genome Architecture". In: *Trends in Genetics* 35.8, pp. 565–578. DOI: 10.1016/j.tig.2019.05.003.

Murigneux, Valentine et al. (2020). "Comparison of long-read methods for sequencing and assembly of a plant genome". In: *GigaScience* 9.12. giaa146. ISSN: 2047-217X. DOI: 10.1093/gigascience/giaa146.

Mölder, F. et al. (2021). "Sustainable data analysis with Snakemake". In: *F1000Res* 10, p. 33.

Nagarajan, Niranjan and Mihai Pop (2013). *Sequence assembly demystified*. DOI: 10.1038/nrg3367.

Navarro, Caryn (2017). "The Mobile World of Transposable Elements". In: *Trends in Genetics* 33.11, pp. 771–772. DOI: 10.1016/j.tig.2017.09.006.

Nurk, Sergey et al. (2020). "HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads". In: *Genome Research* 30.9, pp. 1291–1305. ISSN: 1088-9051. DOI: 10.1101/gr.263566.120.

Peñalba, Joshua V. and Jochen B. W. Wolf (2020). "From molecules to populations: appreciating and estimating recombination rate variation". In: *Nature Reviews Genetics* 21.8, pp. 476–492. ISSN: 1471-0056. DOI: 10.1038/s41576-020-0240-1.

Peona, Valentina, Matthias H. Weissensteiner, and Alexander Suh (2018). *How complete are "complete" genome assemblies?—an avian perspective*. DOI: 10.1111/1755-0998.12933.

Petersen, Malte et al. (2019). "Diversity and evolution of the transposable element repertoire in arthropods with particular reference to insects". In: *BMC Evolutionary Biology* 19.1, p. 11. ISSN: 1471-2148. DOI: 10.1186/s12862-018-1324-9.

Petrov, D. A. et al. (2011). "Population Genomics of Transposable Elements in Drosophila melanogaster". In: *Molecular Biology and Evolution* 28.5, pp. 1633–1644. ISSN: 0737-4038. DOI: 10.1093/molbev/msq337.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Ravindran, Sandeep (2012). "Barbara McClintock and the discovery of jumping genes". In: *Proceedings of the National Academy of Sciences* 109.50, pp. 20198–20199. ISSN: 0027-8424. DOI: 10.1073/pnas.1219372109.

Rezvoy, Clément et al. (2007). "MareyMap: An R-based tool with graphical interface for estimating recombination rates". In: *Bioinformatics* 23.16, pp. 2188–2189. DOI: 10.1093/bioinformatics/btm315.

Rice, Edward S and Richard E Green (2019). "New Approaches for Genome Assembly and Scaffolding". In: *Annual Review of Animal Biosciences* 7, pp. 17–40. DOI: 10.1146/annurev-animal-020518.

Riddle, Nicole C et al. (2011). "Plasticity in patterns of histone modifications and chromosomal proteins in Drosophila heterochromatin". In: *Genome Research* 21.2, pp. 147–163. DOI: 10.1101/gr.110098.110.

Rizzon, Carène et al. (2002). "Recombination Rate and the Distribution of Transposable Elements in the Drosophila melanogaster Genome". In: *Genome Research* 12, pp. 400–407. ISSN: 1088-9051. DOI: 10.1101/gr.210802..

Robert L. Nussbaum, M.D.F.F., R.R. McInnes, and H.F. Willard (2015). *Thompson & Thompson Genetics in Medicine*. Saunders W.B. Elsevier Health Sciences. ISBN: 9781437706963.

Rowan, Beth A et al. (2019). "An Ultra High-Density Arabidopsis thaliana Crossover Map That Refines the Influences of Structural Variation and Epigenetic Features". In: *Genetics* 213.3, genetics.302406.2019. ISSN: 0016-6731. DOI: 10.1534/genetics.119.302406.

RStudio, Inc (2014). *Shiny: Easy web applications in R*. URL: http://shiny.rstudio.com.

Sardell, Jason M. and Mark Kirkpatrick (2020). "Sex differences in the recombination landscape". In: *American Naturalist* 195.2, pp. 361–379. ISSN: 00030147. DOI: 10.1086/704943.

Sato, Shusei et al. (2012). "The tomato genome sequence provides insights into fleshy fruit evolution". In: *Nature* 485.7400, pp. 635–641. DOI: 10.1038/nature11119.

Schueler, M G et al. (2001). "Genomic and genetic definition of a functional human centromere". In: *Science* 294.5540, pp. 109–115. DOI: 10.1126/science.1065042.

Shen, Chao et al. (2017). "Genome-wide recombination rate variation in a recombination map of cotton". In: *PLoS ONE* 12.11, e0188682. DOI: 10.1371/journal.pone.0188682.

