



HAL
open science

Methods for genome-wide structural variation inference in skimming data using graph pangenome

Thi Minh Nguyet Dang

► **To cite this version:**

Thi Minh Nguyet Dang. Methods for genome-wide structural variation inference in skimming data using graph pangenome. Genetics. Université de Montpellier, 2022. English. NNT : 2022UMONG079 . tel-04067063

HAL Id: tel-04067063

<https://theses.hal.science/tel-04067063v1>

Submitted on 13 Apr 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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

En Génétique et Génomique

École doctorale n°584 GAIA : Biodiversité, Agriculture, Alimentation, Environnement, Terre, Eau

Unité de recherche DIADE : Diversité, Adaptation et Développement des Plantes

Méthodes d'inférence des variations structurelles à l'échelle du
génomme dans les données séquençage basse profondeur à l'aide du
graphe de pangénomme

Methods for genome-wide structural variation inference in skimming
data using graph pangenome

Présentée par DANG Thi Minh Nguyet
le 20 septembre 2022

Sous la direction de Dr. François SABOT

Devant le jury composé de

M. Rayan CHIKHI, Chercheur, Institut Pasteur

M. Matthias ZYTNIKI, Chargé de recherche, INRAE,MIAT

Mme. Malika AINOUCHE, Professeur, Université de Rennes 1

Mme. Clementine VITTE, Chargée de recherche, Université Paris-Saclay, INRAE, CNRS,
AgroParisTech

M. Jean-Christophe GLASZMANN, Chercheur, CIRAD, AGAP

M. François SABOT, Directeur de recherche, IRD Montpellier

Rapporteur

Rapporteur

Examinatrice

Examinatrice

Examineur

Directeur de thèse



UNIVERSITÉ
DE MONTPELLIER

Acknowledgement

First and foremost, I would like to show my greatest appreciation to my supervisor, Dr. François Sabot, who always supports me throughout my research. We have been working together since my Master 2 Internship, when I discovered the pangenome concept the first time. Even our discussions are often short in term of duration, it is always full of ideas, creativity and joy. Thank you for inspiring me to be a researcher, who want to look at any matters from different perspectives. Beside being an amazing supervisor, he provides me guidance to many aspects of life. It is not easy to stay in a foreign country, especially during the pandemic era, but I truly enjoy my 3 year PhD thank to him.

Also, I would like to take this opportunity to thank the RICE team and then later DYNADIV team at UMR DIADE, IRD Montpellier for providing me a good research environment. DYNADIV is a dynamic team consisting of many researchers, technicians, postdocs and students who I can find to exchange knowledge on many research topics. I want to thank Éloi Durant, my PhD binôme, who always "got my back" and supports me whenever I need (also thank you for bearing my craziness and stupidity sometimes, or many times). I want to thank Christine Dubreuil Tranchant, Julie Orjuela and Clothilde Chenal who works in the domain and can exchange their knowledge with me. Thank you Dr. Guillaume Gautreau for your interest in my research topics and always propose questions when I were in conferences.

For my collaborators, I would like to show my appreciation to Dr. Séverine Chambeyron, Dr. Yuki Ogyama, Mourdas Mohamed from Institute of Human Genetics at CNRS, and others for the opportunity to work with an interesting model, *Drosophila* and transposable elements. I learned a lot and improve my research skill sets through our work. In addition, I also want to thank my friends and colleagues at N2TP Technology Solutions, especially Tuan Do, for long nights of programming and optimization together. N2TP team provided me change to work in a collaborative development environment.

Furthermore, I also want to send my gratitude to Dr. Jean-Christophe Glaszmann (CIRAD, AGAP), Dr. Rayan Chikhi (Institut Pasteur), Dr. Matthias Zytnicki (INRAE, MIAT), Prof. Malika Ainouche (Université de Rennes 1), Dr. Clementine Vitte (Université Paris-Saclay, INRAE, CNRS, AgroParisTech) for their kindness to be members in my thesis defense.

For financial support, I want to show my gratitude to France Excellence Scholarship and French Embassy in Vietnam to providing me the ability to pursuit my PhD.

Personally, I want to express my heartfelt appreciation to my family members: my parents, my brother and sister, my niece and nephew for loving me and supporting me unconditionally.

I want to express my gratitude towards all my friends for supporting me or bearing my "annoyance" during this 3 years. Thank you for sharing the same roof with me: Hong Anh, Trung, Nam and Tram. Thank you for somehow doing your PhD in the same time (or about to experience the PhD life) and share the same rhyme of life with me: Ha, Thai, Minh and Viet Anh. Thank you for working together and experience new challenges with me during these years: Tuan, Nhung and Phong. I also want to thank other friends and colleagues who often being around in Montpellier to share lunch, dinner or picnic time

together: Ai My, Ngan, Hieu, Doaa, Trixie, Rayan, Sonia, Francis, Vincent, Patrick and Valerie... And also for other friends who never tired of telling "Finish your PhD first and comeback home, we bring you to good places to eat and to play": Hoang Anh, Quynh Anh, Phuong Anh (suddenly I recognize that your first name is always "Anh"). Thank you, Chi Chi, for your presence in many restless nights running for deadline together, I hope that you will graduate soon.

Thank you this 3 year PhD, to show me new experience of doing research and to show me that I have love and support from everyone around.

So, thank you !!!

P.s: Sending thanks to all the lovely cats around my life, Mun, Shirel and Muoi

Résumé

Pour comparer plusieurs génomes, un génome de référence linéaire a souvent été utilisé comme système de coordonnées pour décrire les gènes, les variations et autres annotations fonctionnelles entre individus. Cependant, il a été démontré que cette référence unique n'était pas suffisante pour appréhender toutes les variations génomiques existantes telles que les variations du nombre de copies (CNV), les variations de présence/absence (PAV) ou les variations structurales (SV) de manière plus générale. Pour surmonter cette limitation, le concept de pangénome, composé d'un génome core et d'un génome accessoire, a été appliqué pour étudier un groupe de génomes.

Le modèle de données basé sur un graphe généré par l'incorporation incrémentale des informations d'alignement de génomes est l'une des nouvelles approches pour représenter les informations du pangénome. Un graphe de séquences contient des nœuds qui sont étiquetés avec des séquences de nucléotides, les liens entre les nœuds servant d'arêtes. La chaîne de nœuds successifs dans un graphe de génome est considérée comme un chemin. En général, le graphe de séquence est bidirectionnel. Le graphe génomique est approprié pour représenter un pangénome puisque chaque chemin représente un individu dans la population étudiée.

Pour étudier la variation structurale d'un pangénome, plusieurs méthodes ont été développées (GraphTyper, BayesTyper, ou vg toolkit) principalement sur les problèmes de génotypage. En conséquence, ces outils fonctionnent en fonction d'un graphe construit à partir de variants connus puis du génotypage basé sur le réaligement des lectures cartographiées, la distribution des k-mer, la couverture des lectures et le graphe d'alignement du génome entier. Cependant, il existe encore certaines limitations dans la présentation des variants structurales imbriqués ou l'identification des orthologues.

Dans ce manuscrit, je discute différents cas pour travailler avec les variations structurales et la façon dont nous pouvons bénéficier du graphe de génome pour améliorer l'inférence des variations structurales.

J'ai d'abord étudié un cas complexe de variations structurales, les éléments transposables, pour souligner les limites de l'utilisation d'un seul génome linéaire de référence. J'ai donc proposé une solution pour travailler avec le graphe de génome tout en profitant des outils des génomes linéaires de référence. J'ai développé des outils à cette fin, PARROT et BioGraph.jl. J'ai également développé GraphInfer pour l'inférence de structures basée sur le graphe pour les données à faible couverture. Ces outils et ce pipeline de travail ont été appliqués au riz asiatique comme exemple.

Mots clés: variation structurale, pangénome, graphe de génome

Abstract

To compare multiple genomes, a linear reference genome is generally used as a coordination system to describe genes, variations and other functional annotations across individuals. However, this single reference was shown to be limited to grasp every existing genomic variation such as copy number variations (CNV), presence/absence variations (PAV) or more general structural variations (SV). To overcome this limitation, the concept pangenome made of a core-genome and a dispensable genome was applied to investigate a group of genomes.

Graph-based data model generated by incrementally incorporating genome-to-graph alignment information was one of the novel approaches to represent pangenome information. A sequence graph contains nodes that are labelled with nucleotides sequences and the linkages among nodes serve as edges. The chain of successive nodes in a genome graph is a path. To represent an individual, generally, the sequence graph is bidirected. Genome graph is suitable for representing a pangenome since each path can demonstrate an individual in the studied population.

For studying structural variation in a pangenome, several methods were developed (GraphTyper, BayesTyper, and vg toolkit). These tools mostly focus on genotyping problems. Correspondingly, these tools function depending on a graph built from known variants then genotyping based on mapped read realignment, k-mer distribution, read coverage or whole-genome alignment graph. However, there are still some limitations in presenting nested structural variants or identification of orthologs.

In this manuscript, I would like to discuss different cases to work with structural variations and how we can benefit from genome graph to improve structural variation inference.

I firstly studied a complex case of structural variation, transposable elements, to emphasize the limits of using a single linear reference genome. Hence, I proposed a solution to work with genome graph while still taking advantages tools for linear reference genomes. I developed tools and package for that purposed namely PARROT and BioGraph.jl. I also developed GraphInfer for structure inference based on graph for low coverage data. All the tools and work pipeline was applied on Asian rice as example.

Keywords: structural variation, pangenome, genome graph, transposable elements

Résumé français des travaux de thèse

La conservation de la biodiversité à tous les niveaux (écosystèmes, espèces et gènes) est importante pour notre santé, notre richesse, notre environnement, notre alimentation et tous les autres services dont nous dépendons. Cependant, ce n'est que récemment qu'il est devenu possible d'étudier la diversité intraspécifique, composante génétique majeure de la biodiversité. Spécifiquement, les technologies de séquençage en lectures longues ont montré que les membres d'une même espèce présentent de nombreuses variations génomiques structurales : délétions, insertions, duplications, variations du nombre de copies, inversions et translocation. Mes travaux de recherche se concentrent sur le développement de méthodes pour détecter ces variations structurales par différents mécanismes et approches. Au cours de mon doctorat, j'ai travaillé sur un cas spécifique d'éléments transposables et sur une approche globale pour la détection et l'inférence de variations structurales *via* les graphes de pangénome.

Les éléments transposables (ETs) sont les principaux composants des génomes, et se multiplient dans les génomes hôtes au fil des générations, modifiant ainsi de nombreuses fonctions de régulation, de transcription et de protéines. Ils sont des sources majeures de variations structurales potentiellement létales (insertion *de novo*, délétion, recombinaison ET-ET), mais la détection de leur transposition reste un défi en raison de leur nature répétitive et des limites de la technologie de séquençage en lectures courtes. Avec mes collègues, nous proposons une méthode combinant l'assemblage de génome et de mapping de lectures longues pour surmonter ces limitations, et fournir une détection du mouvement des ETs à l'échelle globale et à l'échelle individuelle.

Les lectures provenant d'individus de même génération ont été assemblées et organisées face au génome de référence pour obtenir un nouveau génome. Les variants globaux ont été détectés à partir de ces assemblages en utilisant TrEMOLO [1], qui récupère les positions d'insertion et de délétion. L'insertion d'une copie unique au sein d'une population a été faite en re-mappant les lectures ayant servi à l'assemblage contre le génome assemblé et en appelant les variants, dans les deux cas alignés sur la base de données ETs du laboratoire de Bergman. En appliquant le pipeline sur *Drosophila melanogaster* et *D. simulans*, nous avons pu identifier des séquences ETs mobiles dans les deux espèces de manière précise. Notre travail est actuellement optimisé pour les génomes de *Drosophila* ; des efforts supplémentaires sont nécessaires pour adapter cette approche à des échelles plus grandes.

Les progrès de séquençage de nouvelle génération, en particulier Illumina, permettent d'avoir des dizaines, voire des centaines de génomes individuels d'une même espèce à coût modique. Par conséquent, le contenu génomique de l'espèce entière, son pangénome, est accessible. Un pangénome comprend deux compartiments, le génome coeur, qui comprend les séquences présentes chez tous les individus, et le génome dispensable ou variable [2]. Il a été démontré que ce dernier a un impact énorme sur l'adaptation et constitue un réservoir de gènes d'intérêt [3]. Le pangénome est souvent représenté sous

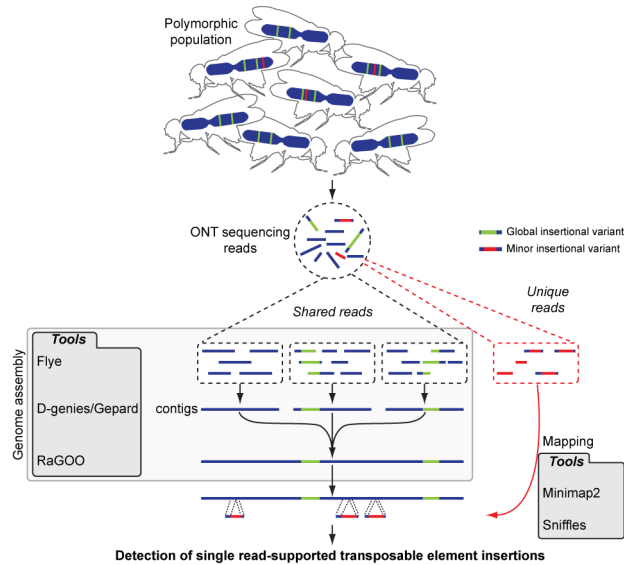


Figure 1: Approche de détection des mouvements d'éléments transposables avec TrEMOLO [1]

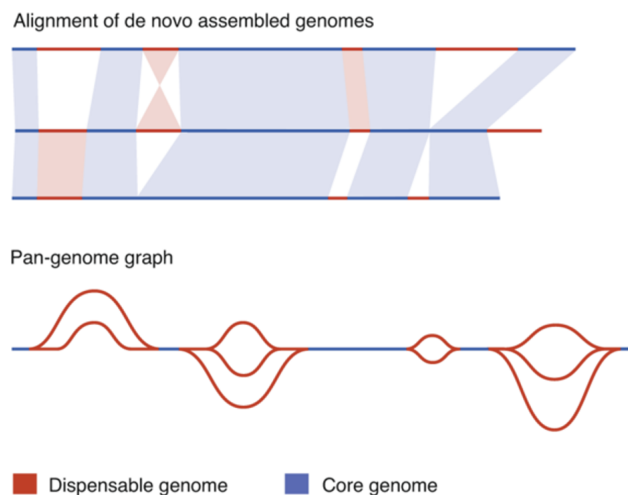


Figure 2: Concept de graphe du pangéome

forme de graphe (Figure 2) dans lequel une séquence génomique individuelle est un chemin unique traversant l'ensemble du graphe, toute variation étant représentée par une boucle [4]. Cependant, l'approche actuelle de graphe de pangéome repose sur l'utilisation de génomes entiers ou au moins de contigs de haute qualité. Je développe un pipeline composé de 3 outils qui infèrent des variants structuraux sur des données basse couverture à partir d'un graphe de pangéome issu de références de haute qualité (Figure 3).

Le premier outil, PARROT (<https://github.com/nguyetdang/PARROT>), rassemble des informations à partir du graphe et de ses données de mapping individuelles pour profiler le chemin linéaire des individus utilisés dans la construction du graphe. Ces informations sont utilisées par BioGraph.jl (<https://github.com/nguyetdang/BioGraph.jl>) pour extraire le plus long chemin représentatif linéaire comme génome synthétique. Celui-ci permet de bénéficier des outils conventionnels tout en apportant plus d'informations qu'une séquence unique. Le 3e outil, GraphInfer (<https://github.com/nguyetdang/GraphInfer>) prédit les variations structurales sur les individus basse couverture.

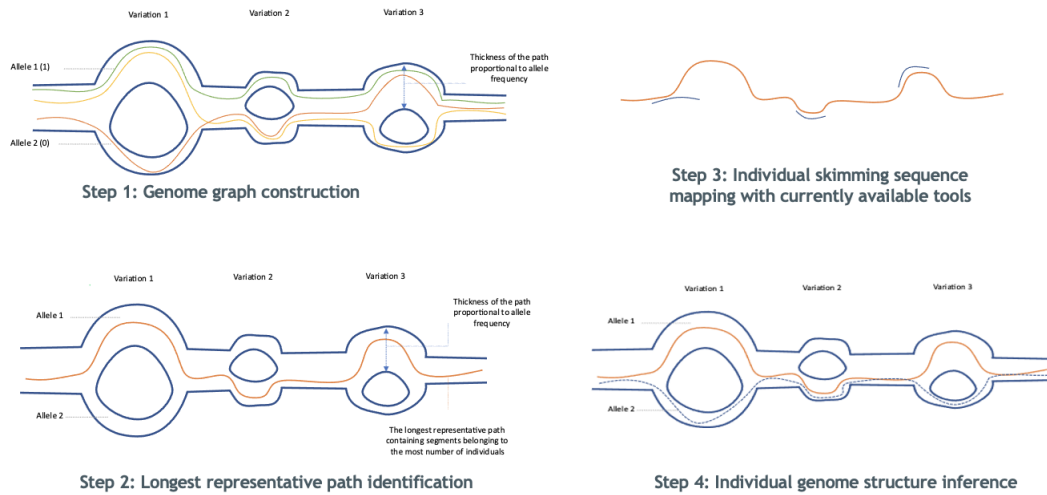


Figure 3: Approche basée sur le graphique de l'inférence de la variation structurale des données d'écrémage

En appliquant cette approche sur le riz asiatique (*Oryza sativa*), j'ai généré un génome synthétique construit à partir des séquences de haute qualité de 13 individus, d'une taille de 463 MB, environ 100 MB plus long que la référence actuelle. Je vais continuer à travailler sur l'application de mes outils pour prédire la variation structurale d'une collection de riz vietnamien ayant des données de séquençage à faible couverture.

Contents

Acknowledgement	ii
Résumé	iii
Abstract	iv
Résumé français des travaux de thèse	v
1 General Introduction	1
1.1 Introduction to structural variation	1
1.1.1 Classification of structural variants	1
1.1.2 Structural variation identification approaches	2
1.1.3 Experiment-based structural variant calling	2
1.1.4 Computational-based structural variant calling	3
1.2 Pangenome	4
1.2.1 Pangenome concept	4
1.2.2 Pangenome construction methods	5
1.3 Genome Graph	8
1.3.1 Types of genome graph	8
1.3.2 Graphical Fragment Assembly formats	9
1.3.3 Graph construction pipelines	10
1.3.4 Limitations	11
1.4 Structural variation based on genome graph	11
1.5 Study model: Vietnamese rice collection	12
1.6 Objectives of my PhD	13
2 Methods to detect transposable elements contents and dynamics	15
2.1 Context	15
2.2 Strategies	16
2.3 Conclusion	16
2.4 Perspectives	16
2.5 Personal implication	17
2.6 Scientific impacts on my PhD work	17
3 Linearisation of the genome graph	41
3.1 Context	41
3.2 Strategies	42
3.3 Conclusion	44
3.4 Perspectives	44
3.5 Personal implication	44
3.6 Scientific impacts on my PhD work	45

4	Structural variation inference of skimming data	55
4.1	Context	55
4.2	Strategies	55
4.3	Conclusion	56
4.4	Perspectives	56
4.5	Personal implication	56
4.6	Scientific impacts on my PhD work	56
5	Discussion and Perspectives	60
5.1	Discussion	60
5.2	Perspectives	62
5.3	Brief on my PhD	63
	Bibliography	64
A	Other publications	73
A.1	Published paper	73
A.2	Paper in submission	92
A.3	Paper in preparation	92
B	Conferences	93
B.1	Oral presentation	93
B.2	Poster presentation	104
C	Trainings	110

List of Figures

1.1	Simple structural variations [16] Deletion/insertion indicates the gain/loss of DNA sequences. Mobile element shows the movement of DNA fragments. Tandem duplication means the gain of a new copy of a DNA fragment at the same position, while in the case of interspersed duplication, the copy is in another position. Inversion is when the DNA fragment change its direction and translocation is when the fragment move from one to another location.	2
1.2	Examples of complex genomic rearrangement events [14] Complex genomic rearrangement is when various genomic events can happened at once. It can be rearrangement at chromosomal scale such as chromothripsis, chromoanasythesis and chromoplexy.	3
1.3	Structural variants identification from sequencing data [14] In the case of short-read and long-read sequencing data, variants can be identified by aligning reads against the reference genome. In case of copy number variants, the information can be extracted from investigating read coverage.	4
1.4	A pangenome and its compartments [37]. A pangenome is defined as the container of all genomic content available in a population. There are two main compartments: The core genome includes all genomic content shared among all individuals and the remaining part is the dispensable genome.	5
1.5	Assemble-then-annotate pangenome construction. Reads from individual genomes are assembled. Then, through annotation step, list of genes in each individual is identified. Comparing the gene list allows the construction of the pangenome.	6
1.6	Assemble-then-map pangenome construction. Reads from individual genomes are assembled to yield the assembly of each individual. The assembled genome are aligned against each other to identify the common and uncommen region to construct the pangenome.	6
1.7	Metagenomic-like pangenome construction. Reads of all individuals are gathered and assembled to construct an assembly of the studied population. Individual reads are then mapped against the obtained assembly to identify the shared compartment and the dispensable one.	7
1.8	Map-then-assemble pangenome construction. In case of having a reference genome, individual reads are aligned against the reference to generate the uncomplete version of the pangenome. Unmapped reads are assembled to yield individuals contigs. These contigs are then aligned against each other to complement the information in the core and the dispensable compartments.	7

1.9	Representation of of variants in linear and graphical representation [43]. R indicates the reference genome while A, B, C, D represent individual genomes. (i) While aligning against the linear reference genome, variants in mapped regions can be identified and documented while variants in unmapped regions cannot be represented in the linear reference. (ii) In graphical model, variants in individual genomes can be depicted through different nodes and paths, hence, capturing all the sequence information	8
1.10	Various types of genome graph [43]. (i) De Bruijn graph, (ii) Directed genome graph, (iii) Bidirected genome graph	9
1.11	Genome graph construction with pggp and minigraph pipelines (<i>adapted figure from the github of each tool</i>). In pggp pipeline, sequences from different individuals are aligned against a reference genome. Variants are called and variation graph is built by augmenting the reference with variant information. In minigraph pipeline, taking an individual sequence as the skeleton, sequences from other individuals are incrementally aligned against the skeleton/graph. The common regions are collapsed on the graph while the uncommon regions induces novel edges on the graph.	10
3.1	Input (GAF) and output (PAV) of PARROT. The GAF file is obtained from minigraph by mapping individual genomes against the graph and the presence/absence matrix file can be obtained after processing with PARROT.	42
3.2	Possible types of graph documented in rGFA format. Each dot represents the node in the graph, equivalent to DNA fragment. Each edge depict the connection between the node. The DNA fragment within each dot can be in forward and reverse directions, hence, yield a bidirected genome graph.	43

Glossary

bp	base pair.
CNVs	copy number variants.
DNA	deoxyribonucleic acid.
FISH	Fluorescence <i>in situ</i> hybridization.
GFA	Graphical Fragment Assembly.
kbp	kilo base pair.
LINEs	Long Interspersed Nuclear Elements.
LTR	Long-Terminal Repeat.
Mbp	mega base pair.
NGS	next generation sequencing.
PAV	Presence/absence variation.
piRNA	PIWI-interacting RNA.
SNPs	single nucleotide polymorphisms.
SV	structural variation.
TEI	transposable element insertion.
TEs	transposable elements.
WGS	Whole genome sequencing.

Chapter 1

General Introduction

1.1 Introduction to structural variation

In the bloom of genomic research and data, it has been shown that many phenotypic polymorphisms are caused by genomic variations such as single nucleotide polymorphisms (SNPs) [5, 6] and structural variation (SV) [7]. Although SNPs, which are the substitution of a nucleotide by another in a DNA sequence, was previously supposed to be the major form of genomic variation [8], recent studies about SV have gradually revealed the importance of SV in different phenomena such as adaptive evolution and species diversification [9]. Initially, structural variation were defined as the alteration of DNA segments longer than 1 kb in the genome and microscopically detectable [10]. The advancement in genomic technologies such as the next generation sequencing (NGS) allowed to obtain not only the detection and the effects of smaller SV [11] but also the extent of intra- and interspecific SV [12, 13]. In general, by now, genomic alteration event concerning rearrangement of sequences longer than 50 bp are considered as structural variation.

1.1.1 Classification of structural variants

By comparing the genomic content between the reference and its alternative genome after rearrangement events, SVs can be considered as balanced if its genomic content remains unchanged, as for inversion and translocation, or unbalanced if there is a change in its genomic content (insertion, deletion, copy number variation).

Based on the mechanism of action, a structural variant can be classified as a simple event or a complex one [14]. Simple events are defined as genomic rearrangements containing a few number of breakpoints and resulting junctions (see Figure 1.1) such as deletions (2 breakpoints, 1 junction), insertion (1 breakpoint, 2 junctions). SVs caused by mobile elements (transposons/transposable elements) are simple events as they lead to insertion events, and occasionally followed by further genomic alteration at the target sites [15]. Another type of simple event is duplication. However, the number of breakpoints and junctions in this case depends on the number of inserted positions, for example, interspersed duplication, where two copies are not in the same position, requires more breakpoints and junctions formed than tandem duplication, where two copies are next to each other. The same scenario is applied for copy number variations where the number of copies determines the number of breakpoints and junctions. In addition to copy number variants CNVs, when a segment of a chromosome change to its opposite direction or move to another position in a distal region within the same chromosome or in another chromosome, inversion and translocation respectively, a simple event occurs [14].

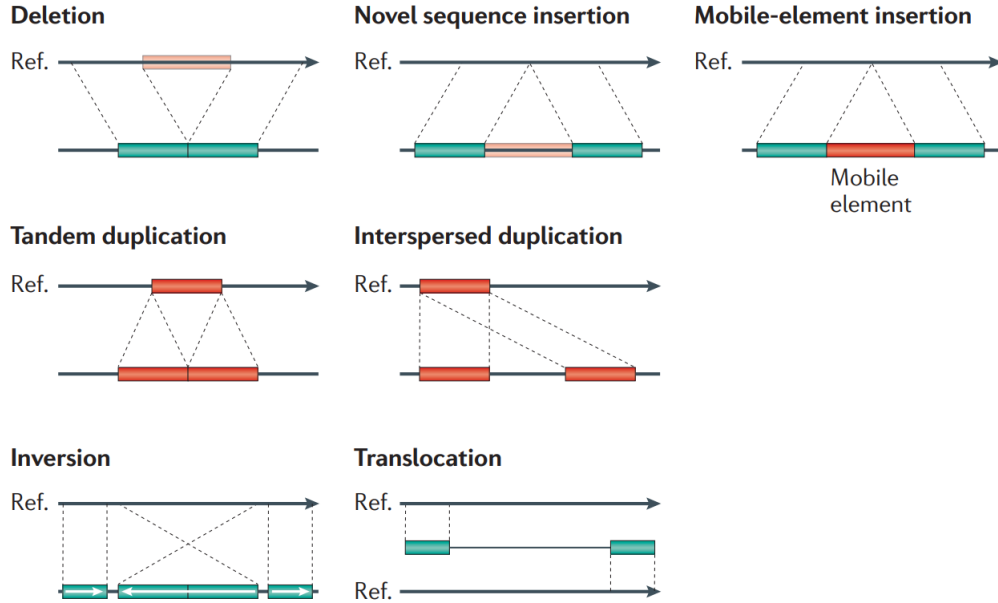


Figure 1.1: **Simple structural variations** [16] Deletion/insertion indicates the gain/loss of DNA sequences. Mobile element shows the movement of DNA fragments. Tandem duplication means the gain of a new copy of a DNA fragment at the same position, while in the case of interspersed duplication, the copy is in another position. Inversion is when the DNA fragment change its direction and translocation is when the fragment move from one to another location.

Although a majority of structural variants are simple events, complex ones involve multiple genomic rearrangements, resulting in more genomic breakpoints and junctions [17]. Some complex patterns are comprised of simple events such as duplication-normal-duplication/duplication [18] as in Figure 1.2. Another example is the massive chromosomal rearrangement occurring in chromothripsis, chromoplexy, and chromoanagenesis [19, 20]. Even these chromoanagenesis event causes catastrophic alteration in the chromosome, their appearance is very rare [20].