Siberchicot, Aurélie et al. (2017). "MareyMap Online: A User-Friendly Web Application and Database Service for Estimating Recombination Rates Using Physical and Genetic Maps". In: *Genome Biology and Evolution* 9.10, pp. 2506–2509. ISSN: 1759-6653. DOI: 10.1093/gbe/evx178.

Sievert, Carson (2020). *Interactive Web-Based Data Visualization with R, plotly, and shiny*. Chapman and Hall/CRC. ISBN: 9781138331457. URL: https://plotly-r.com.

Silva-Junior, Orzenil B. and Dario Grattapaglia (2015). "Genome-wide patterns of recombination, linkage disequilibrium and nucleotide diversity from pooled resequencing and single nucleotide polymorphism genotyping unlock the evolutionary history of Eucalyptus grandis". In: *New Phytologist* 208.3, pp. 830–845. DOI: 10.1111/nph.13505.

Sohn, Jang Il and Jin Wu Nam (2018). "The present and future of de novo whole-genome assembly". In: *Briefings in Bioinformatics* 19.1, pp. 23–40. ISSN: 14774054. DOI: 10.1093/bib/bbw096.

Stapley, Jessica et al. (2017). "Variation in recombination frequency and distribution across eukaryotes: patterns and processes". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 372.1736, p. 20160455. ISSN: 0962-8436. DOI: 10.1098/rstb.2016.0455.

Stumpf, Michael P.H. and Gilean A.T. McVean (2003). "Estimating recombination rates from population-genetic data". In: *Nature Reviews Genetics* 4.12, pp. 959–968. DOI: 10.1038/nrg1227.

Termolino, Pasquale et al. (2016). "Insights into epigenetic landscape of recombination-free regions". In: *Chromosoma* 125.2, pp. 301–308. DOI: 10.1007/s00412-016-0574-9.

Thurmond, Jim et al. (2019). "FlyBase 2.0: The next generation". In: *Nucleic Acids Research* 47.D1, pp. D759–D765. DOI: 10.1093/nar/gky1003.

Treangen, Todd and Steven Salzberg (2011). "Repetitive DNA and next-generation sequencing: Computational challenges and solutions". In: *Nature reviews Genetics* 13, pp. 36–46. DOI: 10.1038/nrg3117.

Vanrobays, Emmanuel, Mélanie Thomas, and Christophe Tatout (2017). "Heterochromatin Positioning and Nuclear Architecture". In: *Annual Plant Reviews online*. Chichester, UK: John Wiley & Sons, Ltd, pp. 157–190. DOI: 10.1002/9781119312994.apr0502.

Wajid, Bilal and Erchin Serpedin (2016). "Do it yourself guide to genome assembly". In: *Briefings in Functional Genomics* 15.1, pp. 1–9. ISSN: 2041-2657. DOI: 10.1093/bfgp/elu042.

Wee, Yongkiat et al. (2019). "The bioinformatics tools for the genome assembly and analysis based on third-generation sequencing". In: *Briefings in Functional Genomics* 18.1, pp. 1–12. ISSN: 2041-2649. DOI: 10.1093/bfgp/ely037.

Weinstock, George M. et al. (2006). "Insights into social insects from the genome of the honeybee Apis mellifera". In: *Nature* 443.7114, pp. 931–949. DOI: 10.1038/nature05260. arXiv: NIHMS150003.

Weller, Mathias, Annie Chateau, and Rodolphe Giroudeau (2015). "Exact approaches for scaffolding." In: *BMC bioinformatics* 16 Suppl 1.Suppl 14, S2. ISSN: 1471-2105. DOI: 10.1186/1471-2105-16-S14-S2.

Wells, Jonathan N. and Cédric Feschotte (2020). "A Field Guide to Eukaryotic Transposable Elements". In: *Annual Review of Genetics* 54.1, annurev–genet–040620–022145. ISSN: 0066-4197. DOI: 10.1146/annurev-genet-040620-022145.

Zerbino, Daniel R. and Ewan Birney (2008). "Velvet: Algorithms for de novo short read assembly using de Bruijn graphs". In: *Genome Research* 18.5, pp. 821–829. ISSN: 10889051. DOI: 10.1101/gr.074492.107.

Zhang, Dabao (2017). "A Coefficient of Determination for Generalized Linear Models". In: *American Statistician* 71.4, pp. 310–316. DOI: 10.1080/00031305.2016.1256839.

Zimin, Aleksey V. et al. (2013). "The MaSuRCA genome assembler". In: *Bioinformatics* 29.21, pp. 2669–2677. ISSN: 13674803. DOI: 10.1093/bioinformatics/btt476.