1.1.2 Structural variation identification approaches

The advancement in technology leads to the evolution in SVs identification strategies. Starting from the use of light microscopy to detect large rearrangements in metaphase, at the moment, it is the sequencing era and structural variant calling through computational approaches. In this section, SVs identification approaches is separated into experimental-based and computational-based methods with the latter increasing in popularity and becoming the main methodologies in recent years.

1.1.3 Experiment-based structural variant calling

In previous century, chromosomal banding techniques, in which chromosomes were stained and unique banding patterns were formed based on the DNA content and the type of dyes, were used to compare genomic profile among individuals [21]. Another study of Speicher, Ballard, and Ward [22] using chromosomal banding techniques on karyotyping of human chromosomes depicted translocations and deletions on abnormal samples. Due to its low-resolution, this technique is better suitable for determining large chromosomal rearrangement.

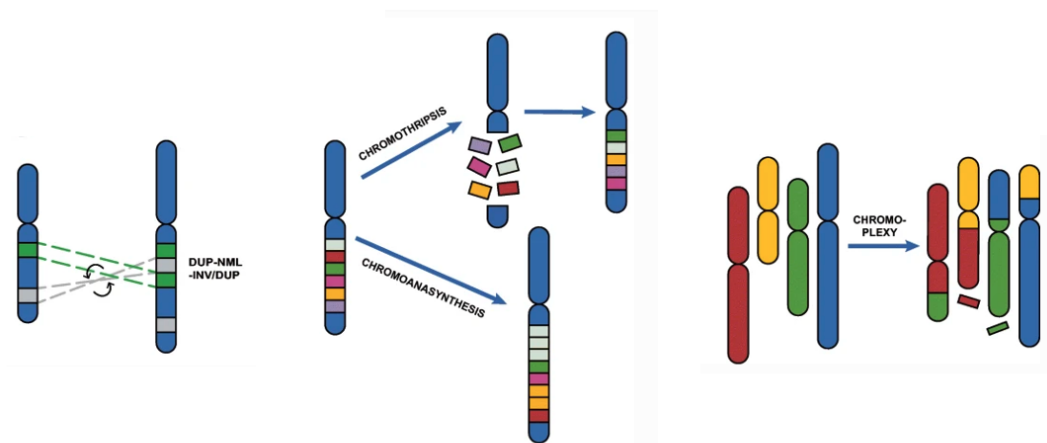


Figure 1.2: **Examples of complex genomic rearrangement events** [14] Complex genomic rearrangement is when various genomic events can happen at once. It can be rearrangement at chromosomal scale such as chromothripsis, chromoanasythesis and chromoplexy.

The later trend in experimental-based approaches for SVs determination is hybridization and mapping techniques, which include two common techniques: fluorescence *in situ* hybridization (FISH) and high-throughput optical mapping [14].

In FISH, chromosomal DNA is complemented with nucleotides hybridizing fluorescent probes. The fluorescent signals are then imaged and quantified by microscope to identify potential mis-localization signal [23]. FISH technique allows the identification of a wide range of SV types such as translocations [23], sub-telomeric rearrangements [24], chromothripsis events or even highly repetitive regions [25]. The resolution of FISH detection can be up to 100 kbp with high accuracy and low false positive rates [26].

High-throughput optical mapping creates optical maps by using fluorescent labeling of nicking restriction enzyme sites along a DNA fragment having length from 300 kbp to 3 Mbp. The information of restriction sites are then extracted and undergo *de novo* assembly to compare with the reference genome to identify regions containing structural variations [27, 28]. Optical mapping can also determine a wide range of rearrangements, however, it does not provide information at nucleotide levels of the whole genome. It can be consider as a hybrid method between experimental-based and computational-based approaches.

1.1.4 Computational-based structural variant calling

While experimental-based approaches mostly take into account fluorescent signals to access structural variation, computational-based approaches work with sequencing data which allow to identify SV breakpoints at single nucleotide level. The advancements in sequencing technology lead to cost-effective and large-scale discovery of SVs using whole-genome sequencing (WGS). Genomic sequencing methodologies are commonly separated into short- and long-reads technologies, which later affect computational methods and algorithms to identify structural variants. They can either be the alignment of sequencing reads to a reference, or a *de novo* assembly of a genome followed by the comparison to a reference to infer SVs. Generally, sequencing techniques and computational-based variant calling methods allow the detection of SVs at nucleotide resolution on the whole genome.

In more details, short-reads techniques identify SVs by finding evidence of genomic rearrangement from read pairs (SV present between paired-end reads), split reads (SV

present on the sequenced portion of the reads), read depth (number of reads mapped to a specific regions of the reference genomes) or assembly (contigs produced from reads) [29, 30]. However, it is challenging for short-read techniques to identify structural variants in long repetitive regions since during assembly process, reads shared similarity are normally collapsed to form contigs. In addition, such sort reads SV detection may produce a high number of false positives and negatives [1].

This obstacles can be overcome by long-read sequencing technologies since they provide reads that are able to span the entire size of majority genomic SVs. The two most commonly used long-read sequencing techniques are single-molecule real-time sequencing by Pacific Biosciences (PacBio) and Nanopore sequencing by Oxford Nanopore Technologies (ONT) by which read length can reach to more than 10,000 bp [31]. The main issues of these techniques are their high native error rates per base (ONT reads at 5-25% [28] and 0.5% for Q20+, and PacBio reads at 13% [32], 1% for HiFi). At the moment, computational methods and tools for long-read data focus on reducing technological error rates. For structure variants discovery, SV calling tools use evidence from split reads, soft-clipped reads and *de novo* assembly. Approaches to identify SVs from sequencing data is depicted in Figure 1.3.

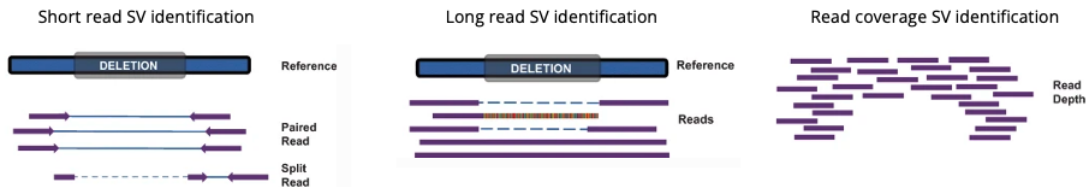


Figure 1.3: **Structural variants identification from sequencing data** [14] In the case of short-read and long-read sequencing data, variants can be identified by aligning reads against the reference genome. In case of copy number variants, the information can be extracted from investigating read coverage.

1.2 Pangenome

One of the most common approaches for sequencing analysis in diversity study is to compare genome sequencing reads (both short and long reads) against a linear haploid reference genome originating from a single individual or a consensus of several individuals. This linear reference genome can only represent a single allele out of the entire genetic variation of the studied population. While aligning against the reference genome, sequencing reads containing alternative alleles might be mismatched, thus, obtain lower mapping score than the same allele of the reference. This reference bias can cause missing alternative alleles, mistaking heterozygous site as homozygous site [33], or influencing allele frequencies [34]. Another possibility for diversity study is to compare pairwise almost complete genomes through genome assembly of reads. However, when more than one alternate genome is involved it is much more complex to see and visualise or even store the common and variable regions. In this context, the pangenome concept was proposed to overcome the mentioned obstacles.

1.2.1 Pangenome concept

The concept of pangenomics was first introduced for bacteria in more than 15 years ago [2], in which the researchers proposed that a pangenome included a core genome (shared genes among all individuals), a dispensable genome (composed of common genes

in some but not all individuals), and individual-specific genomes. Since then many terms were developed and used in various pangenome studies, for example, supragenome, distributed and unique genes [35], and flexible regions [36]. Simply define, originating from the initial explanation, a pangenome consists of two compartments: a core genome containing common sequences among (almost) all individuals, and a dispensable genome having the remaining genomic content found in the studied group (see Figure 1.4). Indeed, due to technical and sampling limitations, it is uncertain to confirm if a gene or a sequence is specific to an individual, hence, the individual-specific genomes is not taken into account [37]. In summary, a pangenome is the union between a core and a dispensable genome, hence, can be used to document not only the common genomic features of a studied group but also its diversity.

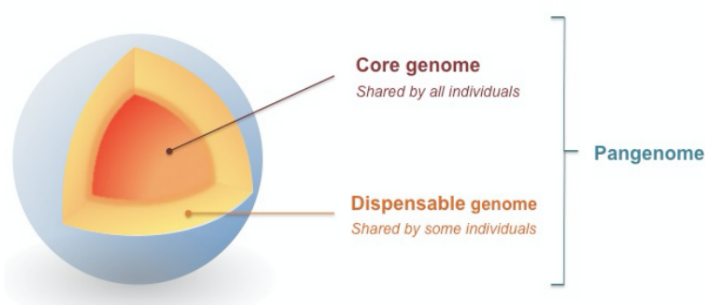


Figure 1.4: **A pangenome and its compartments [37]**. A pangenome is defined as the container of all genomic content available in a population. There are two main compartments: The core genome includes all genomic content shared among all individuals and the remaining part is the dispensable genome.

1.2.2 Pangenome construction methods

At the beginning, the common framework of constructing a pangenome was initiated with the *de novo* assembly of all reads in each individual. After obtaining assemblies, annotation was performed to identify genes encoded by each individual, and homologous gene groups recovered. Based on the present/absent of genes between individuals, the core and the dispensable genome were defined and used in further analyses (see Figure 1.5). This approach was and is still applied widely in many studies related to microorganisms (Rasko et al., 2008). However, the pangenome structure is here only accessible at the genic level.

To overcome the limitation mentioned above, an optimization of this method based on *de novo* assemblies has been proposed [38]. After obtaining the scaffold/chromosome assemblies, multiple alignment/mapping of genomic regions in different individuals were performed to identify the common regions of all individuals in the group. Then, these common regions became the core genome, and the remaining was defined as the dispensable genome (Figure 1.6). This optimization allowed the access to the genotype of a population at DNA level.

Nonetheless, approaches related to *de novo* assemblies requires high quality sequencing data for each individual, especially in the case of higher eukaryotes with polyploidy. In 2015, a metagenomic-like approach was used to construct a pangenome analysis on a combination low-coverage data of approximately 1,500 rice genomes [39]. Whole genome sequences of all individuals were assembled at once and the contigs were later re-assigned to each individual through mapping of single individual data on the metagenomic assemblies (see Figure 1.7). Even this approach works with low-coverage data, the appearance of chimerical contigs might affect the accuracy of analysis.

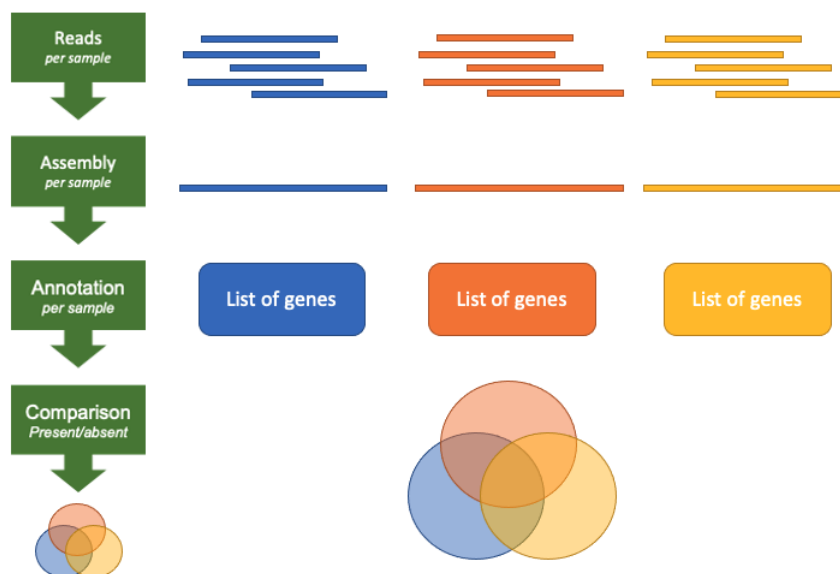


Figure 1.5: **Assemble-then-annotate pangenome construction.** Reads from individual genomes are assembled. Then, through annotation step, list of genes in each individual is identified. Comparing the gene list allows the construction of the pangenome.

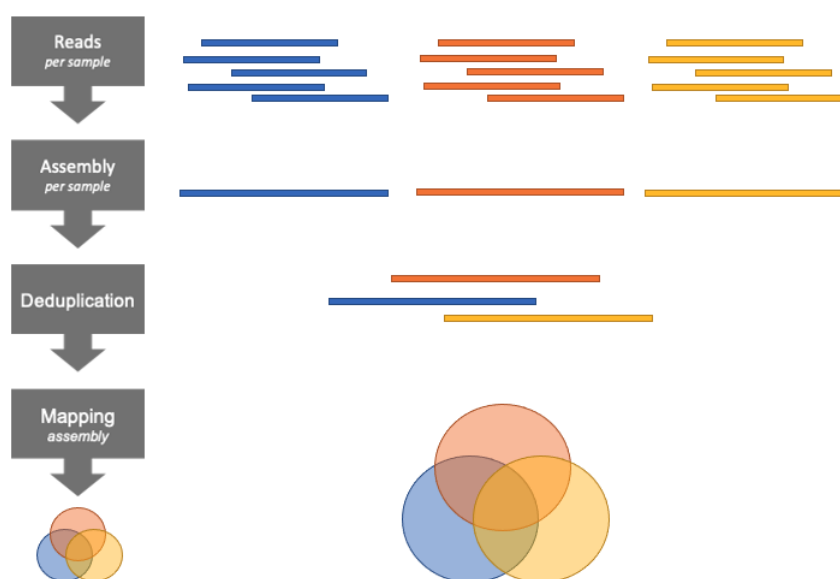


Figure 1.6: **Assemble-then-map pangenome construction.** Reads from individual genomes are assembled to yield the assembly of each individual. The assembled genome are aligned against each other to identify the common and uncommen region to construct the pangenome.

In addition, *de novo* assembly approaches demanded computation resources to deal with big amount of data. Hence, an alternative method starting with the mapping process of sequencing reads to identify the common genomic regions then the assembly step of unmapped data was suggested (map-then-assemble; [40]). The unmapped data can be assembled per individual or through iterative mode [41, 42] after each alignment on the generated panreference (see Figure 1.8). Even this approach is less time-consuming, it is generally used with short reads and might lead to short contigs that may impair further analysis.

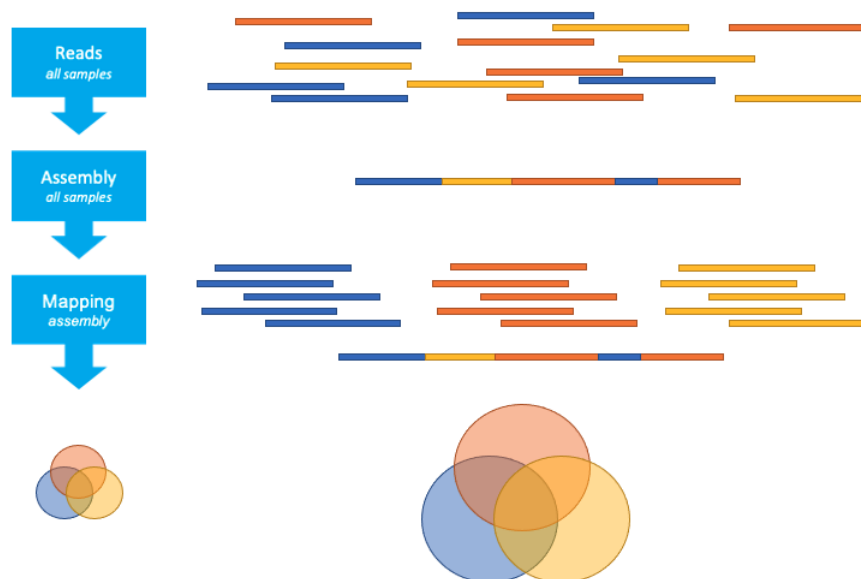


Figure 1.7: **Metagenomic-like pangenome construction.** Reads of all individuals are gathered and assembled to construct an assembly of the studied population. Individual reads are then mapped against the obtained assembly to identify the shared compartment and the dispensable one.

Current standard methods still require either high quality sequencing data or enormous computational resources or both. To adapt the use of pangenome concept into studying large genomic variations for low coverage sequencing data, new approach should be considered.

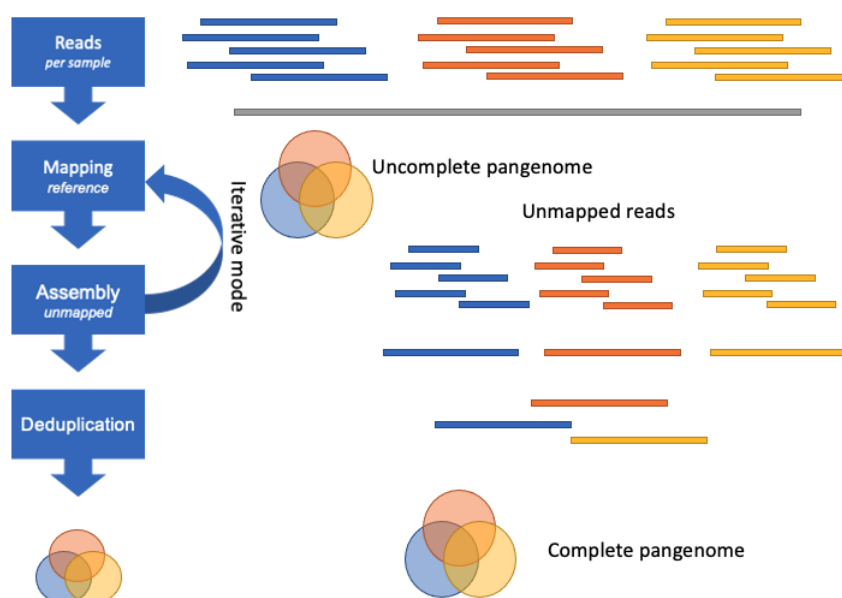


Figure 1.8: **Map-then-assemble pangenome construction.** In case of having a reference genome, individual reads are aligned against the reference to generate the uncomplete version of the pangenome. Unmapped reads are assembled to yield individuals contigs. These contigs are then aligned against each other to complement the information in the core and the dispensable compartments.

1.3 Genome Graph

While a linear reference can act as a consolidate coordinate system to address genomic elements, graph can integrate multiple paths and can explain complex relationships among sequences (see Figure 1.9).

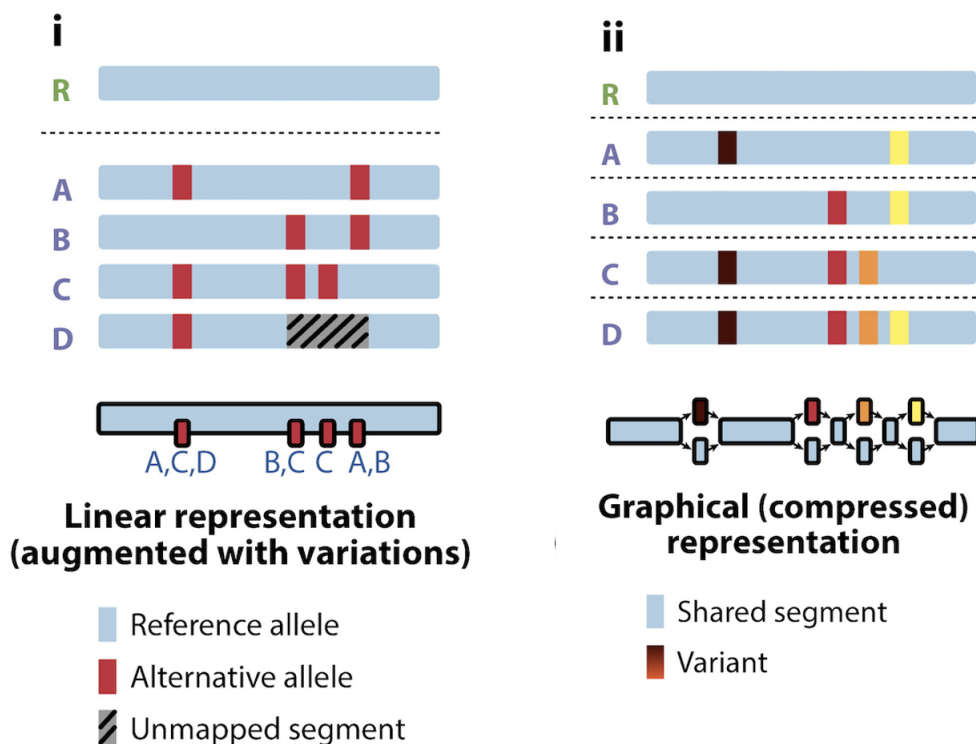


Figure 1.9: **Representation of of variants in linear and graphical representation** [43]. R indicates the reference genome while A, B, C, D represent individual genomes. (i) While aligning against the linear reference genome, variants in mapped regions can be identified and documented while variants in unmapped regions cannot be represented in the linear reference. (ii) In graphical model, variants in individual genomes can be depicted through different nodes and paths, hence, capturing all the sequence information

In this section, I would like to discuss different types of genome graph and how to build one. In addition, some limitation in the current usage of genome graph are also discussed.

1.3.1 Types of genome graph

Graphs have been used as a representation of genomic sequences since the development of assembly concept. In de Bruijn graphs (see Figure 1.10-(i)), each node represents a k -mer (sequence of k nucleotides) and each edge depicts how these k -mer are connected. The overlap between two consecutive k -mers is defined as 1 [44].

In the case of genome graph, the size of variants can be varied. Hence, each node can depicts a DNA sequence and edges in the graph shows how these DNA sequences are connected as in Figure 1.10-(ii). Multiple sequence alignment can be indicated in these sequence graphs. In the case of representing the direction of DNA strands or inversion, bidirected genome graph [45] can be applied as in Figure 1.10-(iii). Node connected to more than one outward edge are considered as intersection points between connected subsequences [46].

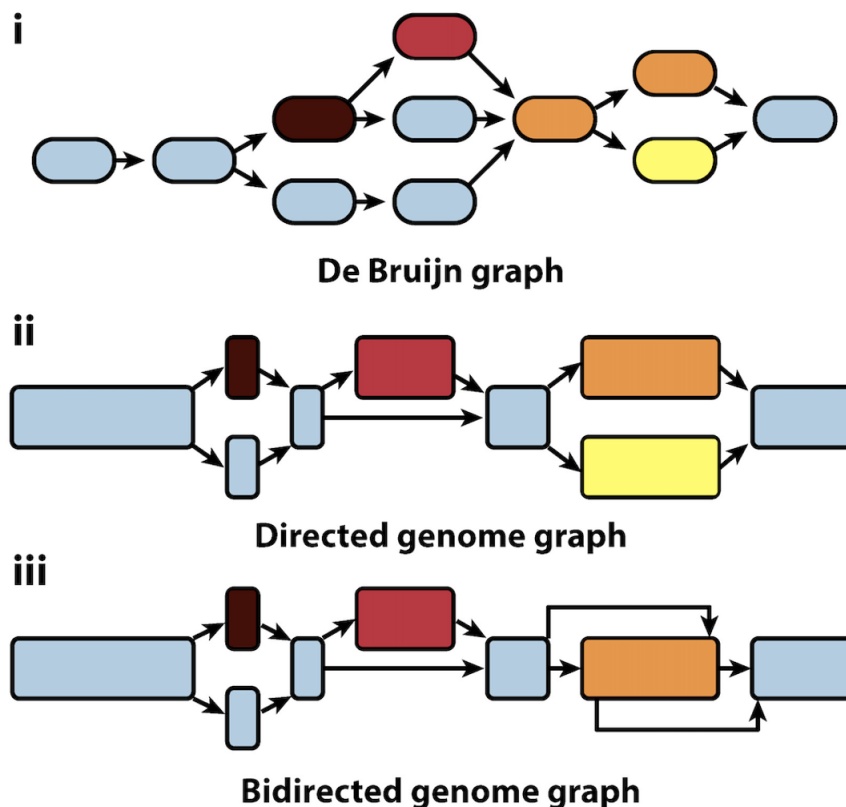


Figure 1.10: **Various types of genome graph** [43]. (i) De Bruijn graph, (ii) Directed genome graph, (iii) Bidirected genome graph

1.3.2 Graphical Fragment Assembly formats

Genome graph can be represented in Graphical Fragment Assembly (GFA) format (<https://github.com/GFA-spec/GFA-spec>). The purpose of the format is to help user access assembly, variation and splicing of the graphs. There are two main versions of GFA formats, which are GFA1 and GFA2.

The GFA1 was initially suggested by Heng Li and it contains the followings components: segment (DNA sequence), link (overlap/connection between two segments), jump (connection between two oriented segments, does not imply direct adjacency between segments), containment (overlap between two segments where one is contained in the other), path (ordered list of segments), walk (ordered list of oriented segments, representing haplotype information) (<https://gfa-spec.github.io/GFA-spec/GFA1.html>). There is also rGFA format which is a subset of GFA1 in which only segments (S-lines) and links (L-lines) are noted.

GFA2 was proposed by vg team to store a sequence graph at any stage of assembly such as graph of overlaps or final assembly of contigs and multi-alignment (<http://gfa-spec.github.io/GFA-spec/GFA2.html>). GFA2 documents more complexity in the graph, for example, fragments which can be described as DNA sequence belonging to a segment, sequences in external collection, interval of segment that external string contributed to or interval of fragment that contributes to the segment. Another element is gaps to estimate the distance between two segment sequences. Moreover, there are groups decoded as U-line and O-line for naming subgraph within the graph.

1.3.3 Graph construction pipelines

There are 3 common pipelines for genome graph construction: vg [47] and pggp methods, minigraph [48] and minigraph cactus. An example to build the human pangenome by using all the 3 pipelines can be refer in this recent preprint in July 2022 [49].

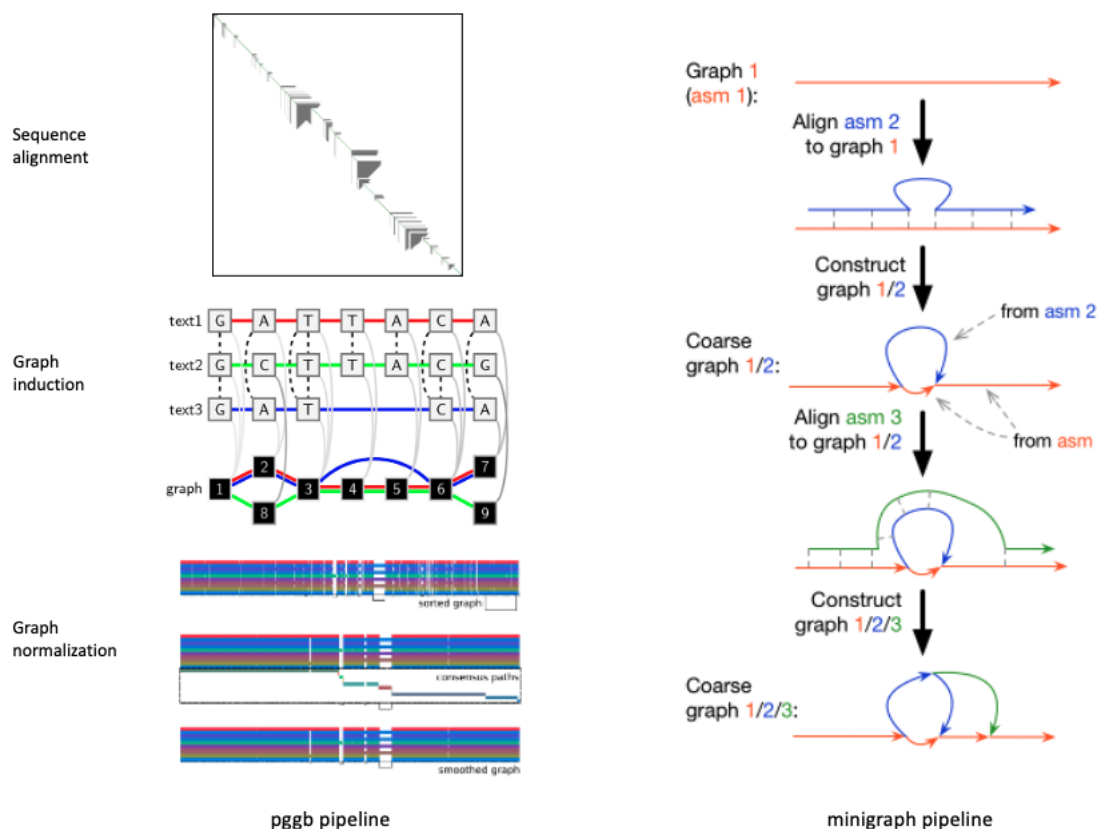


Figure 1.11: **Genome graph construction with pggp and minigraph pipelines** (adapted figure from the github of each tool). In pggp pipeline, sequences from different individuals are aligned against a reference genome. Variants are called and variation graph is built by augmenting the reference with variant information. In minigraph pipeline, taking an individual sequence as the skeleton, sequences from other individuals are incrementally aligned against the skeleton/graph. The common regions are collapsed on the graph while the uncommon regions induces novel edges on the graph.

In more details, vg and pggp methods build genome graph from a collection of sequence based on variation graph model. Since the released of vg [47], variation graph model is composed of nodes (labeled by sequences and id), edges (connect two nodes) and paths (describing genomes, sequence alignments and annotation walk through the nodes and edges). Starting with a reference, vg constructs the genome graph by incorporating information from VCF files of individuals in a studied population. Meanwhile, pggp can start with index FASTA sequences to construct a locally directed and acyclic graph while maintain large structure variation. The pipeline depends on sequence alignment, hence, allow the access of base-level relationship among segments. However, the constructed graph might contain high number of nodes and affect run time. Then, it is necessary to normalize the graph to smooth small loops.

The principle of minigraph is aligning query sequence against a sequence graph and progressively augments an existing graph with long query sequence distinct from the graph [48]. The input of minigraph is FASTA files. The construction is highly order

dependent. Starting from one sequence, large variations of the next sequence are added to form the graph, then, each subsequence of each successive sequence is combine into the graph. Shared subsequences are collapsed on the graph while the remaining parts form new bubbles. minigraph only refer to variations larger than 100 bp. The constructed graph from minigraph is stored in rGFA format. The ideal of pggb and minigraph pipelines are illustrated in Figure 1.11.

Cactus [50] take a phylogenetic tree as a guide to create multiple alignments. minigraph then uses the order from previous result to construct the graph. minigraph Cactus pipeline has two mode: Pangenome mode and Progressive mode. The difference between these two mode is that in Pangenome mode, the first genome used to construct the graph does not collapse against itself. Therefore, the coordinate system of the graph is in accordance with the first genome. As a result, output from this pipeline can be project to VCF to work with vg toolkit and also compatible with rGFA format.

1.3.4 Limitations

At the moment, the standard to represent and to construct a genome graph is still in discussion. GFA1 and GFA2 exist in parallel since each serves its own purpose. To be specific, GFA1 is familiar for minigraph and GFA2 is used in pggb/vg ecosystem. The interchange between the two is still limited since there are only a few tools work with both format or to convert from one to another such as gfakluge [51] and gfapy [52]. Due to the difference between the two formats and their principle of genome graph construction, developers who want to develop tools for genome graph, they have to choose either to work with one type genome graph or another. Moreover, for pangenome analysis of a given species, all 3 genome construction methods coexist as in the Human Pangenome Reference preprint [49].

Complex structure within the graph is problematic, especially when internal cycles involve due to lack of tools for alignment. Most of the current alignments works only with directed acyclic graphs (PaSGAL [53]) or unroll cycles (vg toolkit [47]). HybridSPAdes [54] proposed it was possible to align reads against cyclic graph, however, the run time will be polynomial. GraphAligner [55] is a fast alignment algorithm built on a minimizer-based seeding method to compute edit distance (quantifying how many operators required to compute from one string to another). However, edit distance among sequence strings affects variant calling since long INDELS might not be discovered. In term of dealing with cycler, GraphAligner finds super bubbles, then creates a pseudo linear positions for nodes before the alignment to treat cycle cases. A^* seed heuristic algorithm [56] were proposed to reduce the influence of edit distance principles, however, cycles were not discussed in this case. To date, a lot of efforts are still required to optimize graph alignment issues.

Another tradeoff that needs to be considered while working with genome graph is the resolution and computational time. For example, variation graph from vg and pggb can provide graph having base-level variants while in minigraph, only structural variation are documented. Hence, it is recommended to smooth the graph from vg and pggb to optimize run time.

1.4 Structural variation based on genome graph

The variant calling process can be divided into two steps: identification of potential variation sites and determination of the genotypes at those sites. Based on the structure of a genome graph, a node containing more than one outward edge is considered as the site of variation. Based on that idea, the location of structural variation can be determined. For example, minigraph [48] extracts the coordinate and sequences information of bubbles

in the genome graph as list of structural variants.

Other tools for genotyping includes GraphTyper [57, 58], BayesTyper [59], Paragraph [60] and vg toolkit [47]. GraphTyper realign short reads sequences against a variation graph representing a pangenome and possible haplotype as graph path. Variants were determined on a sliding window scales. BayesTyper also works with variation graphs constructed from a set of input variants and a reference genome. They use k -mers in the sequences to build the profile for each individuals by traversing through the graph and genotyping was perform by combining k -mer counts from the individual’s haplotype and counts from noise process. Paragraph evaluate short-reads map at breakpoints of the graph to identify structure variants in the graph. vg decomposes the bubble sites and investigate reads mapped against the reference and the variant paths in variation graph to identify which path was taken by the individuals.

Most of genotyping tools work with variation graph built by a reference genome and a VCF file containing known variants. However, the position of SVs is vastly dependant on alignment and SVs caller, which is an in-discussion issue in genome graph analysis. Therefore, using only VCFs for structural variant integration reduce the ability to work with complex structure such as nested SVs [48] whose coordinates is not only in the reference but in other individuals. Another issue can come from graph construction. For example, in vg, a segment in the graph can appear in multiple paths, allowing different regions to collapse into one segment, even the reference. The graph then become a collapsed graph where the discover of orthologs is challenging since multiple sequences from the same sample might go through the same segment.

Furthermore, in the validation, most of these tools are working with samples having long-read or high-coverage short-reads sequences. There is no work on the genotyping for individuals possessing only skimming data. This tendency raises the need to benefit genome graph for structural variants determination and genotyping of low-coverage sequenced individuals.

1.5 Study model: Vietnamese rice collection

Rice (*Oryza sativa*) is one of the most essential crops in the world, especially for Asian and African countries. In Vietnam, for instance, rice is not only the major food staple, but also the main agricultural export and resources, accounting for 85% of the total agricultural land. Because of the topographical properties in Vietnam, rice is cultivated in various ecosystems: irrigated, rainfed lowland, flood-prone, upland and mangrove. From these cultivation conditions, the diversity of rice genotypes and phenotypes have been selected and accumulated through generations to cope with these distinct environments. Hence, it is considered that Vietnam is one of the main centers of genetic diversity for rice germplasms, particularly traditional rice varieties [61, 62]. Until now, more than 3,000 rice varieties have been collected through Vietnam. Despite the potential richness of genetic resources, only a small proportion (15%) of the diversity has been used in modern breeding program [62].

In a previous study [63], a panel of 182 Vietnamese rice accessions were used to study the genetic diversity of Vietnamese rice landraces. By applying genotyping-by-sequencing (GBS) on this panel, a dataset of 21,814 single nucleotide polymorphisms (SNP) was identified and then used to represent the genetic diversity of Vietnamese rice landraces. Due to the low number of detected variants, the resolution of the SNP-array in this study is not sufficient.

Therefore, a whole-genome sequencing data of 168 Vietnamese rice landraces that belong to the collection mentioned above and a Vietnam’s popular commercial variety (Khang Dan 18) along with four references (Nipponbare, Azucena, IR64 and IR29) was

prepared by Illumina HiSeq7 2000/2500. To access the genomic content at a reasonable price, each individual was sequenced at low coverage ($\sim 2-3x$). Among them, 6 samples were sequenced at high coverage ($\sim 50x$). Those sequences were used to generate a SNP dataset of higher resolution on the rice genome by inference approaches. Because of the low coverage sequencing reads, it is difficult to detect large genomic variations such as structural variations or presence/absence variations (PAV).

The objective of my research is to find a method to predict large genomic variations such as structure variations and presence/absence variations on low-coverage sequencing data and then apply it on the whole genome sequencing of Vietnamese rice landraces.

1.6 Objectives of my PhD

In the first year, my PhD topic was "Development of AI methods for pangenomics on populational sequences data-driven" in which I worked on developing AI methods to predict structural variations for low-coverage sequencing data based on pangenome.

My approach was to first build a high-quality pangenome for rice using a *map-then-assemble* approach from high-depth sequencing individuals obtained in 3K rice genome project. Then, I would create a presence/absence profile of individuals with low coverage sequencing data and compare these profile against the build pangenome from the high-coverage sequenced individuals. I intended at this time to implement an artificial intelligence models to predict presence/absence value of low-coverage individual afterwards by taking advantage of neural network and single matrix decomposition methods.

However, during the data preparation process, I found out that the presence/absence matrix obtained from my skimming data have too many missing value. So that, the prediction result might not be significantly meaningful. Therefore, I decided to drop the idea of performing prediction based on presence/absence matrix.

in the same time, I collaborated on a project dedicated to transposable elements detection using long read sequencing data. During this study, I was confronted to the limits of genome assembly sequences in terms of representation of the different haplotypes presents in the sequencing dataset.

Thus, I redirect my research topic to genome graph, which is an approach to represent a (pan)genome. Going through different bibliography resources, I found that this approach to be more suitable for structural variation inference of skimming data, as it provides many benefits such as richer structures comparing to the reference genome or haplotype embedded information within the graph. However, there are still many open questions related to alignment, variant calling, variant annotation or genotyping issues [64]. To fit with my purpose of performing structural variant inference of skimming data, my strategy contains the following steps:

- Step 1: Constructing a genome graph of reference genome
- Step 2: Extracting the most representative path of the graph and allowing it to act as a reference genome. In this case, we will be able to take advantage of both the tools use for linear reference genome and the powerful advantages of genome graph.
- Step 3: Perform variant inference of individuals having skimming data by mapping sequencing reads against the extracted representative path. Unmapped reads will be retrieved and aligned back to the graph to identify which structure they are following.

For that purpose, I used the rice as data model, as this species has already more than 12 high-quality genome sequences, with a quite important level of structural variations and a huge set of low-coverage sequencing data available.

After reformulate my PhD research project, I come up with 3 objectives that will be explained more clearly in the following chapters:

- Explore structure events in a common case, transposable elements, with the limit of the sequence-based approaches
- Implement a linearisation of the pangenome graph structure for a better all-day usage by biologists
- SV inference in low-coverage data using the linear pangenome

All of these approaches lead to tool development (TrEMOLO, PARROT/BioGraph and GraphInfer, respectively), and provide a framework to deal with low-coverage or population sequencing data in order to explore their structural variations.

Chapter 2

Methods to detect transposable elements contents and dynamics

In this chapter, I studied a specific case of structural variation identification which is transposable elements. This is the work I finished in the 1st year of my PhD. This work allowed me to understand different scenario of structural variant identification and further more, to work with data coming from various sequencing technology (long-read and short-read both included in the study). I improved not only my programming skills through the study and also the ability to work in a collaborative development environment. In addition, the result from the study illustrates the limit of sequence-to-sequence comparison and the missing haplotypes. This raise the need of having a better system to depict the genomic content of a population, for example, a genome graph.

Article published in Cells. DOI: 10.3390/CELLS9081776 Received: 27 June 2020; Accepted: 23 July 2020; Published: 25 July 2020.

2.1 Context

Transposable elements (TEs) are mobile DNA fragments having the ability to propagate independently through genomes. The length of TEs can vary from 100 bp to 10,000 bp or even larger sometimes [65]. In term of classification, TEs are generally classified in two classes depending on the transposition intermediate, and are subsequently grouped into DNA transposons, Long-Terminal Repeat (LTR) elements and Long Interspersed Nuclear Elements (LINEs) [66, 67]. One of their main impacts is gene expression variation, which makes it an interesting topic in many research since its first discovery in previous century [68].

In eukaryotic genomes, transposable element insertion (TEI) is a common structural variation type [69, 67]. Different experimental methods were applied to identified TEI sites and to understand the invasion of TEs in the genome. However, due to the technical limitations linked to the repetitive nature of TEs, their mechanisms of action still remain in questions. In the sequencing era, analysing Illumina short-read data from a pools of individuals allowed the determination of TEI frequency in natural samples [70]. Indeed, the main issues of using short -read is under-representation of TEIs and noise caused by background errors [71].

Drosophila melanogaster acts as a model organism in biology research besides *Homo sapiens*, *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Anopheles gambiae* and *Fugu rubripes*. It possesses many beneficial characteristics for genetic and genomic research such as a small genome size, relatively easy husbandry and a range of very sophisticated genetic tools [72]. In our study, we first overcome the technical issue by developing strategies to generate *de novo* assemblies of high quality long-read sequencing data and

to analyze the presence of TEs at high and low frequencies in *Drosophila* population. We then applied our strategies on investigation of PIWI-interacting RNA (piRNA), which is a repetitive region on *Drosophila* genome that cannot be determined by short-read data. piRNA represses TE mobility and maintains genome integrity [73].

2.2 Strategies

Stable strains of wide-type *D. melanogaster* and *D. simulans* were used to describe the general landscapes of TEIs. To compare TEs present at high and low frequencies, isogenic wild-type strain of *D. melanogaster* and an unstable line with a succession of transpositions busts having Piwi knockdown were generated for the study.

TE content was estimated using two following methods on forward reads: TEcount module of TEtools and dnaPipeTE. TEI site estimation was detected by mapping Illumina short-read against the FlyBase reference genome and TE sequence library. Pairs of reads having one end mapped to the reference genome and the other end to a TE sequence were retrieved. For each pair, the position of the reference genome mappable read were documented and used for the estimation of the insertion site.

Long-read nanopore data was submitted to Flye v2.6 for assemblies, then polished using four round of RACON v1.3.2 and corrected by RaGOO v1.1. with *Drosophila* reference genome. RepeatMasker and Dfam database was used to identified TE content at this step. Global structural variant was identified by in-house svTEidentification.py tool (available at <https://github.com/DrosophilaGenomeEvolution/TrEMOLO>). For minor insertion variant detection, raw long reads were mapped against its corresponding assembly and SV calling was performed using Sniffles v1.0.10. Sequences longer than 1000 bp were aligned against LTR subset of TE database <https://github.com/bergmanlab/transposons>. Specific criteria for section of the global structural variant and minor insertion variant was indicated in the paper. The method was validated by comparing the genome size, TE content and TEI site estimation obtained by short and long-read sequencing in *D. melanogaster* and *D. simulans*.

2.3 Conclusion

Comparing the landscapes of the most recent TEIs in wild-type strains of *D. melanogaster* and *D. simulans* suggests that the similarity between TE sequences of the two species are more than previous thought. DNA transposons display higher intra-family sequence divergence than LTR elements, suggesting that elements of this group invaded the genome more recently than other types of transposable elements. We found that piRNA production associated with TE genome occupancy, hence, the piRNA clusters trap model was not supported here.

Our work emphasized the importance of long-read data in describing TE landscapes at intra-genome scale. By combining both short-read and long-read data, we are able to determine TE content, TEI site, global structural variant and LTR minor insertion variant. Long-read nanopore data are able to study repetitive region like piRNA, which is overcome the limitation of short-read data. In addition, long-read data can also detect singleton reads.

2.4 Perspectives

In the study, due to the strain specificity, the identity of the most recently active TE families is still concerting. The issues might come from the limit of sequence-to-sequence

comparison and the missing haplotypes that cannot represent the whole diversity of the studied groups. For later phase of the research, we might consider using genome graph analysis to overcome the issues.

In term of development, features to identify populational variations (and even somatic variations) is developed, in which, TEs that are outsiders or represent in only a part of the population will be identified.

2.5 Personal implication

Many authors contributed to this publication, for the details, the contribution of each author is indicated in the paper. I retrieved the data and perform data curation and analysis including: mapping and variant calling for both long and short reads. I perform global structural variation detection and LTR minor insertion variant detection. I worked with Mourdas Mohamed for the retrieving of paired-end reads in which only one of them was mapped against the reference and another mapped with the TE database to identify TE content and TEI sites. My scripts were used for further development of the work pipelines and TrEMOLO.

2.6 Scientific impacts on my PhD work

This work highlight the limitation of using only a single linear reference genome for genomic analysis. The limits is significantly emphasized in complex case such as transposable elements. Therefore, the needs for another representation of genomic content and analysis tools is in highly demand.

Article

A Transposon Story: From TE Content to TE Dynamic Invasion of *Drosophila* Genomes Using the Single-Molecule Sequencing Technology from Oxford Nanopore

Mourdas Mohamed ^{1,†}, Nguyet Thi-Minh Dang ^{2,†}, Yuki Ogyama ¹, Nelly Burlet ³, Bruno Mugat ¹, Matthieu Boulesteix ³, Vincent Mérel ³, Philippe Veber ³, Judit Salces-Ortiz ^{3,4}, Dany Severac ⁵, Alain Pélisson ¹, Cristina Vieira ³, François Sabot ², Marie Fablet ^{3,*} and Séverine Chambeyron ^{1,*}

¹ Institute of Human Genetics, UMR9002, CNRS and Montpellier University, 34396 Montpellier, France; mourdas.mohamed@igh.cnrs.fr (M.M.); yuki.ogyama@igh.cnrs.fr (Y.O.); bruno.mugat@igh.cnrs.fr (B.M.); alain.pelisson@igh.cnrs.fr (A.P.)

² IRD/UMR DIADE, 911 avenue Agropolis BP64501, 34394 Montpellier, France; dangminhnguyet09@gmail.com (N.T.-M.D.); francois.sabot@ird.fr (F.S.)

³ Université de Lyon, Université Lyon 1, CNRS, Laboratoire de Biométrie et Biologie Evolutive UMR 5558, 69622 Villeurbanne, France; nelly.burlet@univ-lyon1.fr (N.B.); matthieu.boulesteix@univ-lyon1.fr (M.B.); vincent.merel@etu.univ-lyon1.fr (V.M.); philippe.veber@univ-lyon1.fr (P.V.); judit.salces@ibe.upf-csic.es (J.S.-O.); cristina.heddi@univ-lyon1.fr (C.V.)

⁴ Institute of Evolutionary Biology (IBE), CSIC-Universitat Pompeu Fabra, 08003 Barcelona, Spain

⁵ MGX-Montpellier GenomiX, c/o Institut de Génomique Fonctionnelle, CNRS, INSERM, Université de Montpellier, 34094 Montpellier, France; Dany.Severac@mgx.cnrs.fr

* Correspondence: marie.fablet@univ-lyon1.fr (M.F.); severine.chambeyron@igh.cnrs.fr (S.C.); Tel.: +33-47-243-2916 (M.F.); +33-43-435-9949 (S.C.)

† These authors contributed equally to this work.

Received: 27 June 2020; Accepted: 23 July 2020; Published: 25 July 2020



Abstract: Transposable elements (TEs) are the main components of genomes. However, due to their repetitive nature, they are very difficult to study using data obtained with short-read sequencing technologies. Here, we describe an efficient pipeline to accurately recover TE insertion (TEI) sites and sequences from long reads obtained by Oxford Nanopore Technology (ONT) sequencing. With this pipeline, we could precisely describe the landscapes of the most recent TEIs in wild-type strains of *Drosophila melanogaster* and *Drosophila simulans*. Their comparison suggests that this subset of TE sequences is more similar than previously thought in these two species. The chromosome assemblies obtained using this pipeline also allowed recovering piRNA cluster sequences, which was impossible using short-read sequencing. Finally, we used our pipeline to analyze ONT sequencing data from a *D. melanogaster* unstable line in which LTR transposition was derepressed for 73 successive generations. We could rely on single reads to identify new insertions with intact target site duplications. Moreover, the detailed analysis of TEIs in the wild-type strains and the unstable line did not support the trap model claiming that piRNA clusters are hotspots of TE insertions.

Keywords: transposable elements; ONT; *Drosophila melanogaster*; *Drosophila simulans*; piRNA

1. Introduction

Transposable elements (TEs) are major components of almost all eukaryotic genomes [1,2]. They can be separated into three main groups that include several TE superfamilies and families:

DNA transposons, Long-Terminal Repeat (LTR) elements, and Long Interspersed Nuclear Elements (LINEs) [2,3]. Different methods (e.g., Southern blotting [4,5], in-situ hybridization on polytene chromosomes [6,7], and PCR [8,9]) were first used to estimate TE content in *Drosophila* genomes and to understand how TEs invade and shape genomes by affecting genome function and evolution. However, technical problems linked to TE repetitive nature and diversity have not allowed for the reaching of firm conclusions and many questions about TE biology remain unanswered.

Then, next-generation short-read sequencing technologies allowed for characterizing the global TE content within and between related species. Moreover, the high coverage provided by Illumina sequencing led to the identification of consensus sequences for each TE family. Several computational methods were developed, such as RepeatExplorer [10] and dnaPipeTE [11], to analyze Illumina data from different *Drosophila* species, and to study TE biology at the populational level.

In TE biology, it is also important to estimate the TE insertion (TEI) rate to determine the degree of polymorphism within and between populations. This is an indicator of the activity level of each TE family and can help to date transposition events [12,13]. Illumina sequencing of pools of individuals allowed for determining TEI frequency in natural samples (from large numbers of individuals to populations) [14]. To study individual TEI, dedicated software tools were developed (e.g., TIDAL [15], T-Lex/T-Lex2 [16], PopoolTE2 [17]) based on the analyses of: (1) the TEI junction and flanking sequences (split-reads), (2) the paired-end information, (3) the depth of coverage, or (4) a mix of these three criteria. However, these approaches revealed only a portion of the repetitive sequence landscape, and they detect many false positives due to various factors. The first one is linked to the library preparation and PCR amplification that lead to the generation of PCR chimera and thus false positive insertions [18], or to biased sequence representativity (AT- and GC-rich sequences are less represented in Illumina sequencing). The second factor is inherent to the sequencing size (short reads) that does not span more than 400 bp, thus hindering the full sequencing of any repeat or variation larger than this size, especially insertions [19]. The third one is related to the difficulty in detecting TEIs occurring at low frequency in an individual or population. Indeed, these TEIs are usually under-represented in the sequencing data and generally confused with background errors [18]. The comparison of different methods to identify TEIs shows very small levels of overlap [20]. Another weak point of the Illumina sequencing technology is that the insertion size and sequence are not accessible, because this approach generally only gives the global position.

Long-read, or third-generation, sequencing technology might improve the detection of long structural variants and thus of TE variations, and also reduce the detection of false positives/false negatives. This technology should allow for the identification of full copies. Indeed, long-read sequencing methods generate individual reads that are mostly longer (>15 kb) than many of the repeats (TE sequences are generally smaller than 10 kb). Moreover, it solves the problems linked to PCR-based library preparation because it relies on direct DNA sequencing without amplification. However, the main drawback of long-read sequencing, such as the Oxford Nanopore Technology (ONT), is the high rate of single read sequencing errors (3 to 8% for the recent sequencing and base calling) that could introduce bias in data interpretation. This problem is partially solved by increasing the coverage and by improving the final assembly quality by polishing, thus providing an almost perfect genome sequence. Such an approach, based on PacBio sequencing, has already allowed the detection of 38% more TEIs in *Drosophila* chromosome 2 L compared with the available short-read sequencing estimates [21]. Different *Drosophila* genome assemblies using ONT sequencing have also been reported [22,23]. Long-read sequencing methods allow almost complete chromosome-scale genome assemblies, instead of the fragmented draft genomes provided by short reads. Therefore, the assembled individual genomes can be directly compared without the need for any reference genome and their relative structural variants can be scored without biases (or very few). In addition, long-read sequencing of genomes should allow identifying real TEI sites and accurately determining TE copy number at the inter- and intra-population levels. This approach might also help to analyze

repetitive regions like PIWI-interacting RNA (piRNA) clusters that contribute to maintaining genome integrity by repressing TE mobility.

Here, we developed strategies to generate *de novo* assemblies of high quality long-read sequencing data, suitable for genomic analyses of TEs present at high and low frequencies in *Drosophila* populations. We first validated our method by comparing the data (genome size, TE content and TEI site estimation) obtained by short and long-read sequencing in *D. melanogaster* and *D. simulans*, two closely related species, but that may vary in TE content [24,25]. We found that, although the *D. simulans* genome contains a large number of old and degraded TE copies, among the most recent pool of insertions, DNA transposons display higher intra-family sequence divergence than LTR elements, suggesting that elements of this group invaded the genome more recently than DNA transposons. Moreover, we observed that piRNA production correlates with TE genome occupancy. When considering the most recent pool of TE insertions, we could not find convincing evidence supporting the piRNA clusters trap model [26,27]. Finally, we developed and validated an approach to identify TEI that occur at low frequencies in a population.

2. Materials and Methods

2.1. *Drosophila* Strains

The wild-type *D. melanogaster* and *D. simulans* strains from natural populations were kept at 24 °C in standard laboratory conditions on cornmeal–sugar–yeast–agar medium. The eight samples of *D. melanogaster* and *D. simulans* natural populations were collected using fruit baits in France (Gotheron, 44°56'0"N 04°53'30"E—"goth" lines) and Brazil (São Jose do Rio Preto 20°41'04.3"S 49°21'26.1"W—"sj" lines) in June 2014. Two isofemale strains per species and geographical origin were established directly from gravid females from the field (French *D. melanogaster*: dmgoth63, dmgoth101; Brazilian *D. melanogaster*: dmsj23, dmsj7; French *D. simulans*: dsgoth613, dsgoth31; Brazilian *D. simulans*: dssj27, dssj9). Brothers and sisters were then mated for 30 generations to obtain inbred strains with a very low amount of intra-line genetic variability.

A previously published *D. melanogaster* laboratory line [28] was used for Piwi knockdown (piwi KD) in adult follicle cells. This line carries three components: (i) a GAL4 UAS activator driven by the follicle cell-specific traffic jam (tj)-promoter (tj-GAL4), (ii) an UAS short hairpin(sh)-piwi that induces Piwi RNAi, and (iii) the ubiquitously expressed thermo-sensitive GAL4-inhibitor GAL80^{ts}. At 20 °C, GAL80^{ts} sequesters GAL4, preventing sh-piwi expression. At 25 °C, GAL80^{ts} is partially inactive, allowing some GAL4-driven expression of sh-piwi in somatic follicle cells. The resulting partial Piwi depletion allows for the derepression of at least two LTR families (ZAM and gtwin) in follicle cells and their integration as new proviruses in the progeny genome [28]. The polymorphism of this line was partially reduced by isolating a single pair of parents and the line was thereafter stably maintained at 20 °C as a large population (more than 500 progenitors at each generation). The G0 and G0-F100 genomic libraries were prepared shortly after isolation of this line and at the hundredth generation, respectively. Soon after isolation of this isofemale line, a subset of individuals at the pupal to early adult stages was shifted to 25 °C for 5 days, and this was repeated for at least 500 flies for 73 successive generations of partial Piwi KD. Then, after six more generations of stabilization at 20 °C, a third genomic library, called G73, was generated.

2.2. Genome Size Estimations

Flow cytometry: genome size was estimated according to [29] using fresh samples of 4-day-old females heads with 10 replicates (five heads per replicate) for each *Drosophila* wild-type strain.

findGSE: k-mer distribution was established from the Illumina reads using findGSE [30]. Briefly, adaptors were first removed from the reads with Skewer version 0.2.2 (paired-ends) or NxTrim version 0.4.3-6eb8d5e (mate pairs), when necessary. Reads were then treated essentially as previously described [31] to remove duplicates, filter out reads mapping to reference mitochondrial genomes

(GenBank AF200854.1 and AF200828.1 [32]) or microbial contaminants. This allowed for establishing the 21-mer distributions from which genome sizes were estimated using findGSE [30] with default parameters, except for dmsj23 in which the k-mer distribution clearly displayed a peak corresponding to heterozygous regions and was thus treated accordingly.

2.3. Illumina Sequencing

Wild-type strains: DNA was extracted from 3 to 5-day-old females for each strain using the Qiagen DNeasy Blood&Tissue kit (# 69506) and following the manufacturer's instructions. Genomic DNA (1.5 µg) was fragmented for a target insert size of 300 base pairs and sequenced by paired-end Illumina HiSeq (125 bp reads). Library and sequencing were performed by the GeT-PlaGe facility, Génopole Toulouse (France).

2.4. DNA Isolation, Oxford Nanopore MinION Sequencing and Base Calling

DNA was extracted from ~100 males from each wild-type and from the Piwi KD lines using the Qiagen DNeasy Blood&Tissue kit. The genomic DNA quality and quantity were evaluated using a NanoDrop™ One UV-Vis spectrophotometer (Thermo Fisher Scientific, Waltham, MA, USA) and a Qubit® 1.0 Fluorometer (Invitrogen, Carlsbad, CA, USA), respectively. Three micrograms of DNA were repaired using the NEBNext FFPE DNA Repair Mix (NEB M6630). End repair and dA-tailing were performed using the NEBNext End repair/dA-tailing Module (E7546, NEB). Ligation was then performed with the Ligation Sequencing Kit 1D (SQK-LSK108, ONT, for G0, and SQK-LSK109 ONT for wild type strains, G73 and G0-F100 samples). MinION sequencing was performed according to the manufacturer's guidelines using R9.4.1 flow cells (FLO-MIN106, ONT) and a Nanopore MinIon Mk1b sequencer (ONT) controlled by the ONT MinKNOW software (version 18.3.1 for G0, version 19.05.0.0 for isogenic wild-type strains, and version 19.10.1 for the G73 and G0-F100 samples). Base calling was performed after sequencing using Albacore (version 2.3.3) for G0, and the GPU-enabled guppy basecaller in high accuracy mode for isogenic wild-type strains (version 3.1.5), G73 (version 3.3.3) and G0-F100 samples (version 3.4.4).

2.5. TE Content and TEI Site Estimates from Illumina Sequencing

TE abundance was estimated using forward reads and two methods: the TEcount module of TEtools [33] and dnaPipeTE (v1.0.0 and v1.3.1) [11]. TEcount estimates TE abundance by quantifying reads that map to a set of known TE sequences, here the rosetta fasta file [34]. This tool was run using default parameters and Bowtie2 (v2.2.4) [35,36]. dnaPipeTE assembles repeated sequences from a subsample of reads (<1x) and quantifies reads mapping to these sequences to estimate TE abundance. dnaPipeTE was used with the following parameters: -sample_number 2, -genome_coverage 0.25). Concerning the genome size option, 175 Mb and 147 Mb were used for *D. melanogaster* and *D. simulans* samples, respectively. The rosetta fasta file was used as library [34].

TEIs were detected in Illumina sequencing data using a dedicated mapping-based algorithm similar to that implemented in PoPoolationTE2 [17] with paired-end reads as input, FlyBase reference genomes (ftp://ftp.flybase.net/genomes/Drosophila_melanogaster/dmel_r6.16_FB2017_03/fasta/dmel-all-chromosome-r6.16.fasta.gz and ftp://ftp.flybase.net/genomes/Drosophila_simulans/dsim_r2.02_FB2017_04/gtf/dsim-all-r2.02.gtf.gz), and the TE sequence library at https://github.com/bergmanlab/transposons/raw/e2a12ff708c42dce5b15d6af290506d78021212/releases/D_mel_transposon_sequence_set_v10.1.fa. Sequencing reads are mapped to the reference genome and TE sequences using Bowtie2 (version 2.3.3) [36]. Then, the algorithm scans the resulting Binary Alignment/Map (BAM) files for pairs in which one end matches to the reference genome, the other end to a TE sequence, and the pair cannot be mapped concordantly to the genome. For each pair, the position of the genome-mappable read is noted, and positions are clustered in order to have no read further apart than 100 bp in that cluster. Each cluster is then interpreted as an insertion, the position of which is the mean of the position of the reads it contains, and the strength of which is evaluated on

the basis of the number of reads it contains. For the purpose of this study, only insertions that were supported by at least 50 reads were retained. Unlike PoPoolationTE2, the insertions detected with this procedure correspond to occurrences absent from the reference genome.

2.6. Small RNA Extraction and Sequencing

For small RNA sequencing, two replicates per strain were prepared. Small RNA was isolated from 50 pairs of ovaries using HiTrap Q HP anion exchange columns (Cytiva, Velizy-Villacoublay, France) as described in [37], and the eluate was run on a 10% TBE urea gel (Thermo Fisher Scientific). Small RNA size selection (18–50 bp) was performed on gel at the sequencing facility. Quality was checked with the Bioanalyzer small RNA kit (Agilent, Santa Clara, CA, USA). Library construction was performed using the TruSeq Small RNA Library kit (Illumina, San Diego, CA, USA) and sequenced (1 × 50 single reads) on an Illumina HiSeq 4000 at the IGBMC Microarray and Sequencing facility. Adapter sequences were removed using cutadapt [38]. Size selection was then performed using PRINSEQ lite version 0.20.4 [39]. All subsequent analyses were built upon small RNA counts after normalization according to the miRNA amounts, as described in [34].

2.7. Genome Assembly

Raw nanopore reads were QC checked using Nanoplot v1.10.2 (<https://github.com/wdecoster/NanoPlot>) for sequencing run statistics. Reads with QC < 7 were removed by the sequence provider (Montpellier Genomix, Montpellier, France) before QC. For each dataset, mean length, N50 reads, total reads and bases are listed in Table S1 and Table 1. Reads were submitted to Flye v2.6 [40] with standard options, except `-plasmids` and `-threads 16`. Raw contigs were polished using four rounds of RACON v1.3.2 [41] with standard options and 20 threads (`-t` option; the required mapping was performed using minimap2 [42] v2.16 and `-x map-ont -t 20` options). At each step, basic assembly metrics (N50, length, L50) were recorded using Assembly-Stats v1.0.1 (<https://github.com/sanger-pathogens/assembly-stats>). Once polished, assemblies were visually inspected using D-genies v1.2.0 [43], and incongruencies manually corrected using samtools v1.9.0 [44], `faidx` command for sequence extraction, and Gepard [45] v1.4.0 for visual determination of breaking points. The corrected assemblies underwent super scaffolding using RaGOO v1.1 [46] with `-s` (structural variants (SV)) and `-t 4`, using the specific reference genome (from FlyBase): *Dmel_R6.23* for G0 and *D. melanogaster* samples, *Dsim_r2.02* for *D. simulans*, and the previously assembled G0 for G73 and G0_F100 samples. Once the assembly was finalized at the chromosome scale, a Benchmarking Universal Single-Copy Orthologs (BUSCO) analysis [47] using the gVolante web service [48] was performed using the BUSCO v2/v3 option and the *Arthropoda* reference set (Figure 1). TE content was estimated in the corresponding chromosome assemblies using RepeatMasker (<http://www.repeatmasker.org>) and the Dfam database [49].

Table 1. Statistics for the de novo assemblies before scaffolding. All lengths are expressed in bases. The Benchmarking Universal Single-Copy Orthologs (BUSCO) score indicates the “complete hit” level.

Name	Size	Nb contig	Mean Length	Longest	N50	L50	BUSCO Score, %
dmgoth101	130,483,042	1213	107,571	20,963,225	14,899,963	4	c: 98.6
dmgoth63	134,481,426	1005	133,812	22,615,553	16,996,519	4	c:98.03
dmsj23	131,331,777	1094	120,047	22,945,221	10,553,205	5	c:98.5
dmsj7	131,360,683	1197	109,742	18,094,419	6,212,683	7	c:98.7
dsgoth31	135,039,133	822	164,281	27,577,085	17,530,992	4	c: 98.3
dsgoth613	132,908,190	918	144,78	22,559,698	16,120,890	4	c:98.6
dssj27	134,309,820	866	155,092	27,370,717	20,976,825	3	c:98.6
dssj9	142,009,588	508	279,546	27,589,620	19,611,840	4	c:99
G0	127,415,251	642	198,466	5,037,957	1,208,862	33	c:93.7
G0-F100	139,374,117	836	166,715	17,781,420	9,085,947	6	c:98.97
G73	144,335,962	584	247,15	24,539,270	12,530,957	4	c:98.7

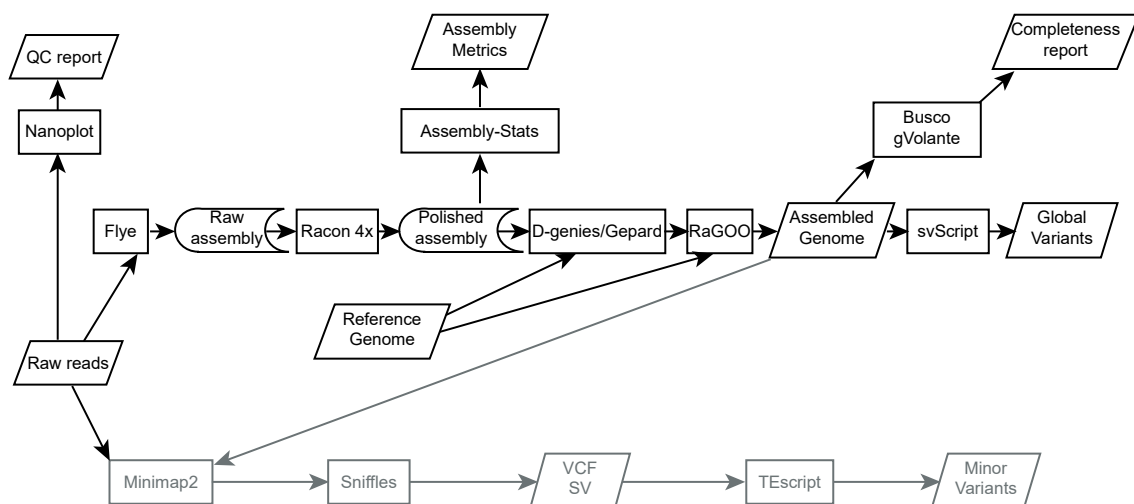


Figure 1. Schematic of the method used for genome assembly and for transposable element insertion (TEI) detection. Global variants (black) were detected from genome assemblies, and minor variants (gray) by remapping reads in these assemblies. The reference genomes used for RaGOO scaffolding were Dmel_R6.23 for G0 and for wild-type *D. melanogaster* strains, Dsim_R2.02 for wild-type *D. simulans* strains, and the G0 assembly for G73 and G0-F100.

2.8. Global Structural Variant Detection

Global variant detection (i.e., variants common to most genomes of a considered sample compared with the reference genome, see below) was performed using the svTEidentification.py tool (available at <https://github.com/DrosophilaGenomeEvolution/TrEMOLO>). Briefly, this tool recovers the insertion and deletion positions and creates the associated fasta sequence, based on the Assemblytics report from the RaGOO scaffolding (the deletions are extracted from the reference and the insertions from the new assembly). Once the fasta file corresponding to the SVs was recovered, these sequences were matched with the Basic Local Alignment Search Tool, nucleotide to nucleotide (BLASTN)+ v2.4.0 to a specified TE database. Hits larger than 80% of the TE sequence and identical to more than 80% at the nucleotide level were considered as candidate for new TE insertions/deletions (TEI/TED) in the G0, G0-F100 and G73 samples. For wild-type strains, new insertions/deletions were detected without any filter. The potential candidates were then listed in a tabular format that included their position, size and percentage of size or similarity compared with the reference TEs. The used TE database was a collection of the reference TEs from Bergman's laboratory (<https://github.com/bergmanlab/transposons>) and from previously published data [50].

2.9. LTR Minor Insertional Variant (LTR MIV) Detection

Each raw long read was mapped using minimap2 v2.16 (-ax map-ont -t 16 as options) to the assembly corresponding to that set of long reads. After recovering the sam file, samtools v1.10.0 was used to compress and sort the sam file in BAM with samtools view and samtools sort (basic options, but with 16 threads), and the MD tag was added using samtools calmd. Then, SV were detected in the resulting sorted BAM file using Sniffles v1.0.10 with at least 1 read and -report_seq -s 1 -n -1 as parameters [51]. These sequences longer than 1000 bp were aligned with BLASTN v2.4.0+ (-outfmt 6) to the LTR subset (60 families) of the database used before. A nucleotide alignment of more than 94% identity and a minimum of 90% of the total length of the TE consensus sequence were then considered as criteria to validate a putative LTR minor insertion variant (LTR MIV), if the length of the variant did not exceed the total size of the TE by more than 18 nt. This corresponds to the largest target site duplication (TSD) ever reported to flank any LTR TE [52]. All codes are available in a snakemake file at <https://github.com/DrosophilaGenomeEvolution/TrEMOLO>.

2.10. Fluorescent In Situ Hybridization on Polytene Chromosomes

Polytene chromosomes were squashed from salivary glands of third instar male larvae. *NotI* and *PstI* restriction enzymes were used to extract a fragment of the ZAM *pol* gene from a previously published plasmid [53]. The probe was labeled with digoxigenin-11-dUTP using the Nick Translation Mix (Roche #11 745 816 910), and signals were detected with anti-digoxigenin-rhodamine Fab fragments (Roche). The fluorescent in situ hybridization method was adapted from a previously described protocol [54].

2.11. Automatic Identification of the Target Site Duplication for LTR MIV

The putative LTR MIVs matching to six LTR families (blood, gtwin, mdg3, ZAM, roo, and copia) were studied. One read supporting each MIV, previously extracted in a fasta file, was compared by BLASTN v2.4.0+ with the corresponding consensus sequence. To automatically check for the presence of a TSD, the positions of the 5' and 3' end of the TE alignment were determined within the read. 30nt-long sequences upstream and downstream the putative insertion site were extracted and were aligned to detect the presence, on both sides of the insertion, of a short duplication, the size of which was previously reported by [55] for ZAM and by [52] for the other TEs. The resulting TSD sequences were then extracted and used to create sequence logos with WebLogo (<https://github.com/WebLogo/weblogo>). All scripts and codes for this automatic extraction are available at the project GitHub.

2.12. piRNA Cluster Identification in the Assembled Genomes

To determine the piRNA cluster localization in genome assemblies, a previous annotation of piRNA clusters in the *D. melanogaster* Dmel_R6.04 genome release was used [56]. The flanking genes for each of the 153 major piRNA clusters were identified, their sequence was extracted and mapped to the new reference using BLASTN to locate the limits of the corresponding piRNA clusters in the corresponding assemblies. When only a single gene could be used as border, the piRNA cluster length described in [56] was used to define the other border. Bona fide piRNAs were extracted from the previously published G0 small RNA-seq library [28], and from each of the small RNA-seq libraries presented here, as reads longer than 23 nt that do not map (bowtie -best) to sequences of other known small RNAs (downloaded from FlyBase [57] and MirBase [58]). These selected small RNA reads were then mapped to the corresponding assemblies using Bowtie 1.2.2 [59]. Bowtie parameters were selected to keep only reads that display unique alignments and <2 mismatches (-best -v 2 -m 1). The positions of uniquely mapped reads were determined in the assembly, and sequences with more than 500 reads were conserved and compared to the piRNA cluster coordinates determined in the assembly of that line. Table S4 shows the list of the 42 piRNA clusters corresponding to the best piRNA producers in the G0 line. The coordinates of these 42 regions were then determined in the G73 and G0-F100 assemblies. For wild-type strains, the piRNA abundance was computed within 1 kb windows.

2.13. Comparison of ZAM Sequences

After obtaining the corresponding region of the ZAM insertions the fasta sequence was extracted (using bedtools getfasta) and compared with the ZAM sequence at a global level using redotable v1.1.

3. Results and Discussion

3.1. Using Oxford Nanopore Technologies (ONT) to De Novo Assemble the Highly Contiguous Genomes of Several Isogenic Wild-Type Strains and of one Unstable Line

The ONT-based single-molecule long-read sequencing data provided between 5 and 24 million reads, with a depth of coverage ranging from 40x to 196x (mean = 130x), and a N50 ranging from 3.7 to almost 20 kb (mean = 11 kb) (QC 7 reads only; Table S1). The N50 large range was explained by the different methods used for genomic DNA extraction and ligation (Materials and Methods). Our assembled genome procedure is summarized in Figure 1. To compare our data with the reference

D. melanogaster and *D. simulans* genomes, whole genome alignments and local dot plots were performed using D-genies and Gepard, respectively (Figure S1).

A strong correspondence was observed between most de novo assemblies and the corresponding reference genome, except for the G73 and dsgoth31 assemblies in which incongruent contigs were detected. These incongruent contigs were manually broken at the discrepancy points (Figure S1) and the final statistics for the de novo assemblies were obtained using Assembly-Stats (Table 1).

Using our approach based only on ONT data, the N50 ranged from 1.2 Mb (L50 of 33 contigs) to 21 Mb (L50 of 3 contigs). The previously described de novo *D. melanogaster* hybrid assembly obtained using BioNano and assembly merging [23] reported a N50 of 9 Mb (L50 of 6 contigs) for the raw data, and a N50 of 21.3 Mb (L50 of 3 contigs) after merging. Moreover, the BUSCO score of their hybrid assembly was 97.2% after Illumina polishing, while the BUSCO score of our assemblies ranged from 93.7% to almost 99% (98.5% for the reference Dmel_R6.23 assembly [23]) only with RACON polishing. This comparison indicates that our assemblies are of high quality, and that RaGOO use as scaffolder allowed obtaining high-quality assemblies at the chromosome scale.

3.2. Estimation of Genome size Using Different Methods

To determine the quality of the ONT-based assemblies of the isogenic wild-type *D. melanogaster* and *D. simulans* genomes, their sizes were compared to the genome sizes estimated with two other approaches: findGSE (based on k-mer estimation) and flow cytometry (Table S2).

Genome size estimates varied between 142 and 144 Mb (flow cytometry) and 129 and 132 Mb (findGSE) for the *D. simulans* strains and between 162 and 163 Mb (flow cytometry) and 133 and 137 Mb (findGSE) for the *D. melanogaster* strains after excluding dmsj7. The k-mer distribution obtained for this strain was much more scattered than the others, and resulted in a k-mer-based genome size estimate of 147 Mb, most probably an artefact. The size estimated obtained using the ONT data ranged between 131 and 142 Mb for the wild-type *D. simulans* strains and between 130 and 134 Mb for the *D. melanogaster* strains, with similar values for the final assemblies. The correlation coefficients were significant only between the ONT-based and the flow cytometry estimates for *D. melanogaster* ($r = 0.9675$, $p = 0.0325$), but not *D. simulans* (flow cytometry: $r = 0.7564$, $p = 0.2436$; findGSE: $r = 0.1237$, $p = 0.8763$). The correlation only with the flow cytometry estimate indicates that the different genome compositions, and probably the different amounts of heterochromatin affect the estimations obtained by findGSE. The genome size estimates obtained with findGSE were globally more similar than those obtained using the de novo assembly approach, but no correlation was observed between these values, probably due to the different amounts of repeats present in the various strains. In conclusion, genome size estimations present several biases in function of the used method, and ONT assemblies seem to give values close to those obtained by flow cytometry, which is a more global method.

3.3. Comparison of TE Abundance in the Isogenic Wild-Type Strains Measured by Illumina and ONT Sequencing

To validate the ONT approach, the TE abundance in the isogenic wild-type *D. melanogaster* and *D. simulans* strains was evaluated using dnaPipeTE [11] and TEcount [33] for Illumina sequencing data, and RepeatMasker for ONT assembled chromosomes (Figure 2). Overall, TE content (expressed as genome percentage) was often higher when estimated using dnaPipeTE (Illumina data) (Wilcoxon matched-pairs signed rank test; $p = 0.0234$) than with RepeatMasker (ONT assemblies) (Wilcoxon matched-pairs signed rank test; $p = 0.0156$). This might be explained by the fact that unlike the RepeatMasker TE database, dnaPipeTE is based on the de novo detection of TEs and the local assembly of TE families, independently of a previously annotated reference genome, thus recovering the maximum number of reads that correspond to known and unknown TEs. In agreement, the correlation was higher between the results obtained with RepeatMasker (ONT data) and the results obtained with TEcount, which is based on the read similarity against a curated database of known TEs [34] ($r = 0.8921$, $p < 0.0001$), than with dnaPipeTE ($r = 0.8504$, $p < 0.0001$) (Figure 2, right panel). As previously reported,

the LTR group was more abundant than the LINE and DNA transposon groups in all *Drosophila* genomes (see [60] for a review).

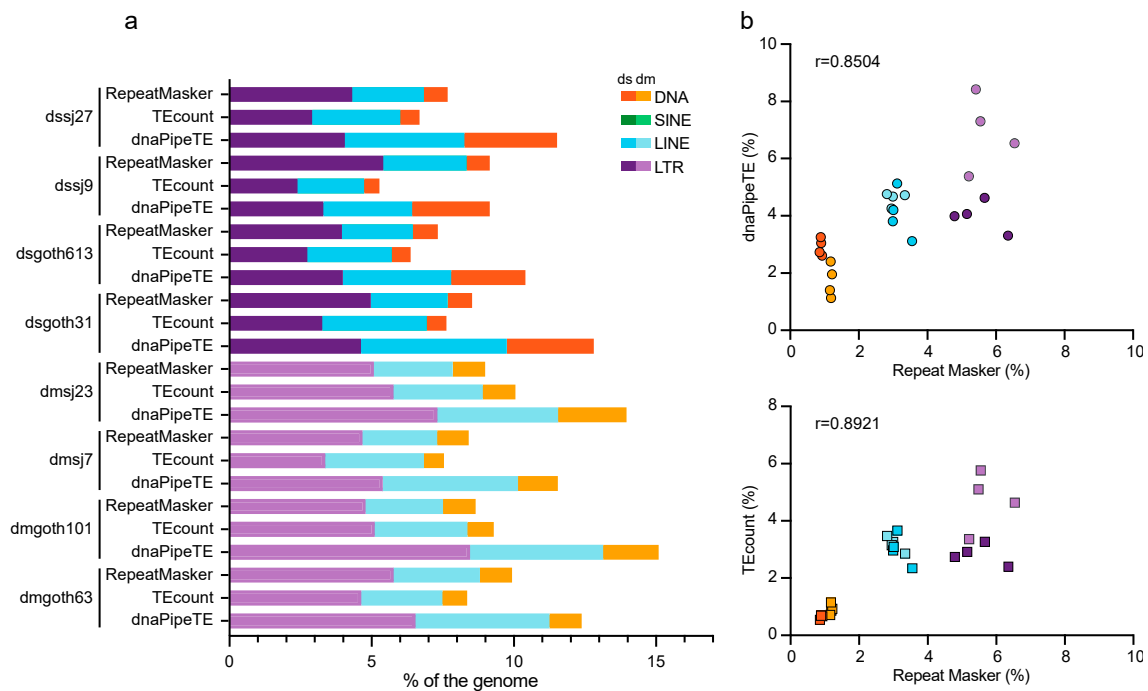


Figure 2. Estimation of the TE percentage in the *D. melanogaster* and *D. simulans* genomes (isogenic wild-type strains). (a) Estimation of the TE percentage using RepeatMasker (ONT chromosome assemblies), and dnaPipeTE or TEcount (Illumina reads). (b) Correlations between the estimates obtained with the indicated methods.

3.4. Comparison of the TEI Sites Identified in the Isogenic Wild-Type Strains Using the Illumina and ONT Data

Before focusing on the results provided by the ONT approach, we first compared these data to the classically used Illumina results based on discordant pairs of reads (method developed in the laboratory, see Material and Methods). The number of TEI sites tended to be higher when using the Illumina data than ONT data (Wilcoxon paired test, p -value = 0.023). This could be due to the presence of false positives caused by PCR artefacts during the Illumina library preparation [18], and/or to the fact that some TEIs might have been too short (fragmented or partially deleted) to be identified using the assembled ONT data. Using the Illumina approach, TEI numbers were significantly lower in the *D. simulans* than in the *D. melanogaster* strains (Wilcoxon test, p -value = 0.029), but not when using the ONT data (Wilcoxon test, p -value = 0.343) (Figure 3). This may reflect a bias towards *D. melanogaster* sequences in our TE reference file, and/or a long-term difference in TE dynamics between these species [25,61]. Comparisons (chi-square tests) of TEI distributions across TE groups (DNA, LINE, LTR) (see Table S3) showed that in *D. simulans*, the distributions obtained using both approaches were similar. Conversely, in *D. melanogaster*, the TEI number for retrotransposons was significantly higher relative to the other groups, when using the Illumina approach. This may be due to the higher propensity of *D. melanogaster* retrotransposons to be involved in Illumina PCR chimeras [18] because of their higher genome occupancy (Figure 2), and this difference may be amplified by the exponential behavior of the PCR reaction.

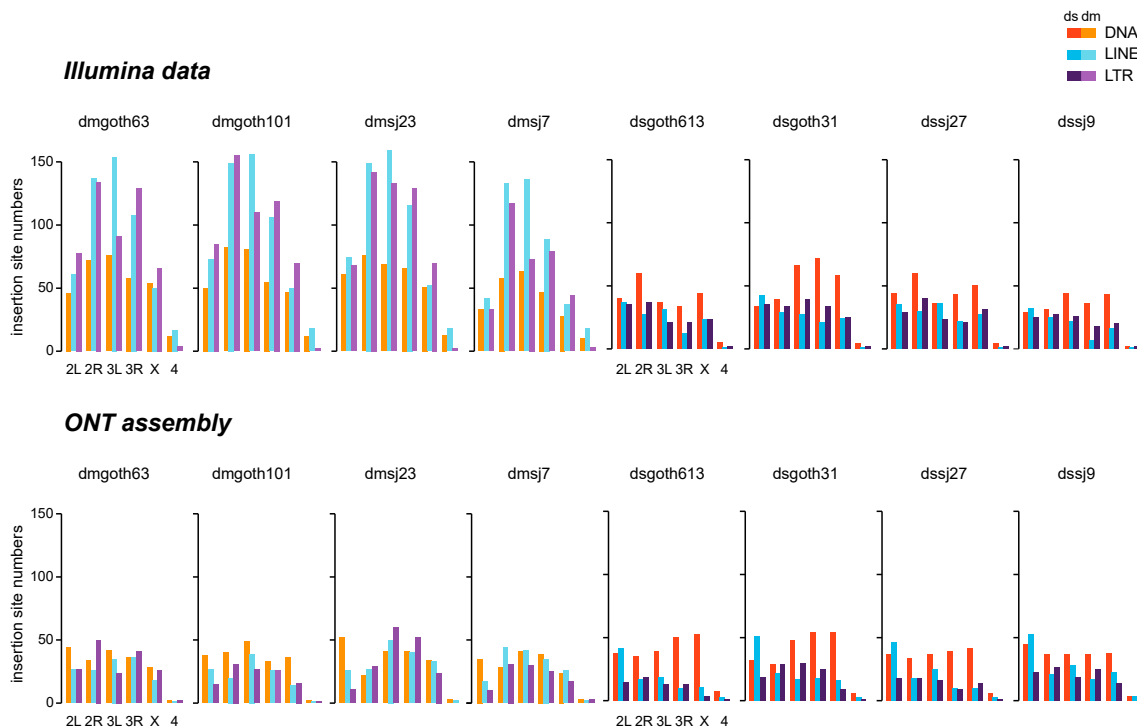


Figure 3. Insertion site numbers for each TE group and per chromosome, determined using Illumina data (upper panels) or Oxford Nanopore Technology (ONT) chromosome assemblies (lower panels).

In the subsequent analyses, only TEIs identified using the ONT approach (i.e., the most reliable set of recent insertions) were considered.

3.5. TEI Landscape in the Isogenic Wild-Type Strains

Using the ONT approach, the de novo genome assembly of each wild-type strain was compared with the reference genome and the detected insertional structural variants were called global variants (see Figure 1). These global variants correspond to the most recent TEIs. On average, there were 492 and 456 global variants in *D. melanogaster* and *D. simulans*, respectively (Table 2).

Table 2. Number of transposable elements insertions (TEIs) identified as global variants in the Oxford Nanopore Technology (ONT) chromosome assemblies.

	dmgoth63	dmgoth101	dmsj23	dmsj7	dsgoth613	dsgoth31	dssj27	dssj9
Total Insertion Number	515	448	550	456	434	496	420	474

DNA transposons were the most abundant group in both species (188 and 215 copies, on average, in *D. melanogaster* and *D. simulans*, respectively), and LTR retrotransposons the least abundant (147 and 117 copies, on average, in *D. melanogaster* and *D. simulans*, respectively). These results may seem in contradiction with the previous data on genome occupancy. However, in this analysis only recent insertions were considered. Moreover, as DNA transposons are in general smaller than LTR retrotransposons, similar levels of genome occupancy correspond to higher copy numbers for DNA transposons than for LTR retrotransposons.

Comparison of the locations of the insertions identified in the chromosome assemblies showed that 22 global variants were present in all four *D. melanogaster* strains, and 23 in all four *D. simulans* strains. These were mainly DNA transposons ($n = 9$ and $n = 10$, respectively). The number of shared pairwise global variants was rather low, roughly 10% of all insertions in most comparisons (Figure 4a).

D. simulans strains appeared equally distant in terms of insertion sites. Conversely, a geographical structuring could be observed in the *D. melanogaster* comparisons: strains from the same population shared more insertion sites than strains from distinct populations.

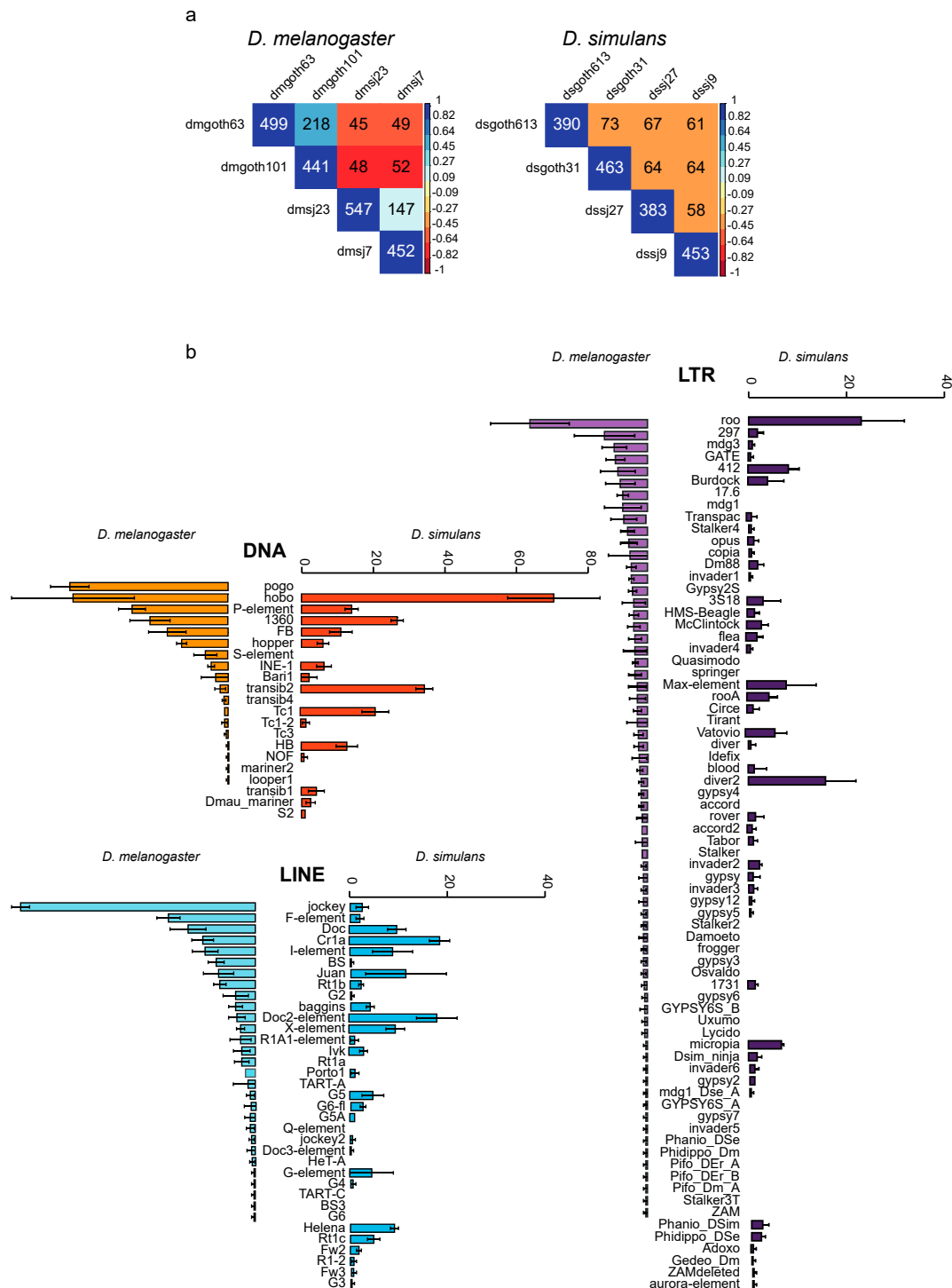


Figure 4. Global variant copy numbers in wild-type *D. melanogaster* and *D. simulans* strains. (a) Number of shared global variants among strains. The color scale (on the right of each panel) shows the distance based on the number of pairwise shared insertions (indicated in black in the figure). Values in white correspond to the total numbers of the identified insertions for the considered strains. (b) Mean TEI numbers for the indicated TE groups computed in the wild-type *D. melanogaster* and *D. simulans* strains based on the ONT chromosome assemblies.

The mean copy numbers for the different TE families were weakly correlated between *D. melanogaster* and *D. simulans* (Figure 4b) (Spearman $\rho = 0.33$, p -value = $1e-4$, across 129 TE families). Few families were found in the *D. simulans* strains but not in the *D. melanogaster* strains, and vice versa. In *D. melanogaster* strains, the most abundant families were roo (mean copy number: 24.00), jockey (mean copy number: 48.00), and pogo (mean copy number: 44.25), for LTR retrotransposons, LINEs, and DNA elements, respectively. In *D. simulans*, they were roo (mean copy number: 23.00), Cr1a (mean copy number: 18.50), and hobo (mean copy number: 70.50). In addition, some TE families displayed different copy numbers across strains. For instance, the 297 family had 18 copies in dmgoth63, 6 in dmgoth101, 6 in dmsj23, and 5 in dmsj7. Such patterns are suggestive of recent, independent activations, or even bursts of some families in specific strains, as suggested by in situ hybridization studies in a large number of samples [62]. Kofler et al. (2015) studied TE patterns in *D. melanogaster* and *D. simulans* field samples using Illumina pool-seq data [63]. By computing the insertion frequencies for each family of a subset of 121 TE families, they established that LTR elements were more frequent in *D. melanogaster* than in *D. simulans* populations, whereas DNA transposons were more frequent in *D. simulans* samples. A similar trend was observed in the present work: 147 LTR retrotransposon insertions in *D. melanogaster* and 117 in *D. simulans* (Wilcoxon test p -value = 0.343); 188 DNA transposon insertions in *D. melanogaster* and 215 in *D. simulans* (Wilcoxon test p -value = 0.029).

3.6. Comparison of TE Dynamics in Isogenic Wild-Type *D. Melanogaster* and *D. Simulans* Strains by Studying TEI Sequences in ONT Assemblies

The major advantage of the ONT approach is its ability to retrieve whole TEI sequences, while short read-based approaches only give access to TE insertion sites. First, the TEI sizes across strains were compared by parsing the BLAST results at the insertion level and by computing the insertion lengths (Figure 5a). The mean insertion lengths (i.e., fragment sizes) significantly varied among TE groups (2-way ANOVA, p -value = $2e-81$), but not between species (2-way ANOVA, p -value = 0.22). LTR retrotransposons were the largest (mean size = 2692 bp), followed by LINEs (mean size = 1290 bp), and DNA transposons (mean size = 1210 bp). The observed absence of difference between species in these global variants differs from what was previously described. Indeed, for a subset of 15 families, Lerat et al. found that TE copies were more internally deleted (i.e., shorter) in *D. simulans* than in *D. melanogaster* [24]. However, analysis of these 15 families using our ONT data indicated that they displayed, on average, longer fragment sizes compared with the other TE families in *D. melanogaster* (Wilcoxon test, p -value = $8e-19$), but not in *D. simulans* (Wilcoxon test, p -value = 0.34) [24]. This suggests that Lerat et al. 2011 focused on TE families that have particularly large copies in *D. melanogaster* [24], probably because they have been more studied in the past due to their easier analysis by in situ hybridization on polytene chromosomes [7,25,64].

Then, the Refiner module of RepeatModeler (<http://www.repeatmasker.org/RepeatModeler>) was used to compute the intra-family sequence divergence (average Kimura distance) (Figure 5b). This measure is a proxy of the time passed since the last transposition wave(s). Overall, these distributions were not significantly different between *D. melanogaster* and *D. simulans* and among TE groups (2-way ANOVA; species effect, p -value = 0.151; group effect, p -value = 0.701), showing that the TE recent dynamics are similar in these two species. However, in *D. simulans*, DNA transposons displayed significantly higher intra-family divergence compared with LTR retrotransposons (Wilcoxon test, p -value = 0.023). This suggests that among the most recent transposition events, DNA transposon insertions occurred slightly less recently in *D. simulans*.

Kofler et al. 2015 assumed that population frequencies of TE insertions provide an estimator for the insertion age. However, we find that their population frequencies were not correlated with our measures of intra-family sequence divergence (Spearman correlation coefficients: -0.714 (p -value = 0.136) and 0.116 (p -value = 0.827) for *D. melanogaster* and *D. simulans*, respectively). We think that intra-family sequence divergence is a more direct estimate of the age of transposition events; however, this discrepancy may also reflect differences in the origins of the sampled flies [61,64]. Alternatively,

it may suggest that other factors influence insertion frequencies, besides the age since the initial transposition burst. In addition, our analysis only included TEIs that are not found in the reference genome, i.e., TEIs that result from transposition events more recent than the set-up of the actual populations. Altogether, while the TE ancient dynamics are different between *D. melanogaster* and *D. simulans* [60], the present results suggest that *D. melanogaster* and *D. simulans* TE landscapes are rather similar when comparing only global variants (i.e., the subset of the most recent insertions). As already proposed [25], this may reveal that the colonization of *D. simulans* genome by TEIs has now reached a state similar to that of *D. melanogaster*, although it started more recently.

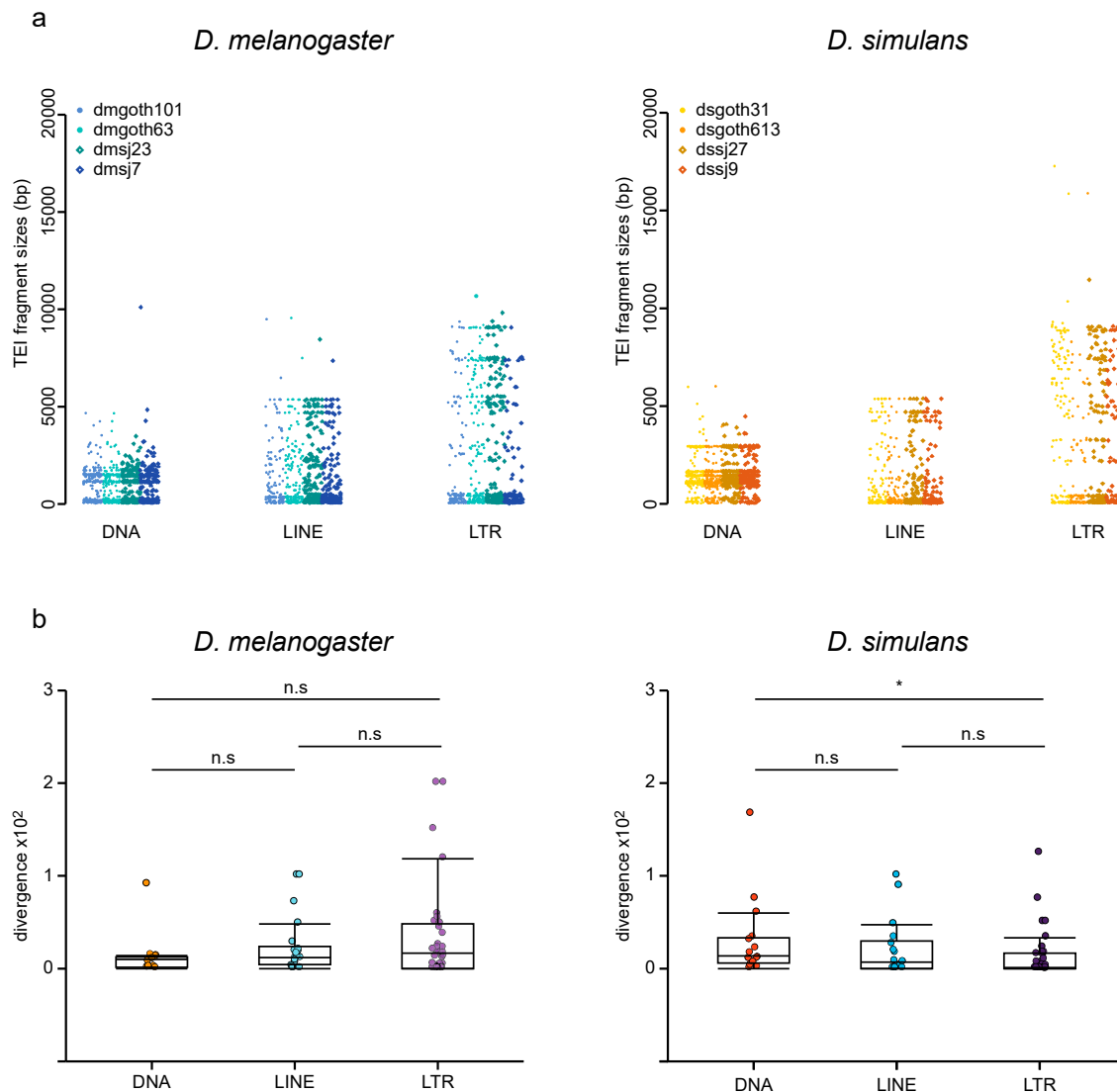


Figure 5. Global variant sequence analysis in wild-type *D. melanogaster* and *D. simulans* strains. (a) Distributions of TEI copy lengths (i.e., fragment size) in bp for all global variants across strains and TE groups. (b) Intra-family sequence divergence (average Kimura distance) computed per strain and per TE family.

3.7. piRNAs, piRNA Clusters and TEIs in Isogenic Wild-Type Strains

Another way to study TE dynamics is to understand the way the production of piRNAs is linked to the TEI type and structure. Indeed, some relationships might exist between piRNA abundance and the recent activity of TEIs, estimated by the intra-family sequence divergence. Therefore, piRNA production, TE length and intra-family sequence divergence were analyzed for each TE

group and strain. This analysis highlighted a significant TE group effect: piRNA counts were higher in retrotransposon families (LTR elements and LINES) than in DNA transposon families (p -value = $2e-9$). Moreover, piRNA counts were significantly and positively correlated with genome occupancy (p -value = $5e-7$), which strongly depends on TE copy number (Figure 6a). The hypothesis that TE copy numbers determine piRNA abundance was previously suggested in *D. melanogaster* [65,66] and is confirmed here also for *D. simulans*. However, it should be noted that genome occupancy accounts only for 6.2% of the total variation of piRNA counts, indicating that many other factors are involved in TE control.

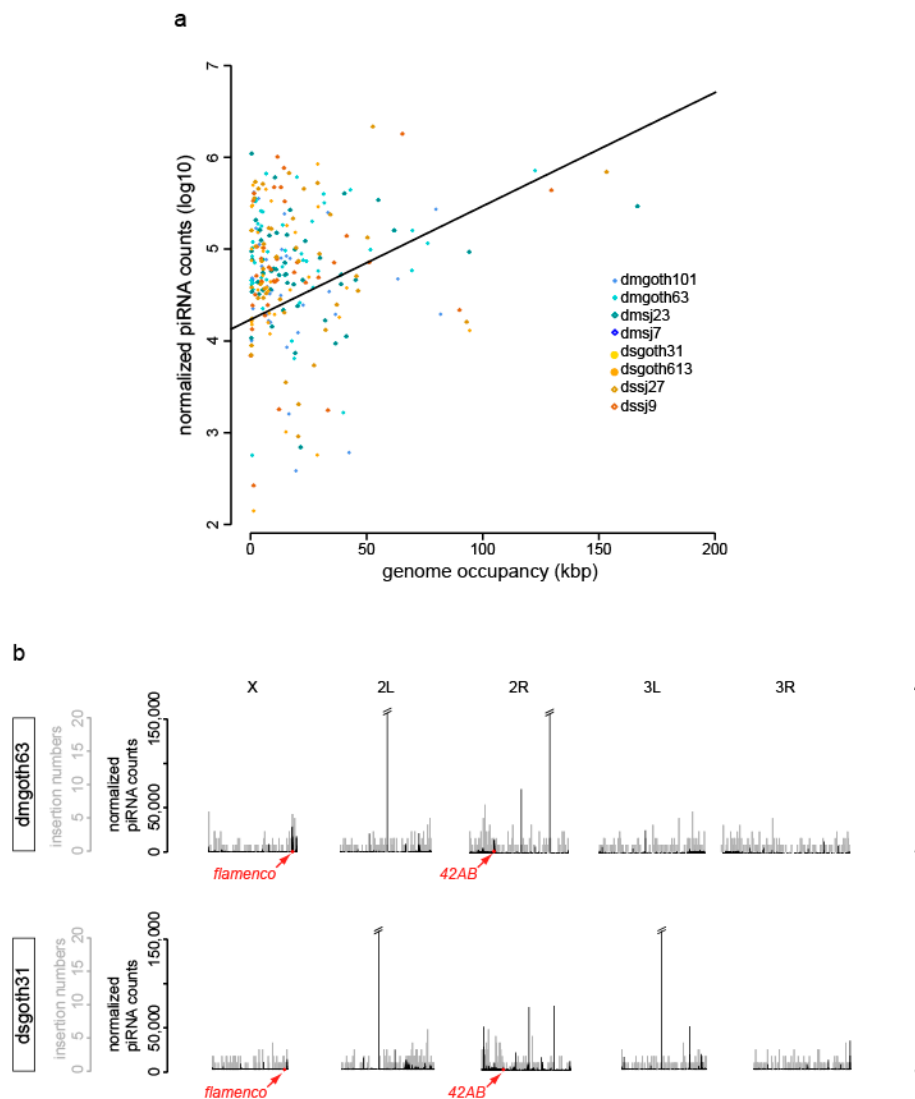


Figure 6. piRNA analyses in wild-type *D. melanogaster* and *D. simulans* strains. (a) Normalized piRNA counts (log10) relative to genome occupancy for all strains and the two species and linear regression curve. Each dot is a TE family. (b) Results for the dmgoth63 and dsgoth31 strains are shown as examples. Uniquely mapping piRNAs along ONT chromosome assemblies (black, normalized piRNA counts). Global variants identified along ONT chromosome assemblies (gray). Red arrows indicate flamenco (X chromosome) and 42AB (2R chromosome). Data for the other strains are provided in Figure S2. The off-scale peaks might correspond to microRNAs that are absent from miRBase.

These observations are also in agreement with the idea that newly integrated copies become piRNA producers [67], and that longer copies produce more piRNAs. It should be noted that retrotransposons are on average longer than DNA transposons.

As ONT assemblies also include piRNA cluster sequences, 42AB and flamenco (the two major piRNA cluster producers in *D. melanogaster*) could be retrieved using their flanking genes (see Material and Methods) [68] from each assembly. Alignment of the uniquely mapped piRNA sequences against the assembly of each wild-type isogenic strain (Figure 6b and Figure S2, black lines) indicated that the regions corresponding to 42AB and flamenco did not display any enrichment in global variant insertion numbers (Figure 6b, gray lines). This indicates that recent TEIs are not specifically enriched in the two major piRNA cluster producers in *D. melanogaster* and *D. simulans* strains. Therefore, the analysis of the de novo assembled genomes to follow the piRNA cluster dynamics in these isogenic wild-type strains did not highlight the previously reported high TEI insertion rate within piRNA clusters [26,50,69,70]. Our data suggests the number of recent TEIs fixed in these piRNA clusters is not different compared with anywhere else in the genome. This discrepancy could be explained by the high frequency of deletions (from several base pairs up to several kilobases) that seems to occur in these regions and that affect ancient TEs, which remain as vestiges in these loci, and also recently inserted TEs [50].

3.8. Recent TEIs May Not Be Frequent Enough to Be Incorporated in the Assembled Genomes

To challenge the ONT assembly approach, a bioinformatic analysis was performed to identify recent LTR TEIs that occurred during the last 73 generations (G73) in the unstable Piwi KD line (Materials and Methods and [28]). As a control, to estimate the basal transposition rate when TEs are normally repressed by the functional piRNA pathway, the genome of the hundredth generation (called G0-F100) after establishment of the stable G0 isofemale line was also sequenced. Using the pipeline for detection of global variants (Figure 1), no new ZAM insertion could be detected in the G73 assembled genome compared with the G0 reference genome. This is not consistent with previous data obtained by PCR quantification of the ZAM copy number [28]. Therefore, *in situ* hybridization analysis was performed to determine whether de novo ZAM insertions were present on polytene chromosomes of G73 male larvae (Figure 7a and Figure S3). This analysis confirmed the presence of the two preexisting ZAM insertions identified on chromosome 2R as global variants in the G0 de novo assembly (compared with the Dmel_R6.23 reference genome). These two insertions were also detected in all three G73 larvae analyzed, as well as many other ZAM signals that were not observed in the G0 samples (Figure 7a and Figure S3). As each of these many G73-specific new ZAM insertions was present in a single larva, they were not incorporated in the G73 de novo assembled genome due to their low frequency, and therefore could not be detected as global variants. Based on the G0 assembled genome, the sequences of the two shared ZAM detected by FISH on chromosome 2R could be accessed. One contained the full length canonical ZAM consensus sequence, while the other displayed an internal deletion (Figure 7b).

3.9. A Long Read-Based Pipeline to Detect Low Frequency TEI Polymorphisms

To determine whether ONT can be used to detect TEIs with a frequency not high enough to be recovered in the assembled haplotype, an approach to identify “minor insertional variants” (MIV) was developed (Material and Methods, paragraph 2.9, and Figure 1 (gray)). Minimap2 was used to map each individual long read to the corresponding assembled genome, and Sniffles to obtain the list of variants that had been neglected during the assembling process. Some of the sequences identified as MIVs matched to the 60 canonical LTR TE consensus sequences (Materials and Methods).

As expected, very few LTR MIVs were detected in the G0-F100 “stable line”. Only copia and roo, which have high transposition rates [71], exhibited more than four variants (14 and 22, respectively) among the 51 LTR MIVs detected (Figure 7c). Also in the G73 line, copia and roo were among the more active LTR families (35 and 48 LTR MIVs among the 274 LTR MIVs detected) (Figure 7c). However, two other LTR families, ZAM and gtwin (51 and 93 LTR MIVs, respectively), showed a 50-fold increase in G73 compared with G0-F100, which is more than an order of magnitude higher than what observed for any other LTR family.

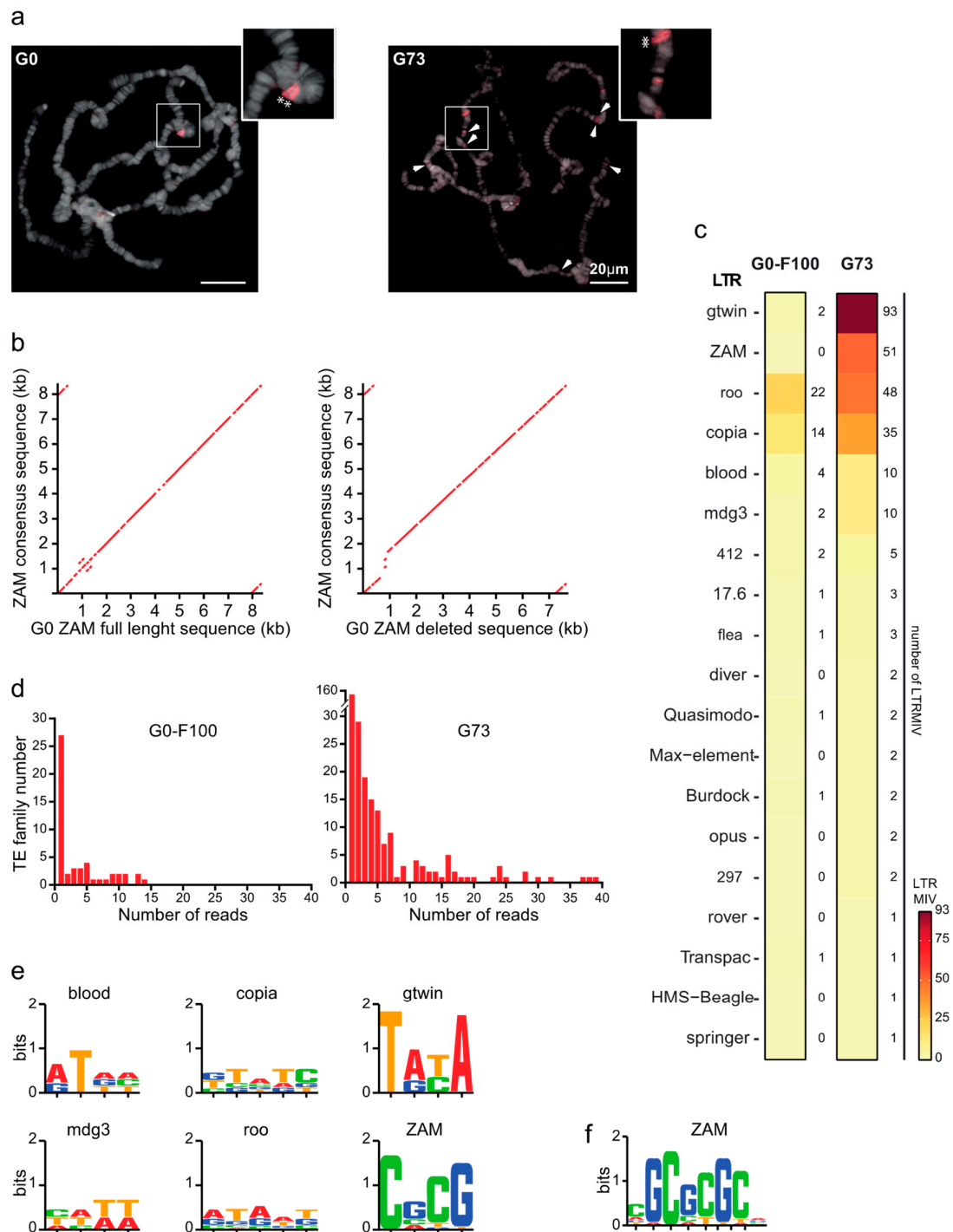


Figure 7. Characterization of the Long-Terminal Repeat minor insertion variant (LTR MIV) in the stable (G0) and unstable (G73) lines. **(a)** ZAM copies visualized by fluorescent in situ hybridization in G0 (left) and G73 (right) polytene chromosomes. The two global variants correspond to non-reference ZAM copies present in G0 and G73 (asterisks in the zoomed images). Arrowheads show the new ZAM insertions in G73. More examples are presented in Figure S3. **(b)** Dot plot of the sequence comparison between the ZAM sequences accessed from the de novo assembled G0 genome and the ZAM consensus sequence. **(c)** Heat map of the LTR MIV detected in the G0-F100 (stable) and G73 (unstable) libraries. **(d)** Histograms showing the number of reads supporting each LTR MIV. **(e)** Sequence logo of TSD defined using the LTR MIV automatic detection procedure. **(f)** the ZAM TSM motif defined using the automatic and manual LTR MIV detection procedures.

The next question was to determine whether the 274 LTR MIVs, present at low frequency in G73, had occurred after the establishment of the isofemale line. Indeed, such insertions could have been already present in G0 at high frequency (and therefore, could have been incorporated in the G0 but not in the G73 assembled genome) or at low frequency (and, therefore, detectable only as MIVs in G0). The first hypothesis was ruled out by comparing global deletions in G73 and G0. Very few G0 insertions were lost in the G73 assembly and they all belonged to five LTR families (mdg3, Transpac, 3S18, blood, and driver) that did not show a large MIV increase in G73 (data not shown). The total absence of LTR MIVs in G0 was not in favor of the second hypothesis.

As a large fraction of the 274 LTR MIVs in G73 were supported by a single read (Figure 7d), the next step was to check whether they were bona fide insertions by looking for insertional hallmarks, such as the target site duplications (TSDs) that occur upon integration as a result of staggered double-strand breaks at this site [72]. Flanking duplications were first detected automatically for each of the top six LTR families (mdg3, blood, copia, roo, ZAM, and gtwin) by aligning the two 30nt-long sequences that flank each putative LTR MIV extracted from the read supporting the variant. This analysis showed that depending on the LTR family, 30–80% of MIVs were flanked by a short duplication of the expected size (4 or 5 nt) (Table 3) [52]. The TSD consensus sequences identified are presented in Figure 7e.

Table 3. Target site duplication (TSD) flanking Long-Terminal Repeat minor insertion variants (LTR MIVs) in the G73 line.

	LTR Family					
	gtwin	roo	ZAM	copia	Blood	mdg3
Total LTR MIV detected (<i>n</i>)	93	48	51	35	10	10
TSD automatic detection (<i>n</i>)	66	15	25	11	8	5
TSD automatic detection (%)	71	31	49	31	80	50
Additional TSD manually detected (<i>n</i>)	NA	NA	23	NA	NA	NA

The failure to automatically detect a TSD for the other LTR MIVs could be due to the frequent sequencing errors, a known ONT drawback. When located in the genome-LTR junction region, such errors, which may include several nt-long indels, could impair the automatic detection of the expected TSD, as shown in Figure S4 for the manual inspection of the 2R-33863 putative ZAM insertion. Even when junctions are correctly determined, a simple sequencing error in one of the duplicated sequences might prevent their perfect matching. However, it was possible to correct the errors present in these single reads by aligning them with the empty genomic target present on the assembled genome (see, Figure S4). Using this method to manually inspect the sequence of all 51 ZAM variant reads, 48 bona fide insertions were identified, as judged by the presence of the expected 4-nt TSD included in a palindromic GC-rich 6-nt target site motif (TSM) (Figure 7f) [52,55].

Therefore, despite ONT low sequencing accuracy, LTR MIVs could be detected with high sensitivity (insertions present in a population at a frequency <1%, because detected as single reads in a 197x average coverage library) and specificity (FDR of 3/51 = 6%).

3.10. Invading LTR Elements Are Not Preferentially Trapped by piRNA Clusters

It is widely assumed that a TE invasion is stopped when a member of the TE family jumps into a piRNA cluster that then triggers the production of piRNAs to repress this TE family (i.e., trap model) [27]. Long-read sequencing data allowed determining whether new insertions accumulated in major piRNA source loci during the 73 generations of LTR TE derepression. Comparison of the 42 major piRNA clusters after their localization in the G0 and G73 assemblies (Table S4) did not highlight any new TEI into any of these piRNA clusters in the G73 assembled genome. However, new insertions that occurred during the 73 generations of piRNA pathway impairment could still segregate as MIVs in the G73 population. Indeed, among the 274 LTR MIVs present in G73, 6.57% (*n* = 18) were located within the 42 major piRNA producers (Figure 8). However, this proportion was very similar to that of the

piRNA cluster size relative to the total de novo assembled genome size (7.36%). Therefore, unlike what expected in the trap model, LTR retrotransposons do not seem to have preferentially accumulated in piRNA clusters during the 73 generations of transposition burst. Specifically, assuming a binomial law with $n = 274$ and $p = 0.0736$ and using a one-tailed test, more than 29 insertions (and not the 18 detected) belonging to many different TE families would have been necessary to validate the hypothesis that piRNA clusters are TE trappers (5% probability threshold).

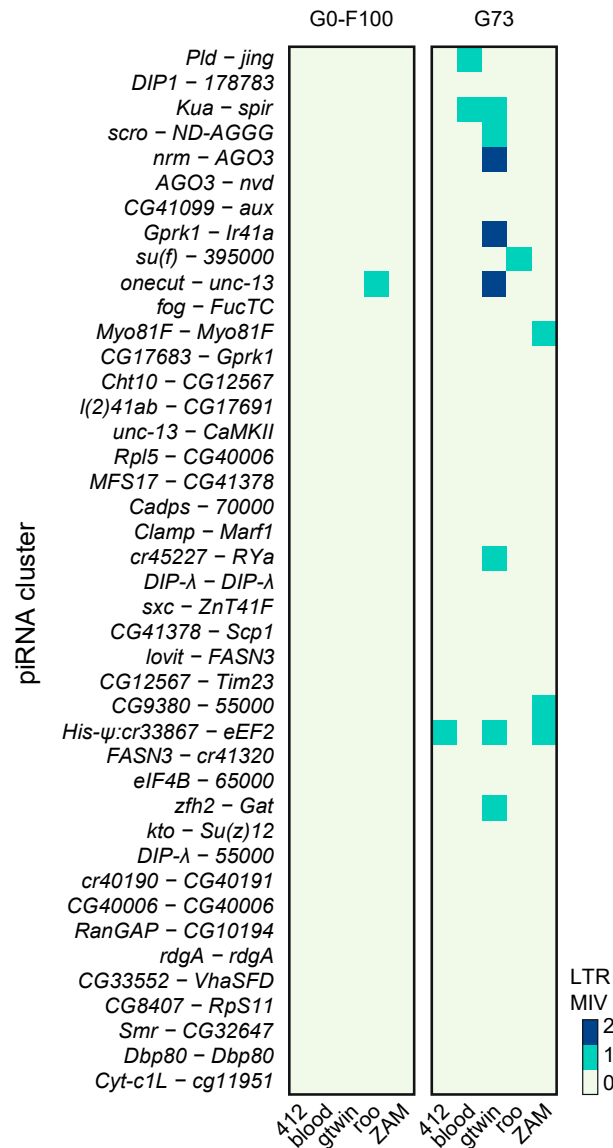


Figure 8. Heat map of the LTR MIVs inserted in piRNA clusters and detected in the G0-F100 and G73 lines.

More than 50% of the LTR MIVs located in piRNA clusters belonged to the gtwin family, suggesting that this family inserts preferentially into piRNA clusters. Indeed, among the 93 gtwin MIVs, 11 (11.8%) were found in piRNA clusters, which is very close to the minimal number ($n = 12$) required to reject the null hypothesis of random insertion in the genome (binomial law with $n = 93$, $p = 0.0736$, and 5% probability threshold). More data on de novo gtwin mobilization are needed to confirm their preferential integration in piRNA clusters during a transposition burst and to support the trap model for this TE family.

4. Conclusions

Our work demonstrates that long reads are crucial in order to finely describe TE landscapes at the intra-genome scale. Using isogenic wild-type strains and an unstable line with a succession of transposition bursts, we could characterize the most common TE variants in different strains and identify TE minor variants observed soon after transposition. The parallel analysis of two close species (*D. melanogaster* and *D. simulans*) and two genetic backgrounds allowed us to show that overall, TE recent dynamics are quite similar between species and among strains. However, there is still some strain specificity concerning the identity of the most recently active TE families. ONT is also a powerful tool to investigate the dynamics of piRNA clusters, which are in general inaccessible using short-read sequencing methods. We show here that recent TEIs are not enriched in piRNA clusters, despite recent bursts of TE transposition. Moreover, ONT allows detecting very recent TEIs that are sequenced as singleton reads.

Supplementary Materials: The following are available online at <http://www.mdpi.com/2073-4409/9/8/1776/s1>, Figure S1: D-genies genome-wide dot plot of ONT assembly contigs versus reference genome, Figure S2: piRNA analyses in wild-type strains, Figure S3: ZAM copies were visualized by fluorescent in situ hybridization on G0 and G73, Figure S4: Alignments of the 2R-33863 insertion variant to the ZAM consensus sequence. Table S1: Statistics about sequencing data. All lengths are expressed in bases. Quality is expressed in standard Phred scale, Table S2: Genome size estimations using different methods, Table S3: Comparison of TEI distributions across TE groups using chi-square tests, Table S4: piRNA cluster coordinates based on flanking genes in de novo assembled genomes.

Author Contributions: Conceptualization, S.C.; Data curation, N.T.-M.D., M.M., F.S., Y.O., N.B., J.S.-O., D.S. and A.P.; Formal analysis, M.M., M.B., M.F., N.T.-M.D., F.S., A.P. and V.M.; Funding acquisition, C.V. and S.C.; Investigation, S.C., M.F.; C.V.; Methodology, P.V., A.P., F.S. and M.F.; Software, M.M., N.T.-M.D., P.V., F.S. and M.F.; Supervision, M.F. and S.C.; Visualization, B.M.; Writing—original draft, C.V., F.S. and S.C.; Writing—review & editing, C.V., F.S., M.F. and S.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Fondation pour la Recherche Médicale, grant number “DEQ20180339167” to S.C., by the ANR Exhyb to C.V., by the CNRS.

Acknowledgments: We thank J. Gonzalez and C. Goubert for the discussion, C. Jourdan and B. Barckmann for G0-F100, G73 and G0 DNA extraction, C. Brun for the polytene mapping, D. Gourion for the modelization and statistics in Section 3.10 and A.S. Fiston-Lavier for the discussions. We thank Ndomassi Tando and the IRD itrop “Plantes Santé” bioinformatic platform for providing HPC resources and support for our research project. T.-M.N.D. was supported by France Excellence. D.S. acknowledges financial support from France Génomique National infrastructure, funded as part of “Investissement d’avenir” program managed by Agence Nationale pour la Recherche (contract ANR-10-INBS-09).”

Conflicts of Interest: The authors declare no competing interests.

Data Availability: Long reads sequencing data used for this study have been deposited at ENA (<https://www.ebi.ac.uk/ena>) under the accession numbers PRJEB39340 and ERP122844. The small RNA-seq datasets and the Illumina DNA-seq datasets were deposited in NCBI SRA (<https://www.ncbi.nlm.nih.gov/sra>) under the accession numbers PRJNA644327 and PRJNA644748, respectively. Release 6.23 of the *D. melanogaster* genome and Release 2.2 of the *D. Simulans* used in this study are available on FlyBase (<http://www.flybase.org>). Bioinformatic scripts and pipelines used for long reads analyses are available at <https://github.com/DrosophilaGenomeEvolution/TrEMOLO> and for small reads Illumina insertion at <https://gitlab.in2p3.fr/pveber/te-insertion-detector/>.

References

1. Biémont, C.; Vieira, C. Genetics: Junk DNA as an evolutionary force. *Nature* **2006**, *443*, 521–524. [[CrossRef](#)]
2. Wicker, T.; Sabot, F.; Hua-Van, A.; Bennetzen, J.L.; Capy, P.; Chalhoub, B.; Flavell, A.; Leroy, P.; Morgante, M.; Panaud, O.; et al. A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.* **2007**, *8*, 973–982. [[CrossRef](#)] [[PubMed](#)]
3. Kapitonov, V.V.; Jurka, J. A universal classification of eukaryotic transposable elements implemented in Repbase. *Nat. Rev. Genet.* **2008**, *9*, 411–412. [[CrossRef](#)] [[PubMed](#)]
4. Brookfield, J.F.; Montgomery, E.; Langley, C.H. Apparent absence of transposable elements related to the P elements of *D. melanogaster* in other species of *Drosophila*. *Nature* **1984**, *310*, 330–332. [[CrossRef](#)] [[PubMed](#)]
5. Black, D.M.; Jackson, M.S.; Kidwell, M.G.; Dover, G.A. KP elements repress P-induced hybrid dysgenesis in *Drosophila melanogaster*. *EMBO J.* **1987**, *6*, 4125–4135. [[CrossRef](#)] [[PubMed](#)]

6. Biémont, C.; Ronsseray, S.; Anxolabéhère, D.; Izaabel, H.; Gautier, C. Localization of P elements, copy number regulation, and cytotype determination in *Drosophila melanogaster*. *Genet. Res.* **1990**, *56*, 3–14. [[CrossRef](#)]
7. Biémont, C.; Monti-Dedieu, L.; Lemeunier, F. Detection of Transposable Elements in *Drosophila* Salivary Gland Polytene Chromosomes by In Situ Hybridization. In *Mobile Genetic Elements: Protocols and Genomic Applications*; Miller, W.J., Capy, P., Eds.; Methods in Molecular Biology; Humana Press: Totowa, NJ, USA, 2004; pp. 21–28, ISBN 978-1-59259-755-0.
8. Ignatenko, O.M.; Zakharenko, L.P.; Dorogova, N.V.; Fedorova, S.A. P elements and the determinants of hybrid dysgenesis have different dynamics of propagation in *Drosophila melanogaster* populations. *Genetica* **2015**, *143*, 751–759. [[CrossRef](#)]
9. Onder, B.S.; Kasap, O.E. P element activity and molecular structure in *Drosophila melanogaster* populations from Firtina Valley, Turkey. *J. Insect Sci. Online* **2014**, *14*, 16. [[CrossRef](#)]
10. Novák, P.; Neumann, P.; Macas, J. Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. *BMC Bioinform.* **2010**, *11*, 378. [[CrossRef](#)]
11. Goubert, C.; Modolo, L.; Vieira, C.; ValienteMoro, C.; Mavingui, P.; Boulesteix, M. De Novo Assembly and Annotation of the Asian Tiger Mosquito (*Aedes albopictus*) Repeatome with dnaPipeTE from Raw Genomic Reads and Comparative Analysis with the Yellow Fever Mosquito (*Aedes aegypti*). *Genome Biol. Evol.* **2015**, *7*, 1192–1205. [[CrossRef](#)]
12. Granzotto, A.; Lopes, F.R.; Lerat, E.; Vieira, C.; Carareto, C.M.A. The evolutionary dynamics of the Helena retrotransposon revealed by sequenced *Drosophila* genomes. *BMC Evol. Biol.* **2009**, *9*, 174. [[CrossRef](#)] [[PubMed](#)]
13. Rebollo, R.; Lerat, E.; Kleine, L.L.; Biémont, C.; Vieira, C. Losing helena: The extinction of a drosophila line-like element. *BMC Genom.* **2008**, *9*, 149. [[CrossRef](#)] [[PubMed](#)]
14. Schlötterer, C.; Tobler, R.; Kofler, R.; Nolte, V. Sequencing pools of individuals—Mining genome-wide polymorphism data without big funding. *Nat. Rev. Genet.* **2014**, *15*, 749–763. [[CrossRef](#)] [[PubMed](#)]
15. Rahman, R.; Chirn, G.; Kanodia, A.; Sytnikova, Y.A.; Brembs, B.; Bergman, C.M.; Lau, N.C. Unique transposon landscapes are pervasive across *Drosophila melanogaster* genomes. *Nucleic Acids Res.* **2015**, *43*, 10655–10672. [[CrossRef](#)] [[PubMed](#)]
16. Fiston-Lavier, A.-S.; Barrón, M.G.; Petrov, D.A.; González, J. T-lex2: Genotyping, frequency estimation and re-annotation of transposable elements using single or pooled next-generation sequencing data. *Nucleic Acids Res.* **2015**, *43*, e22. [[CrossRef](#)] [[PubMed](#)]
17. Kofler, R.; Gómez-Sánchez, D.; Schlötterer, C. PoPoolationTE2: Comparative population genomics of transposable elements using Pool-Seq. *Mol. Biol. Evol.* **2016**, *33*, 2759–2764. [[CrossRef](#)]
18. Treiber, C.D.; Waddell, S. Resolving the prevalence of somatic transposition in *Drosophila*. *eLife* **2017**, *6*, e28297. [[CrossRef](#)]
19. Pollard, M.O.; Gurdasani, D.; Mentzer, A.J.; Porter, T.; Sandhu, M.S. Long reads: Their purpose and place. *Hum. Mol. Genet.* **2018**, *27*, R234–R241. [[CrossRef](#)]
20. Lerat, E.; Goubert, C.; Guirao-Rico, S.; Merenciano, M.; Dufour, A.-B.; Vieira, C.; González, J. Population-specific dynamics and selection patterns of transposable element insertions in European natural populations. *Mol. Ecol.* **2019**, *28*, 1506–1522. [[CrossRef](#)]
21. Chakraborty, M.; VanKuren, N.W.; Zhao, R.; Zhang, X.; Kalsow, S.; Emerson, J.J. Hidden genetic variation shapes the structure of functional elements in *Drosophila*. *Nat. Genet.* **2018**, *50*, 20–25. [[CrossRef](#)]
22. Miller, D.E.; Staber, C.; Zeitlinger, J.; Hawley, R.S. Highly Contiguous Genome Assemblies of 15 *Drosophila* Species Generated Using Nanopore Sequencing. *G3 Genes Genomes Genet.* **2018**, *8*, 3131–3141. [[CrossRef](#)]
23. Solares, E.A.; Chakraborty, M.; Miller, D.E.; Kalsow, S.; Hall, K.; Perera, A.G.; Emerson, J.J.; Hawley, R.S. Rapid Low-Cost Assembly of the *Drosophila melanogaster* Reference Genome Using Low-Coverage, Long-Read Sequencing. *G3 Genes Genomes Genet.* **2018**, *8*, 3143–3154. [[CrossRef](#)] [[PubMed](#)]
24. Lerat, E.; Burlet, N.; Biémont, C.; Vieira, C. Comparative analysis of transposable elements in the melanogaster subgroup sequenced genomes. *Gene* **2011**, *473*, 100–109. [[CrossRef](#)] [[PubMed](#)]
25. Vieira, C.; Lepetit, D.; Dumont, S.; Biémont, C. Wake up of transposable elements following *Drosophila simulans* worldwide colonization. *Mol. Biol. Evol.* **1999**, *16*, 1251–1255. [[CrossRef](#)]
26. 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.* **2006**, *7*, R112. [[CrossRef](#)]

27. Kofler, R. Dynamics of Transposable Element Invasions with piRNA Clusters. *Mol. Biol. Evol.* **2019**, *36*, 1457–1472. [[CrossRef](#)]
28. Barckmann, B.; El-Barouk, M.; Péliesson, A.; Mugat, B.; Li, B.; Franckhauser, C.; Fiston Lavier, A.-S.; Mirouze, M.; Fablet, M.; Chambeyron, S. The somatic piRNA pathway controls germline transposition over generations. *Nucleic Acids Res.* **2018**, *46*, 9524–9536. [[CrossRef](#)]
29. Romero-Soriano, V.; Burlet, N.; Vela, D.; Fontdevila, A.; Vieira, C.; García Guerreiro, M.P. Drosophila Females Undergo Genome Expansion after Interspecific Hybridization. *Genome Biol. Evol.* **2016**, *8*, 556–561. [[CrossRef](#)]
30. Sun, H.; Ding, J.; Piednoël, M.; Schneeberger, K. findGSE: Estimating genome size variation within human and Arabidopsis using k-mer frequencies. *Bioinformatics* **2018**, *34*, 550–557. [[CrossRef](#)]
31. Di Giovanni, D.; Lepetit, D.; Guinet, B.; Bennetot, B.; Boulesteix, M.; Couté, Y.; Bouchez, O.; Ravallec, M.; Varaldi, J. A behavior-manipulating virus relative as a source of adaptive genes for Drosophila parasitoids. *Mol. Biol. Evol.* **2020**. [[CrossRef](#)]
32. Ballard, J.W.O. Comparative Genomics of Mitochondrial DNA in Drosophila simulans. *J. Mol. Evol.* **2000**, *51*, 64–75. [[CrossRef](#)] [[PubMed](#)]
33. Lerat, E.; Fablet, M.; Modolo, L.; Lopez-Maestre, H.; Vieira, C. TEtools facilitates big data expression analysis of transposable elements and reveals an antagonism between their activity and that of piRNA genes. *Nucleic Acids Res.* **2017**, *45*, e17. [[CrossRef](#)] [[PubMed](#)]
34. Roy, M.; Viginier, B.; Saint-Michel, É.; Arnaud, F.; Ratiner, M.; Fablet, M. Viral infection impacts transposable element transcript amounts in Drosophila. *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 12249–12257. [[CrossRef](#)] [[PubMed](#)]
35. Langmead, B.; Wilks, C.; Antonescu, V.; Charles, R. Scaling read aligners to hundreds of threads on general-purpose processors. *Bioinforma. Oxf. Engl.* **2019**, *35*, 421–432. [[CrossRef](#)] [[PubMed](#)]
36. Langmead, B.; Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **2012**, *9*, 357–359. [[CrossRef](#)] [[PubMed](#)]
37. Grentzinger, T.; Armenise, C.; Brun, C.; Mugat, B.; Serrano, V.; Pelisson, A.; Chambeyron, S. piRNA-mediated transgenerational inheritance of an acquired trait. *Genome Res.* **2012**, *22*, 1877–1888. [[CrossRef](#)]
38. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. J.* **2011**, *17*, 10–12. [[CrossRef](#)]
39. Schmieder, R.; Edwards, R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* **2011**, *27*, 863–864. [[CrossRef](#)]
40. Kolmogorov, M.; Yuan, J.; Lin, Y.; Pevzner, P.A. Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* **2019**, *37*, 540–546. [[CrossRef](#)]
41. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv* **2013**, arXiv:1303.3997 (q-bio).
42. Li, H. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* **2018**, *34*, 3094–3100. [[CrossRef](#)] [[PubMed](#)]
43. Cabanettes, F.; Klopp, C. D-GENIES: Dot plot large genomes in an interactive, efficient and simple way. *PeerJ* **2018**, *6*, e4958. [[CrossRef](#)] [[PubMed](#)]
44. Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **2009**, *25*, 2078–2079. [[CrossRef](#)] [[PubMed](#)]
45. Krumsiek, J.; Arnold, R.; Rattei, T. Gepard: A rapid and sensitive tool for creating dotplots on genome scale. *Bioinformatics* **2007**, *23*, 1026–1028. [[CrossRef](#)]
46. Alonge, M.; Soyk, S.; Ramakrishnan, S.; Wang, X.; Goodwin, S.; Sedlazeck, F.J.; Lippman, Z.B.; Schatz, M.C. RaGOO: Fast and accurate reference-guided scaffolding of draft genomes. *Genome Biol.* **2019**, *20*, 224. [[CrossRef](#)] [[PubMed](#)]
47. Seppey, M.; Manni, M.; Zdobnov, E.M. BUSCO: Assessing Genome Assembly and Annotation Completeness. In *Gene Prediction: Methods and Protocols*; Kollmar, M., Ed.; Methods in Molecular Biology; Springer: New York, NY, USA, 2019; pp. 227–245, ISBN 978-1-4939-9173-0.
48. Nishimura, O.; Hara, Y.; Kuraku, S. gVolante for standardizing completeness assessment of genome and transcriptome assemblies. *Bioinformatics* **2017**, *33*, 3635–3637. [[CrossRef](#)]
49. Hubley, R.; Finn, R.D.; Clements, J.; Eddy, S.R.; Jones, T.A.; Bao, W.; Smit, A.F.A.; Wheeler, T.J. The Dfam database of repetitive DNA families. *Nucleic Acids Res.* **2016**, *44*, D81–D89. [[CrossRef](#)]

50. Zanni, V.; Eymery, A.; Coiffet, M.; Zytnicki, M.; Luyten, I.; Quesneville, H.; Vaury, C.; Jensen, S. Distribution, evolution, and diversity of retrotransposons at the flamenco locus reflect the regulatory properties of piRNA clusters. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 19842–19847. [[CrossRef](#)]
51. Sedlazeck, F.J.; Rescheneder, P.; Smolka, M.; Fang, H.; Nattestad, M.; von Haeseler, A.; Schatz, M.C. Accurate detection of complex structural variations using single-molecule sequencing. *Nat. Methods* **2018**, *15*, 461–468. [[CrossRef](#)]
52. Linheiro, R.S.; Bergman, C.M. Whole Genome Resequencing Reveals Natural Target Site Preferences of Transposable Elements in *Drosophila melanogaster*. *PLoS ONE* **2012**, *7*, e30008. [[CrossRef](#)]
53. Arnaud, F.; Peyretailade, E.; Dastugue, B.; Vaury, C. Functional characteristics of a reverse transcriptase encoded by an endogenous retrovirus from *Drosophila melanogaster*. *Insect Biochem. Mol. Biol.* **2005**, *35*, 323–331. [[CrossRef](#)] [[PubMed](#)]
54. Lavrov, S.; Déjardin, J.; Cavalli, G. Combined immunostaining and FISH analysis of polytene chromosomes. In *Methods in Molecular Biology*; Humana Press: Clifton, NJ, USA, 2004; Volume 247, pp. 289–303. [[CrossRef](#)]
55. Leblanc, P.; Dastugue, B.; Vaury, C. The Integration Machinery of ZAM, a Retroelement from *Drosophila melanogaster*, Acts as a Sequence-Specific Endonuclease. *J. Virol.* **1999**, *73*, 7061–7064. [[CrossRef](#)] [[PubMed](#)]
56. George, P.; Jensen, S.; Pogorelcnik, R.; Lee, J.; Xing, Y.; Brassset, E.; Vaury, C.; Sharakhov, I.V. Increased production of piRNAs from euchromatic clusters and genes in *Anopheles gambiae* compared with *Drosophila melanogaster*. *Epigenetics Chromatin* **2015**, *8*, 50. [[CrossRef](#)] [[PubMed](#)]
57. Dos Santos, G.; Schroeder, A.J.; Goodman, J.L.; Strelets, V.B.; Crosby, M.A.; Thurmond, J.; Emmert, D.B.; Gelbart, W.M. FlyBase: Introduction of the *Drosophila melanogaster* Release 6 reference genome assembly and large-scale migration of genome annotations. *Nucleic Acids Res.* **2015**, *43*, D690–D697. [[CrossRef](#)]
58. Kozomara, A.; Griffiths-Jones, S. miRBase: Annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res.* **2014**, *42*, D68–D73. [[CrossRef](#)]
59. Langmead, B. Aligning Short Sequencing Reads with Bowtie. *Curr. Protoc. Bioinforma.* **2010**, *32*, 11.7.1–11.7.14. [[CrossRef](#)]
60. Mérel, V.; Boulesteix, M.; Fablet, M.; Vieira, C. Transposable elements in *Drosophila*. *Mobile DNA*. in press.
61. Vieira, C.; Fablet, M.; Lerat, E.; Boulesteix, M.; Rebollo, R.; Burlet, N.; Akkouche, A.; Hubert, B.; Mortada, H.; Biémont, C. A comparative analysis of the amounts and dynamics of transposable elements in natural populations of *Drosophila melanogaster* and *Drosophila simulans*. *J. Environ. Radioact.* **2012**, *113*, 83–86. [[CrossRef](#)]
62. Vieira, C.; Biémont, C. Geographical variation in insertion site number of retrotransposon 412 in *Drosophila simulans*. *J. Mol. Evol.* **1996**, *42*, 443–451. [[CrossRef](#)]
63. Kofler, R.; Nolte, V.; Schlötterer, C. Tempo and Mode of Transposable Element Activity in *Drosophila*. *PLoS Genet.* **2015**, *11*, e1005406. [[CrossRef](#)]
64. Biémont, C.; Nardon, C.; Deceliere, G.; Lepetit, D.; Lœvenbruck, C.; Vieira, C. Worldwide Distribution of Transposable Element Copy Number in Natural Populations of *Drosophila Simulans*. *Evolution* **2003**, *57*, 159–167. [[CrossRef](#)] [[PubMed](#)]
65. Kelleher, E.S.; Barbash, D.A. Analysis of piRNA-mediated silencing of active TEs in *Drosophila melanogaster* suggests limits on the evolution of host genome defense. *Mol. Biol. Evol.* **2013**, *30*, 1816–1829. [[CrossRef](#)] [[PubMed](#)]
66. Song, J.; Liu, J.; Schnakenberg, S.L.; Ha, H.; Xing, J.; Chen, K.C. Variation in piRNA and Transposable Element Content in Strains of *Drosophila melanogaster*. *Genome Biol. Evol.* **2014**, *6*, 2786–2798. [[CrossRef](#)] [[PubMed](#)]
67. Shpiz, S.; Ryazansky, S.; Olovnikov, I.; Abramov, Y.; Kalmykova, A. Euchromatic Transposon Insertions Trigger Production of Novel Pi- and Endo-siRNAs at the Target Sites in the *Drosophila* Germline. *PLoS Genet* **2014**, *10*, e1004138. [[CrossRef](#)]
68. Brennecke, J.; Aravin, A.A.; Stark, A.; Dus, M.; Kellis, M.; Sachidanandam, R.; Hannon, G.J. Discrete Small RNA-Generating Loci as Master Regulators of Transposon Activity in *Drosophila*. *Cell* **2007**, *128*, 1089–1103. [[CrossRef](#)]
69. Goriaux, C.; Desset, S.; Renaud, Y.; Vaury, C.; Brassset, E. Transcriptional properties and splicing of the flamenco piRNA cluster. *EMBO Rep.* **2014**, *15*, 411–418. [[CrossRef](#)]
70. Duc, C.; Yoth, M.; Jensen, S.; Mouni e, N.; Bergman, C.M.; Vaury, C.; Brassset, E. Trapping a somatic endogenous retrovirus into a germline piRNA cluster immunizes the germline against further invasion. *Genome Biol.* **2019**, *20*, 127. [[CrossRef](#)]

71. Díaz-González, J.; Domínguez, A.; Albornoz, J. Genomic distribution of retrotransposons 297, 1731, copia, mdg1 and roo in the *Drosophila melanogaster* species subgroup. *Genetica* **2010**, *138*, 579–586. [[CrossRef](#)]
72. Craig, N.L. *Mobile DNA II*; ASM Press: Washington, DC, USA, 2002.



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

Chapter 3

Linearisation of the genome graph

In this chapter, I explain the process to build a set of tools for genome graph analysis. This work allowed to investigate different context of pangenomes and how genome graphs were built. Through bibliography research and after discussions with biologists, I decided to come up with the idea to linearize the genome graph while maintaining a way to trace back the coordinate of the graph for further investigation. This strategy allows to take advantages not only well-developed tools for linear reference genome but also the superiority of genome graph.

To fulfill the purpose of this chapter, during the 2nd year of my PhD, I self-taught programming in Julia and learned the optimization process with my collaborators from N2TP Technology Solutions JSC., in Vietnam. Moreover, I gained experience in maintaining packages to keep up with the improvement of Julia programming languages. I studied graph theory and practice graph problems by myself so that I have the ability to raise ideas and implement the development for my tools and packages.

The manuscript for this works is in preparation to be submitted on OUP Bioinformatics as an Original Paper before the end of August, 2022. The content in preparation at the time of submission (13th of July, 2022) is documented at the end of this chapter.

3.1 Context

Currently the 3 main common pipelines to build a pangenome graph are: `vg` [47], `minigraph` [48] and `minigraph-cactus` (<https://github.com/ComparativeGenomicsToolkit/cactus>) pipelines. On one hand, `vg` constructs the graph by incorporating variant information from a VCF file into a reference genome. On the other hand, `minigraph` and `minigraph-cactus` pipeline share the same mechanism of incrementally collapsing common genome fragments to a reference or a graph, then creating bubbles for uncollapsed parts. The difference between `minigraph` and `minigraph-cactus` is that in `minigraph-cactus`, fragments of the selected reference genome will not be collapsed against itself.

Each construction method provides different types of graph owning specific advantages and limitations. For instance, `vg` graph has the ability to store variants at very high resolution up to the SNP, however, it is not always feasible to obtain adequate quality VCF from sequencing data or presenting large SV. Alternatively, `minigraph` and `minigraph-cactus` only requires FASTA input and are quite efficient for large SV, but they need a stable reference sequence to construct the genome graph and only record variants having size above 100 bases. At the moment, there are very few efforts to consolidate the graph genome coming from the 3 pipelines. In the latest publication of the Human Pangenome Project, the researchers thus released all 3 versions of the human genome graph corresponding to 3 pipelines [49].

As genome graph construction methods is still in development, the tools working on genome graph analysis are in high demand. However, there are still many obstacles, such as a consolidate coordinate system, to work through the graph, and most of the studies consider only high-coverage sequencing data. Therefore, I develop tools and packages to satisfy the following purposes: to identify DNA fragment from one individual to another individual; to take advantages of well-developed tools in linear reference genome while maintaining the use of genome graph; to benefit from genome graph for low-coverage data.

3.2 Strategies

For genome graph constructing methods, I started with the "easiest" way to obtain a graph from minigraph, as only FASTA file is required. For the example on Asian rice genome, I used the common reference *Oryza sativa ssp Japonica* cv Nipponbare [74] as the first template, and then 12 platinum rice genomes from main subgroups of Asian rice [75] to build the graph. I add these 12 genomes in an iterative mode based on their proximity to Nipponbare through a Mash [76] analysis. The output of minigraph is a file in rGFA format containing 2 lines: an S-line of each stable segment (nodes) in the graph that can act as a coordinate system and an L-line representing the linkage (edge) among stable segments.

To answer the first purpose of finding the position of a DNA fragment from one individual to another, it is necessary to know which individual possesses which part of the graph. On previous versions of minigraph, it was not possible to directly obtain the path of each individual, therefore, I remapped the sequences of the 13 genomes against the graph to yield a mapping result in GAF format. I developed then PARROT (PAngenome gRaph Related Output Transmutator) tool (<https://github.com/nguyetdang/PARROT>) to handle this question among others. PARROT will read the mapping line of GAF output from each genome and count the number of copy for each stable segments. The output of PARROT is a matrix indicating presence/absence status of each stable segments in each individual in the genome graph (see Figure 3.1)

GAF format when remap reference genomes against obtained graph

```
Os128077RS1_Ctg40      50227  44305  47115  +      >s104335>s104336>s104337>s104339>s104340
Os128077RS1_Ctg40      50227  1103   6038   +      >s147586>s21792>s21793>s164408
```

Presence/absence matrix

SegName	Start	End	SegID	AzucenaRS1	IRGSP-1	Os117425RS1	Os125619RS1	Os125827RS1	Os127518RS1
IRGSP-1.0_Chr1	0	27974	>s1	1	1	1	1	3	1
IRGSP-1.0_Chr1	27974	37449	>s2	1	1	1	1	2	1
IRGSP-1.0_Chr1	37449	37505	>s3	1	1	1	1	2	1
IRGSP-1.0_Chr1	37505	44126	>s4	1	1	1	1	2	1
IRGSP-1.0_Chr1	44126	44282	>s5	1	1	1	1	2	1
IRGSP-1.0_Chr1	44282	54574	>s6	1	1	1	1	2	1
IRGSP-1.0_Chr1	54574	57838	>s7	1	1	1	1	1	2
IRGSP-1.0_Chr1	57838	79333	>s8	1	1	1	1	2	1
IRGSP-1.0_Chr1	79333	79362	>s9	1	1	1	1	1	1

Figure 3.1: **Input (GAF) and output (PAV) of PARROT.** The GAF file is obtained from minigraph by mapping individual genomes against the graph and the presence/absence matrix file can be obtained after processing with PARROT.

With the current output, it is thus possible to know if a stable segment is available in which individuals. In order to know the exact original position, it required to trace back the position of each stable segment in each individuals from the mapping data. This feature will be incorporated in the next updates of PARROT.

For the second purpose, to take advantages of well-developed tools for linear reference genome, my idea was to extract the longest representative path from the graph and used it as a pseudo-reference. This artificial genome contains the segments appeared in the highest number of individuals forming the graph. When I looked at the technical aspect, the input file is in rGFA format built from the whole genome sequencing of 13 individuals, each harboring 12 chromosomes plus sometimes mitochondria and chloroplast sequences. Then, there might not be one single graph inside the rGFA file. Indeed, there might be multiple forms such as lone node, cycle, simple graph or graph with internal cycle (see Figure 3.2).

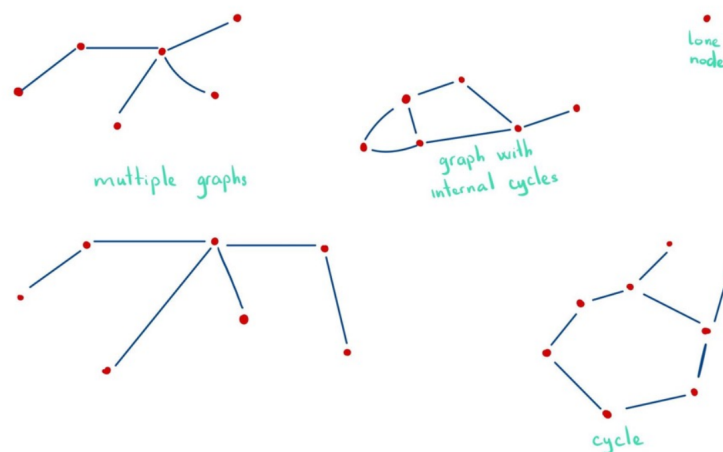


Figure 3.2: **Possible types of graph documented in rGFA format.** Each dot represents the node in the graph, equivalent to DNA fragment. Each edge depict the connection between the node. The DNA fragment within each dot can be in forward and reverse directions, hence, yield a bidirected genome graph.

I decided to separate the graphs stored in rGFA into single components, and then for each simple graph, I find the longest representative path. I performed test on the cycle inside a graph for longest path extraction purpose. However, the algorithm was not optimized and it took a lot of time to run. Hence, I decided not to optimize the cycle problem at the moment. BioGraph.jl find source and sink nodes from the graph and classify them in 3 cases: lone node (without any linkage), lone cycle (without source vertices or sink vertices) and simple graph for remaining cases.

For the longest path extraction the way I defined it, I want to extract the longest linear path containing stable segments appearing in as many individuals as possible. Therefore, based on the presence/absence matrix it obtained from the graph, PARROT also computes the weight value for each segment by counting the number of individuals for each segment and to provide an array. BioGraph.jl (<https://github.com/nguyetdang/BioGraph.jl>) will take this array and incorporate it into the graph result from rGFA: hence, we have a weighted graph object. BioGraph.jl identified the source and sink nodes in each simple graph and run through all the possibilities between a source node and a sink node, and then find out the path having the highest weight value. The detailed algorithms for component extractions and longest path identification are documented in the manuscript. It is noticeable here that PARROT only provides a array for weight values based on the number of individual owning a segment. However, BioGraph.jl accepts other arrays from user as well.

Then, I developed the tools and package them. Eventually, I applied the mentioned strategies on development and perform it on the collection of 13 Asian rice genomes.

In this current report, the results were obtained by using minigraph v.0.15. This version did not have the base calling feature added on April 20th, 2022, and was reported to produce wrong result for complex cases by Heng Li during our discussion at the International Graph Genome Conference 2022 in July and on his personal github (<https://github.com/lh3/minigraph>). Further updates of minigraph was released the following months till the current latest version is minigraph v.0.19 released on June 12th. We will rerun thus the graph construction, genome graph analysis, and longest representative linear path extraction and report the new correct results in the thesis defense and in the submitted manuscript.

3.3 Conclusion

With my colleagues, we have developed the first set of tools to linearize a pangenome at the scale of higher eukaryotes genomes. The linearized sequence can be implement as a "reference" genome in many genomic analysis: diversity, GWAS, precision medicine and so on. It contains additional information comparing to a reference genome since it can incorporate sequences of other individuals used to generate the graph.

The tools PARROT and BioGraph.jl allows to generate the longest linear path based on different criteria, the number of base pairs per node or the number of individual having the considered segment. Users can come up with other formula to set the weight value in order to extract the path that are able to answer their biological questions. The full potential of the linearisation idea nevertheless has not yet been fully discovered. In term of performance, PARROT and BioGraph.jl showed the ability to work with graph having nearly hundred thousand nodes under a short time. To be specific, PARROT generates presence/absence matrix and calculate weight value from a graph of approximately 250,000 nodes and 350,000 edges in 4 hours while BioGraph.jl can find the longest path based on the weight in less than 5 minutes on a computer of 8 core and 16 GB in memory. We will continually to work to improve the run time of PARROT.

We showed a working example of the tools on Asian rice genome and we obtain a new linear genome having approximately 100 Mbp longer than the current reference (with the minigraph erroneous version, however). It is still necessary to evaluate the use of this reference comparing to the currently used *Nipponbare* reference genome.

3.4 Perspectives

In term of development, for PARROT, we will optimize the workflow to obtain the coordinate of each stable segment in each individual. For BioGraph.jl, we will continue to tackle the problem with cycles inside the graph and expand the package to work with more input format such as vg graph.

3.5 Personal implication

I designed the whole working pipelines between minigraph, PARROT and BioGraph.jl with the advices from my supervisor Francois SABOT.

For PARROT, I designed the features and performed the development of the tools in Python. I am also responsible for maintaining the tools.

For BioGraph.jl, I also designed the features and requirements for each features. I also received some recommendation and request from Francois and his colleagues to optimize my workflow. Together with Tuan Do, my collaborator, we performed the programming and maintaining of the package on Julia. We are taking care of the github that hosting BioGraph.jl.

I performed the test on each tool/package and I ran analysis for the whole pipelines and keep it in Jupyter Notebook. With Francois, we tested different scenario to run the tools and reproduce our work on different machines.

The manuscript was written by all the 3 authors.

3.6 Scientific impacts on my PhD work

This research work provides two research tools that are able to work with genome graph in rGFA format with an example on Asian rice genome. PARROT and BioGraph.jl provides the ability to work with graph genome in different aspects: genome analysis (as in Asian rice example), or visualization with panache [77] through the PAV matrix.

Furthermore, this research work provides the research tools and materials for the next objectives of my PhD: a genome graph of high quality Asian rice genomes, a linear representative path usable for structural inference purposes.



Genome Analysis

Linear representative path extraction from genome graph and example on rice

Nguyet Dang ^{1,*}, Tuan Do ^{2, 3} and Francois Sabot ^{1,*}

¹DIADE, Univ Montpellier, CIRAD, IRD, Montpellier 34830, France, ²Faculty of Basic Science, Phenikaa University, Yen Nghia, Ha Dong, Hanoi, 11512, Vietnam. and ³N2TP Technology Solutions JSC., Hanoi, 11512, Vietnam.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Motivation text

Results: Result in a nutshell

Availability: Code, JupyterNotebook

Contact: thi-minh-nguyet.dang@ird.fr or francois.sabot@ird.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The reference genome of each species often serves as a coordination system to describe genes, variations, and other functional annotations across individuals [1]. The advancement of sequencing technologies in the last two decades continually helps to fill the gaps and correcting errors within these references over time [6]. However, these efforts are still not sufficient to represent the whole scenery of sequence diversity within a species. For instance, the current human reference genome did not include approximately 10% of the DNA found in the African pan-genome [11]; or 41.6% of trait-associated single nucleotide polymorphisms (SNPs) could not be found in the reference genome of *Oryza sativa*.

To overcome the limitation, pangenome analysis approach provides the basis to investigate the entire genomic content in a studied population by incorporating multiple genomes in its working pipeline. The concept has been widely applied in bacteria [12], eukaryote, and even complex genomes such as human [8] and plants [13, 10]. Different approaches exist to create a pangenome, but whatever one is selected, the results contain core sequences common among (almost) all individuals, and the dispensable counterparts shared by some individuals. Therefore, the complexity of a pangenome required a data format for its storage, documentation and manipulation.

The most trending idea is to represent the pangenome in a graph-based format, in which identical or similar sequences between genomes are merged into a single representative sequence, or node. This concept can be exemplified for instance by alternative loci depiction in the human reference GRCh38 [5]. However, the adaptation of this graph-based representation into research has been slow due to a lack of tools to deal with

genomic data in such format. Currently, the two most common methods to process pangenome graph data are the variation graph, personified in vg [4], odgi [3] and related tools, and the pangenome graph, with minigraph [7] as the main tool. In details, vg variation graphs are bidirected DNA sequence graphs that combine genetic variations across a population, while minigraph works with data model generated by incrementally incorporating whole genome-to-graph-alignment information. While the former can have a nucleotide resolution, the latter is much more adapted to large structural variations detection. Nevertheless, the software ecosystem to deal with such data structure, while evolving, is not comparable to the standard reference-based one, and is not implemented in an every day basis in biology labs.

In this paper, we propose an approach to extract the linear representative path, which is later usable for standard available genomics tools, from a pangenome rGFA graph built with minigraph, without losing information from the pangenome graph. We demonstrate a procedure from graph generation, from linear representative path extraction to visualization. Our implementations (<https://github.com/nguyetdang/AsianRiceGenomeGraph>) with PARROT (<https://github.com/nguyetdang/PARROT>) and *BioGraph.jl* (<https://github.com/nguyetdang/BioGraph.jl>) can achieve the whole analysis of a population consisting 12 almost gap-free reference genomes sequences of 12 subpopulations of cultivated Asian rice and the *Nipponbare* reference [6] within 12 hours.

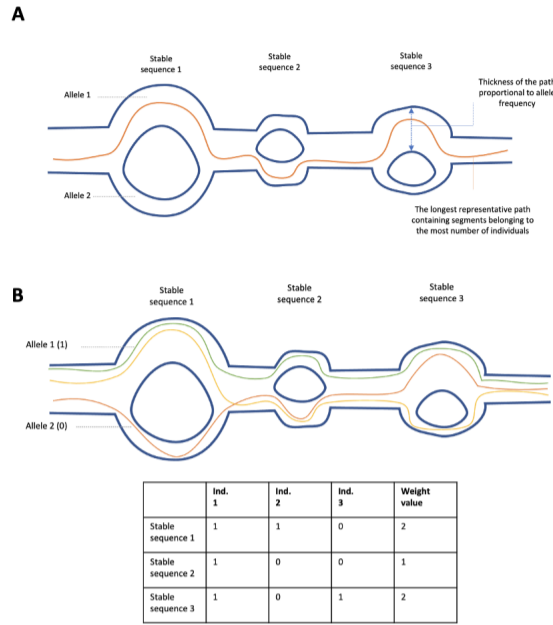


Fig. 1. A. The longest representative path (in red) is the path for which most of the individual paths exist at each node. B. A presence/absence (PAV) matrix can be extracted, provided the genotype for each individual (colored lines) for each node of the original graph.

2 Approach

2.1 Definition of the longest representative path

In the rGFA format output of minigraph, stable sequences (common regions between individuals) are defined and act as a coordinate system. We define the longest representative path in the genome graph as the path containing the stable sequences appearing in the highest number of individuals (Figure 1A). Therefore, the weight value of each node equals the number of individuals harboring the sequences. In order to obtain those values, each individual in the population is remapped against the pre-computed graph. Then, PARROT uses the mapping result as input to count the number of individuals having a considered stable segment. Eventually, a presence/absence matrix and a tabular file of the weighted value is obtained (Figure 1B). These output is required for BioGraph.jl and usable for further analyses. The whole working pipelines is depicted in Figure 2.

2.2 Graph component extraction

Considering the directed graph $rGFA$ from the rGFA input file, we define the set \mathcal{N} - set of all vertices that have outgoing edges and set \mathcal{N}' - set of all vertices that have incoming edges. Using the set operators, we can find the set \mathcal{S} - set of all source vertices (the vertices only have outgoing edges) and the set \mathcal{S}' - set of all sink vertices (the vertices only have incoming edges). We can now extract weakly connected components (a subgraph of the original graph where all vertices are connected by some path, ignoring the direction of edges) using `weakly_connected_components` function of package `Graph.jl` [2]. We classify these subgraphs into three categories based on the number of their vertices, source vertices, and sink vertices as follows, using the algorithm (1):

- Lone Node: only have one vertices
- Lone Cycle: has no source vertices or sink vertices (because sink vertices and source vertices of a cycle is the same)

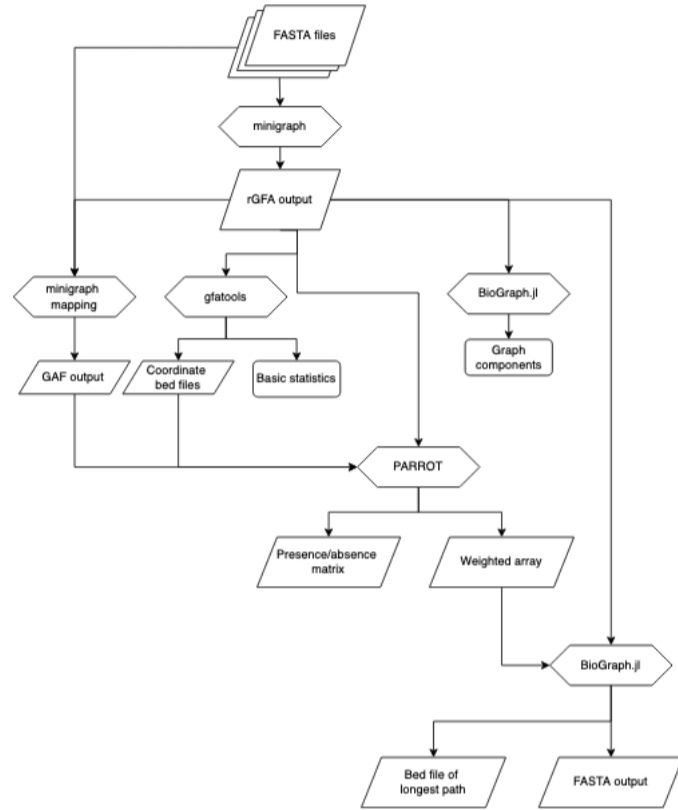


Fig. 2. Main workflow for PARROT/BioGraph.jl analysis.

- Simple Graph: others subgraph induced from GFA

Algorithm 1: Graph Components of GFA

Result: Graph Components of GFA
 init GFA input GFA
 init the $start_nodes$
 init the end_nodes
 Find all $source_nodes$
 Find all $sink_nodes$
 Find all weakly connected components g_com of rGFA
for com in g_com **do**
 if $length(com) = 1$ **then**
 Classify as Lone Node
 else
 Make sub graph induced from rGFA
 Find sub graph sg_source_nodes and sg_sink_nodes
 if $sg_source_node = []$ or $sg_sink_nodes = []$ **then**
 Classify as Lone Cycle
 else
 Classify as Simple Graph
return g_com with classification

2.3 Extract the linear representative path with BioGraph

Consider a simple directed sub-graph induced from GFA graph $G = (V, E)$ with set of vertices V and set of edges E . The number of vertices and edges are denoted by n_v and n_e , respectively. We have to find the

longest path from one vertex in set \mathcal{S}_1 - set of all source vertices to one vertex in set \mathcal{S}_2 - set of all sink vertices. As we need to consider the weight of the last node in the longest path, it is more convenient to add an end vertex to graph G , and one of these sink vertices must link to that end vertex. The new graph $G' = (V', E')$ now have $n_v + 1$ vertices and $n_e + 1$ edges. The adjacency matrix of graph G' will be:

$$\mathbf{A} = \begin{bmatrix} a_{1,1} & a_{1,2} & \dots & \dots & a_{1,n_v} & a_{1,n_v+1} \\ a_{2,1} & a_{2,2} & \dots & \dots & a_{2,n_v} & a_{2,n_v+1} \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ a_{n_v,1} & a_{n_v,2} & \dots & \dots & a_{n_v,n_v} & a_{n_v,n_v+1} \\ a_{n_v+1,1} & a_{n_v+1,2} & \dots & \dots & a_{n_v+1,n_v} & a_{n_v+1,n_v+1} \end{bmatrix}$$

And the corresponding weight matrix of graph G' :

$$\mathbf{W} = \begin{bmatrix} w_{1,1} & w_{1,2} & \dots & \dots & w_{1,n_v} & w_{1,n_v+1} \\ w_{2,1} & w_{2,2} & \dots & \dots & w_{2,n_v} & w_{2,n_v+1} \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ w_{n_v,1} & w_{n_v,2} & \dots & \dots & w_{n_v,n_v} & w_{n_v,n_v+1} \\ w_{n_v+1,1} & w_{n_v+1,2} & \dots & \dots & w_{n_v+1,n_v} & w_{n_v+1,n_v+1} \end{bmatrix}$$

where $w_{i,j}$ is equal to the weight of vertex i of graph G . We can construct a Linear Programming model of longest path problem as follow:

$$\text{find } \mathbf{A} \quad (1)$$

$$\text{maximize } \sum a_{i,j} w_{i,j} \quad (2)$$

$$\text{subject to } a_{i,j} = 0 \text{ if } a_{i,j} \notin E \quad (3)$$

$$a_{i,j} = 0 \text{ or } 1 \text{ if } a_{i,j} \in E \quad (4)$$

$$\sum_{i \in \mathcal{S}_1} \sum_{j \notin \mathcal{S}_1} a_{i,j} = 1 \quad (5)$$

$$\sum_{i \notin \mathcal{S}_2} \sum_{j \in \mathcal{S}_2} a_{i,j} = 1 \quad (6)$$

$$\sum_{i \in \mathcal{S}_2} a_{i,n_v+1} = 1 \quad (7)$$

$$\sum_{i \notin \mathcal{S}_1} a_{i,j} \leq 1 \text{ for all } j \notin \mathcal{S}_1 \cup n_{v+1} \quad (8)$$

$$\sum_{i \notin \mathcal{S}_1} a_{i,j} - \sum_{k \notin \mathcal{S}_1} a_{j,k} = 0 \text{ for all } j \notin \mathcal{S}_1 \cup n_{v+1} \quad (9)$$

As in the LP model above, the decision variable needs to binary with the following constraints:

- Equations 3 and 4 state that the edges in the longest path must exist in the edge of graph G
- Equation 5 states that only one edge is outgoing from the set of source vertices \mathcal{S}_1
- Equations 6 and 7 state that only one edge is incoming to the set of sink vertices \mathcal{S}_2
- Equation 8 states that there is at most one incoming edge for all vertex $j \notin \mathcal{S}_1 \cup n_{v+1}$
- Equation 9 states that if there is an incoming edge, there would be an outgoing edge for all vertex $j \notin \mathcal{S}_1 \cup n_{v+1}$

In most cases, solving this LP model will give an optimal solution for the longest path problem. However, additional feasible solutions for complex graphs will consist of one path and some other cycles. To eliminate these cycles, the sum of the edges in a cycle of length n must be at most $n - 1$. That means we need to add the following constraints to the LP model:

$$\sum_{i \in C} \sum_{j \in C} a_{i,j} \leq \text{length}(C) - 1 \text{ for all } C \in \mathcal{C} \quad (10)$$

where \mathcal{C} is set of all cycles of G . The LP model now has only simple paths as a feasible solution, and the optimal solution will be the longest path. However, finding all cycles in a graph is also an NP-hard problem. That means, with the growth of graph complexity, finding all cycles in a graph becomes more time and computing-intensive. For faster and more reliable implementation, we use the iterative method as in algorithm (2).

Algorithm 2: Iterative Method for Longest Path Problem

Result: Optimal solution for Longest Path Problem

init *has_cycle* = true;

init *cycle_constraints* = [];

init LP problem;

while *has_cycle* **do**

 Solve LP problem;

 Make sub-graph G_{opt} of G from LP optimal solution ;

 Find all cycles of G_{opt} *all_cycles* ;

if *length(all_cycles)* > 0 **then**

for *cycle* in *all_cycles* **do**

 Add constraint 10 of *cycle* to *cycle_constraints*;

 reinit LP problem;

else

has_cycle = false;

return G_{opt} ;

 Making optimal solution from G_{opt} ;

Because the sub-graph G_{opt} in each iteration has only one simple path and cycles, finding the all cycles of G_{opt} can be done easily using *simplecycles* function of package *Graph.jl* [2]. The constraints defined in 10 will be larger at the end of each iteration, allowing the LP model to give an optimal solution in a reasonable time.

3 Materials and Methods

3.1 Data

We take advantage of the availability of the 12 platinum genomes corresponding to the 12 major subgroups of the Asian cultivated rice *Oryza sativa* assembled at the chromosome-scale [13]. We also obtain the sequence of *Oryza sativa Nipponbare* reference genome [6] to use as the initial genome of the graph. The details of each sample can be retrieved at Supplementary Data 1.

3.2 Generation of rice genomes graph

After computing the genetic distance between each genomes using MASH v2.3 [9] under standard conditions, we implemented the rice pangenome graph with minigraph v0.15 [7] with as base the Nipponbare genome, then adding genomes based on the proximity of mash distance (Supplementary Data 2).

3.3 Identification of variation within graph

Structure variants in the genome graph was extracted by feature *bubble* from minigraph. This feature provide the position and FASTA sequences of segments presence in bubbles inside the graph.

3.4 Graph components identification

All graph basic statistics and components from the graph were obtained using the GFAtools suite 0.5 [7]. Lone nodes, lone cycles and simple graph within the rGFA file were classified using *find_graph_component* function in BioGraph.jl. Further information about the graph and its component such as source nodes, sink nodes can be obtained by method *get_summary*.

3.5 Longest path extraction

To calculate the weight value for the extraction of longest linear representative path, the genomes of 12 Asian rice subgroups were re-map against the constructed genome graph. In the same time, a BED file containing coordinate of each stable segment in the graph was built by gfatools. The mapping result in GAF format and the BED file having the coordinate is used by PARROT to calculate the presence/absence matrix. Then, PARROT count the number of individuals having each segment as the corresponding weight value and write an tabular output.

BioGraph.jl receives the weight value in tabular output and the rGFA file as input to extract the longest representative linear path. It returned a FASTA file, a BED file having the coordinate of each stable segment in the FASTA sequence and an array showing how each sequence was connected and its direction. The linear representative graph and the presence/absence matrix are the input for PANACHE [14] to visualize the pangenome in an interactive mode.

3.6 Data availability

The whole approach is available through a Jupyter book at <https://github.com/nguyetdang/AsianRiceGenomeGraph>.

4 Results and Discussion

4.1 Analysis of the rice genome graph

A genome graph built on the *Oryza sativa Nipponbare* reference and 12 platinum genomes of major subgroups in Asian rice was obtained in rGFA format. In this graph, there are 252540 nodes, 357504 edges and 715008 arcs present in the graph. The total segment length is reported to be 549 Mbp with an average segment length of 2173.833 bp. The max degree of the graph is 10, indicating that there is at least a position that unique in for each of 10 individuals or there is at least a locus having 10 alleles. The average degree is 1.416. BioGraph.jl indicates that there are 509 simple graphs and 45 lone nodes. These 45 lone nodes are sequences appeared in only a single individual of the study population.

4.2 Analysis of the longest path

To extract the longest linear representative path, we define the weight value for each segment/node as the number of individuals contain this segments. BioGraph.jl then use these weighted value to calculate the longest path for each of the 509 simple graphs classified in previous part. The final FASTA file obtained for the longest path containing 413 Mbp, which is longer than the *Nipponbare* reference of 362 Mbp.

To be specific, PARROT generates presence/absence matrix and calculate weight value from a graph of approximately 250,000 nodes and 350,000 edges in 4 hours while BioGraph.jl can find the longest path based on the weight in less than 5 minutes on a computer of 8 core and 16 GB in memory. Compare to other tool, gfatk (<https://github.com/tolkitt/gfatk>), a plant organellar graphical fragment assembly toolkit, has an option to



Fig. 3. Visualization of pangenome of Asian rice based on the most representative path - Example of 2 individuals

linearize mitochondria genome. However, it is indicated by the author that the tool is optimized only for small genome (up to 2 Mbp in size). In case of BioGraph.jl, it is possible to work with genome graphs of thousands nodes and longer genome size. The representative linear path extracted by BioGraph.jl and the presence/absence matrix built by PARROT then can be used as input for the pangenome visualization with PANACHE [14] as seen in Figure 3.

4.3 Limitations

At the moment, BioGraph.jl only works with rGFA output, which mean the tools only consider the S and L lines. BioGraph.jl might be suitable for genome graph documented in GFA1 format. In order to work with GFA2, it requires further development. In addition, Biograph.jl does not provide optimal solutions while working with internal cycles inside the graph. This issue will be addressed in the next issue of BioGraph.jl. In case of PARROT, since the output was written line by line, it is necessary to improve the run time of the tool.

5 Conclusion

In this paper, we described the first tool to our knowledge able to linearize a pangenome at the scale of higher eukaryotes genomes, based on either the longest or the most prevalent path. Such linearized pangenome sequence can then be implemented as a "reference" genome in many genomic analyses: diversity, GWAS, precision medicine, and so on. While our algorithm is really efficient, we still has minor issues with bubbles introduced in the graph creation. Our future development will tend to optimize such resolution, as well as to optimize the computational time and the data type input (in particular other graph formats such as from vg graph).

Acknowledgements

Authors acknowledge the ISO 9001 certified IRD itrop HPC (member of the South Green Platform) at IRD Montpellier for providing HPC resources that have contributed to the research results reported within this paper. URLs: <https://bioinfo.ird.fr/> and <http://www.southgreen.fr>.

Funding

ND is supported by a France Excellence PhD Grant.

Conflicts of interests

The authors declare to have no conflict of interest.

References

- [1] J. Craig Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, J. D. Gocayne, P. Amanatides, R. M. Ballew, D. H. Huson, J. R. Wortman, Q. Zhang, C. D. Kodira, X. H. Zheng, L. Chen, M. Skupski, G. Subramanian, P. D. Thomas, J. Zhang, G. L. Gabor Miklos, C. Nelson, S. Broder, A. G. Clark, J. Nadeau, V. A. McKusick, N. Zinder, A. J. Levine, R. J. Roberts, M. Simon, C. Slayman, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. Deng, V. di Francesco, P. Dunn, K. Eilbeck, C. Evangelista, A. E. Gabrielian, W. Gan, W. Ge, F. Gong, Z. Gu, P. Guan, T. J. Heiman, M. E. Higgins, R. R. Ji, Z. Ke, K. A. Ketchum, Z. Lai, Y. Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G. V. Merkulov, N. Milshina, H. M. Moore, A. K. Naik, V. A. Narayan, B. Neelam, D. Nusskern, D. B. Rusch, S. Salzberg, W. Shao, B. Shue, J. Sun, Z. Yuan Wang, A. Wang, X. Wang, J. Wang, M. H. Wei, R. Wides, C. Xiao, C. Yan, A. Yao, J. Ye, M. Zhan, W. Zhang, H. Zhang, Q. Zhao, L. Zheng, F. Zhong, W. Zhong, S. C. Zhu, S. Zhao, D. Gilbert, S. Baumhueter, G. Spier, C. Carter, A. Cravchik, T. Woodage, F. Ali, H. An, A. Awe, D. Baldwin, H. Baden, M. Barnstead, I. Barrow, K. Beeson, D. Busam, A. Carver, A. Center, M. Lai Cheng, L. Curry, S. Danaher, L. Davenport, R. Desilets, S. Dietz, K. Dodson, L. Doup, S. Ferreira, N. Garg, A. Gluecksmann, B. Hart, J. Haynes, C. Haynes, C. Heiner, S. Hladun, D. Hostin, J. Houck, T. Howland, C. Ibegwam, J. Johnson, F. Kalush, L. Kline, S. Koduru, A. Love, F. Mann, D. May, S. McCawley, T. McIntosh, I. McMullen, M. Moy, L. Moy, B. Murphy, K. Nelson, C. Pfannkoch, E. Pratts, V. Puri, H. Qureshi, M. Reardon, R. Rodriguez, Y. H. Rogers, D. Romblad, B. Ruhfel, R. Scott, C. Sitter, M. Smallwood, E. Stewart, R. Strong, E. Suh, R. Thomas, N. Ni Tint, S. Tse, C. Vech, G. Wang, J. Wetter, S. Williams, M. Williams, M. Windsor, E. Winn-Deen, K. Wolfe, J. Zaveri, K. Zaveri, J. F. Abril, R. Guigo, M. J. Campbell, K. V. Sjolander, B. Karlak, A. Kejariwal, H. Mi, B. Lazareva, T. Hatton, A. Narechania, K. Diemer, A. Muruganujan, N. Guo, S. Sato, V. Bafna, S. Istrail, R. Lippert, R. Schwartz, B. Walenz, S. Yooseph, D. Allen, A. Basu, J. Baxendale, L. Blick, M. Caminha, J. Carnes-Stine, P. Caulk, Y. H. Chiang, M. Coyne, C. Dahlke, A. Deslattes Mays, M. Dombroski, M. Donnelly, D. Ely, S. Esparham, C. Fosler, H. Gire, S. Glanowski, K. Glasser, A. Glodek, M. Gorokhov, K. Graham, B. Gropman, M. Harris, J. Heil, S. Henderson, J. Hoover, D. Jennings, C. Jordan, J. Jordan, J. Kasha, L. Kagan, C. Kraft, A. Levitsky, M. Lewis, X. Liu, J. Lopez, D. Ma, W. Majoros, J. McDaniel, S. Murphy, M. Newman, T. Nguyen, N. Nguyen, M. Nodell, S. Pan, J. Peck, M. Peterson, W. Rowe, R. Sanders, J. Scott, M. Simpson, T. Smith, A. Sprague, T. Stockwell, R. Turner, E. Venter, M. Wang, M. Wen, D. Wu, M. Wu, A. Xia, A. Zandieh, and X. Zhu. The sequence of the human genome. *Science*, 2001.
- [2] J. Fairbanks, M. Besançon, S. Simon, J. Hoffiman, N. Eubank, and S. Karpinski. `Juliagraphs/graphs.jl`: an optimized graphs package for the julia programming language, 2021.
- [3] A. Guarracino, S. Heumos, S. Nahnsen, P. Prins, and E. Garrison. `Ogdi`: understanding pangenome graphs. *Bioinformatics*, 2022.
- [4] G. Hickey, D. Heller, J. Monlong, J. A. Sibbesen, J. Sirén, J. Eizenga, E. T. Dawson, E. Garrison, A. M. Novak, and B. Paten. Genotyping structural variants in pangenome graphs using the `vg` toolkit. *Genome Biology*, 21:1–17, 2 2020.
- [5] M. Jäger, M. Schubach, T. Zemojtel, K. Reinert, D. M. Church, and P. N. Robinson. Alternate-locus aware variant calling in whole genome sequencing. *Genome Medicine*, 8, 2016.
- [6] Y. Kawahara, M. de la Bastide, J. P. Hamilton, H. Kanamori, W. R. McCombie, S. Ouyang, D. C. Schwartz, T. Tanaka, J. Wu, S. Zhou, K. L. Childs, R. M. Davidson, H. Lin, L. Quesada-Ocampo, B. Vaillancourt, H. Sakai, S. S. Lee, J. Kim, H. Numa, T. Itoh, C. R. Buell, and T. Matsumoto. Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data. *Rice*, 6(1):1–10, 2013.
- [7] H. Li, X. Feng, and C. Chu. The design and construction of reference pangenome graphs with minigraph. *Genome Biology*, 21:1–19, 12 2020.
- [8] W.-W. Liao, M. Asri, J. Ebler, D. Doerr, M. Haukness, G. Hickey, S. Lu, J. K. Lucas, J. Monlong, H. J. Abel, S. Buonaiuto, X. H. Chang, H. Cheng, J. Chu, V. Colonna, J. M. Eizenga, X. Feng, C. Fischer, R. S. Fulton, S. Garg, C. Groza, A. Guarracino, W. T. Harvey, S. Heumos, K. Howe, M. Jain, T.-Y. Lu, C. Markello, F. J. Martin, M. W. Mitchell, K. M. Munson, M. N. Mwaniki, A. M. Novak, H. E. Olsen, T. Pesout, D. Porubsky, P. Prins, J. A. Sibbesen, C. Tomlinson, F. Villani, M. R. Vollger, H. P. R. Consortium, G. Bourque, M. J. Chaisson, P. Flicek, A. M. Phillippy, J. M. Zook, E. E. Eichler, D. Haussler, E. D. Jarvis, K. H. Miga, T. Wang, E. Garrison, T. Marschall, I. Hall, H. Li, and B. Paten. A draft human pangenome reference. *bioRxiv*, page 2022.07.09.499321, 7 2022.
- [9] B. D. Ondov, T. J. Treangen, P. Melsted, A. B. Mallonee, N. H. Bergman, S. Koren, and A. M. Phillippy. Mash: Fast genome and metagenome distance estimation using minhash. *Genome Biology*, 17, 2016.
- [10] H. Rijzaani, P. E. Bayer, M. Rouard, J. Doležel, J. Batley, and D. Edwards. The pangenome of banana highlights differences between genera and genomes. *Plant Genome*, 15, 2022.
- [11] R. M. Sherman, J. Forman, V. Antonescu, D. Puiu, M. Daya, N. Rafaels, M. P. Boorgula, S. Chavan, C. Vergara, V. E. Ortega, A. M. Levin, C. Eng, M. Yazdanbakhsh, J. G. Wilson, J. Marrugo, L. A. Lange, L. K. Williams, H. Watson, L. B. Ware, C. O. Olopade, O. Olopade, R. R. Oliveira, C. Ober, D. L. Nicolae, D. A. Meyers, A. Mayorga, J. Knight-Madden, T. Hartert, N. N. Hansel, M. G. Foreman, J. G. Ford, M. U. Faruque, G. M. Dunston, L. Caraballo, E. G. Burchard, E. R. Bleecker, M. I. Araujo, E. F. Herrera-Paz, M. Campbell, C. Foster, M. A. Taub, T. H. Beaty, I. Ruczinski, R. A. Mathias, K. C. Barnes, and S. L. Salzberg. Assembly of a pan-genome from deep sequencing of 910 humans of African descent. *Nature Genetics*, 51(1):30–35, 2019.
- [12] H. Tettelin, V. Massignani, M. J. Cieslewicz, C. Donati, D. Medini, N. L. Ward, S. V. Angiuoli, J. Crabtree, A. L. Jones, A. S. Durkin, R. T. DeBoy, T. M. Davidsen, M. Mora, M. Scarselli, I. M. Y. Ros, J. D. Peterson, C. R. Hauser, J. P. Sundaram, W. C. Nelson, R. Madupu, L. M. Brinkac, R. J. Dodson, M. J. Rosovitz, S. A. Sullivan, S. C. Daugherty, D. H. Haft, J. Selengut, M. L. Gwinn, L. Zhou, N. Zafar, H. Khouri, D. Radune, G. Dimitrov, K. Watkins, K. J. O'Connor, S. Smith, T. R. Utterback, O. White, C. E. Rubens, G. Grandi, L. C. Madoff, D. L. Kasper, J. L. Telford, M. R. Wessels, R. Rappuoli, and C. M. Fraser. Genome analysis of multiple pathogenic isolates of streptococcus agalactiae: Implications for the microbial "pan-genome". *Proceedings of the National Academy of Sciences of the United States of America*, 102:13950–13955, 2005.
- [13] Y. Zhou, D. Chebotarov, D. Kudrna, V. Llaca, S. Lee, S. Rajasekar, N. Mohammed, N. Al-Bader, C. Sobel-Sorenson, P. Parakkal, L. J. Arbelaez, N. Franco, N. Alexandrov, N. R. S. Hamilton, H. Leung, R. Mauleon, M. Lorieux, A. Zuccolo, K. McNally, J. Zhang, and R. A. Wing. A platinum standard pan-genome resource that represents the population structure of asian rice. *Scientific Data*, 7, 2020.
- [14] Éloi Durant, F. Sabot, M. Conte, and M. Rouard. Panache: A web browser-based viewer for linearized pangenomes. *Bioinformatics*, 37, 2021.

Supplementary Table 1: Information of Asian rice subgroups used in the study

Variety Name	Genetic Stock ID	Country Origin	15 subpops	Assembly ID	Total sequence length	Gaps between scaffolds	Number of scaffolds	Scaffold N50	Contig N50	
CHAO MEO::IRGC 80273-1	IRGC 132278	Lao PDR	GJ-subtrp	GCA_009831315.1	383,243,376		43	124	11,025,322	11,025,322
Azucena	I1A41685	Philippines	GJ-trop1	GCA_009830595.1	381,570,127		16	53	22,940,949	22,940,949
KETAN NANGKA::IRGC 19961-2	IRGC 128077	Indonesia	GJ-trop2	GCA_009831275.1	382,007,384		9	34	22,679,302	22,679,302
ARC 10497::IRGC 12485-1	IRGC 117425	India	cB	GCA_009831255.1	380,354,188		28	69	17,921,520	17,921,520
IR 64	I1A42114	Philippines	XI-1B1	GCA_009914875.1	387,483,803		91	112	7,352,909	7,352,909
PR 106::IRGC 53418-1	IRGC 127742	India	XI-1B2	GCA_009831045.1	394,601,903		4	44	26,007,178	26,007,178
LIMA::IRGC 81487-1	IRGC 127564	Indonesia	XI-3A	GCA_009829395.1	405,399,143		5	98	27,206,337	27,206,337
KHAO YAI GUANG::IRGC 65972-1	IRGC 127518	Thailand	XI-3B1	GCA_009831295.1	399,249,348		7	53	21,823,919	21,823,919
GOBOL SAIL (BALAM)::IRGC 26624-2	IRGC 132424	Bangladesh	XI-2A	GCA_009831025.1	392,847,014		3	28	29,604,901	29,604,901
LIU XU::IRGC 109232-1	IRGC 125827	China	XI-3B2	GCA_009829375.1	490,154,521		5	580	29,704,550	29,704,550
LARHA MUGAD::IRGC 52339-1	IRGC 125619	India	XI-2B	GCA_009831355.1	391,869,645		4	46	30,747,645	30,747,645
NATEL BORO::IRGC 34749-1	IRGC 127652	Bangladesh	cA2	GCA_009831335.1	384,203,823		4	23	27,825,079	27,825,079

Supplementary Table 2: Mash distance among 12 Asian rice subgroups and *Nipponbare* reference genome

Reference ID	Query ID	Mash-distance	P-value	Matching-hash
IRGSP-1.0.fa	IRGSP-1.0.fa	0	0	1000/1000
Os132278RS1.fa	IRGSP-1.0.fa	0.00550978	0	803/1000
AzucenaRS1.fa	IRGSP-1.0.fa	0.00580821	0	794/1000
Os128077RS1.fa	IRGSP-1.0.fa	0.00607761	0	786/1000
Os117425RS1.fa	IRGSP-1.0.fa	0.00804257	0	731/1000
Os127742RS1.fa	IRGSP-1.0.fa	0.0118927	0	638/1000
Os127518RS1.fa	IRGSP-1.0.fa	0.0126347	0	622/1000
OsIR64RS1.fa	IRGSP-1.0.fa	0.0126347	0	622/1000
Os125619RS1.fa	IRGSP-1.0.fa	0.0127767	0	619/1000
Os132424RS1.fa	IRGSP-1.0.fa	0.0128243	0	618/1000
Os127564RS1.fa	IRGSP-1.0.fa	0.0129198	0	616/1000
Os127652RS1.fa	IRGSP-1.0.fa	0.0133059	0	608/1000
Os125827RS1.fa	IRGSP-1.0.fa	0.0134525	0	605/1000

Chapter 4

Structural variation inference of skimming data

4.1 Context

Along the last 10 years, a lot of sequencing data were generated for multiples individuals within the same organisms, at very various scale in term of samples as well as for depth per sample. While medium and high depth data can be used for partial SV detection, skimming data (i.e. with a mean coverage generally lower than 5x) are generally used only for SNP discovery. All of these analyses are single-reference-based, and thus did not incorporate the pangenomics information. However, the possibility offered by BioGraph.jl and PARROT to work with a linearized pangenome instead of graph opens the door to a new way to infer the structural variations and the PAV from any type of data, and in particular the skimming ones.

In this regard, I was implicated in the analysis of such a dataset for allele discovery and GWAS before my PhD, in a collaboration between IRD and AGI [78]. This dataset of 178 individuals sequenced between 2 to 8x came from a collection gathered between 2016 and 2018 on the Vietnamese rices, and each of them has been phenotyped for various traits. Previous analyses of read mapping and SNP extraction were performed for GWAS analyses [79, 80], but the resolution was low, and such analysis did not take into account the SV. We originally tried to infer more SNPs in missing data through the use of BEAGLE ([81]), but found out that BEAGLE inferred SNP for some samples in regions that are not present in them! Indeed, we have control sequences at high depth (more than 35x) for some samples, and were able to identify the deleted regions in which BEAGLE inferred SNPs.

Thus, we came out with the need of a dedicated tool for inference of SV on skimming data, based on the pangenome structure, that we called GraphInfer.

4.2 Strategies

GraphInfer is designed to infer structure of individuals having skimming sequencing data. The inference depends on a reference genome graph constructed from a population of individuals of the same species to predict the structure of the studied individuals. The longest linear representative path of the genome graph will be extracted so that skimming reads can be mapped and partial structure of the studied individuals can be obtained. To fill the gap, GraphInfer based on a scoring system to predict which path on the graph the studied individual should take. At the moment, the default scoring system defines that the output structure is the one containing the highest number of segment mapped by skimming reads. GraphInfer top up the weight value for those segments with a value equal

to the number of nodes in the genome graph in order to assure these segments appearing in the output structure. The weight value of other segments remain the same as in the input graph. The algorithm to extract the longest path of BioGraph.jl were used here to infer the structure.

4.3 Conclusion

However, GraphInfer still not reach the nucleotide resolution for SV/SNP that could be required for functional analysis, because of the double limit of the graph resolution as well as the data type. In addition, GraphInfer is limited to the variations already present in the graph and thus cannot identify new variants, at the opposite of tools such as FrangiPAN [42] for who works with high depth data.

4.4 Perspectives

As a main perspective, we will work in the near future on the optimization of the score of PAV, in order to be able to propose a reliable tool. This update will be available for the submission of the manuscript, scheduled for the end of November.

4.5 Personal implication

There are 3 authors working together to develop GraphInfer. François and I come up with the idea of structure inference for skimming data based on genome graph and we optimize the process by extracting the longest path.

I designed and develop the features of GraphInfer in Python with the help of Tuan Do for the reading input features and intergration of Julia package BioGraph into the tools.

I prepared the test data and I ran the analysis for validation of the tools. I documented my analysis in Jupyter Notebook for distribution purpose.

The manuscript was written by all the 3 authors.

4.6 Scientific impacts on my PhD work

GraphInfer can be seen as the tool able to answer mostly to my PhD question. This tool can reconstruct with quite high accuracy the global genome structure of any sample sequenced at low depth. Thus, any can found in a skimming data collection just with a single mapping analysis which of her/his samples harbor which already identified gene, and so on.

It has limits, intrinsic and extrinsic, but proposes for the first time to reconstruct the global structure of any sample sequenced at low coverage as soon as a representative pangenome graph for this species is available.

Genome Analysis

GraphInfer: Structural inference for skimming data based on genome graph and its linear representative path

Nguyet Dang^{1,*}, Tuan Do^{2,3} and Francois Sabot^{1,*}

¹DIADE, Univ Montpellier, CIRAD, IRD, Montpellier 34830, France, ²Faculty of Basic Science, Phenikaa University, Yen Nghia, Ha Dong, Hanoi, 11512, Vietnam. and ³N2TP Technology Solutions JSC., Hanoi, 11512, Vietnam.

* To whom correspondence should be addressed.

Associate Editor: XXXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Genome graph proposes a new approach to represent diversity of a studied group or a species since it possesses the ability to store most of genomic variations. While multiple methods are developed for structural variation analysis on the graph, there is still a lack of tools implemented for low-coverage sequencing data due to their intrinsic noise and uncertainty characteristics.

Results: We introduce here GraphInfer, a tool for inference of genome structure on skimming data using a genome graph and its representative linear path. The tool takes advantages of well-developed approaches for structural variation inference of low-coverage on linear reference genome while still maintaining the advantages of genome graph.

Availability: GraphInfer can be accessed at <https://github.com/nguyetdang/GraphInfer> under the MIT License.

Contact: thi-minh-nguyet.dang@ird.fr, francois.sabot@ird.fr

Supplementary information: Examples on how to work with GraphInfer are available at <https://github.com/nguyetdang/GraphInferExample>

1 Introduction

The widespread of massive reduction in the cost of DNA sequencing and the advancements in sequencing technologies promise an increasing understanding of population genomics studies. However, before implementing any research, it is important to make decision how the resources are used among different aspects such as: breadth of the coverage, depth of coverage and number of sequenced samples. In the coming years, low-coverage sequencing (skimming data) has been shown a cost-effective strategy to capture SNP variations across the entire genome at population scale, from which the result can be meaningful in further studies: population structure investigation [17], conservation biology [5] or even ancient DNA [1]. However, due to the low depth nature, genomic structural inference from skimming data is not reliable and normally required specific strategies and tools to be determined, especially for genomic regions containing big structural variants [14].

In most cases, structural inference from skimming data requires a reference genome as a template to map short-read sequence data from each individual. Otherwise, short-read data might be aligned against the reference genome of a closely related species [13]. Since the reference genome is often built from high-quality sequencing of a single individual or from consensus sequences of individuals in a species, it lacks the ability to depict alternative alleles and genome structures, and mismatched sequencing reads might be discarded from post-mapping analysis [12]. This reference bias vastly affect the structure inference of individuals having low-coverage sequencing reads due to reducing number of reads.

Genome graph is a powerful representation to overcome the reference genome bias since it possesses the ability to store not only individuals genomic profiles in the graph [11], but also all the structural variants among individuals [6]. Besides graph genome construction, a lot of efforts are put into structural variant genotyper based on graph resulting in tools. Depending on the type of graph used, there are corresponding tools for structural variant inference purposes, for example, Paragraph for sequence graph [2], GraphTyper for directed acyclic graph [3, 4], and vg toolkit for variation graph [6]. However, these tools and approaches are mostly

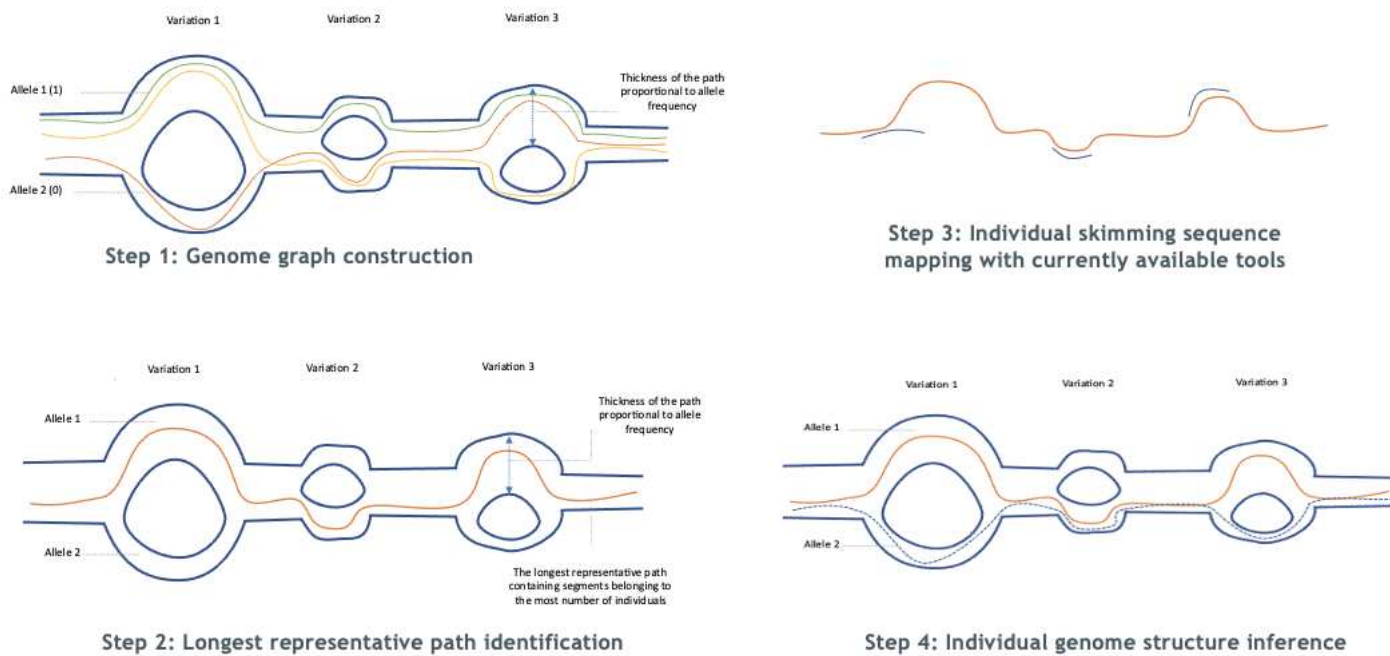


Fig. 1. Four steps to infer structure for skimming data

dedicated for individuals having medium or high coverage sequencing reads. Moreover, the alignment of sequencing data against genome graph is not optimized, especially for the cases of complex regions owning multiple alleles [8, 16, 11], which can affect inference results.

Here we introduce GraphInfer, a tool for inferring genomic structure of low-coverage sequencing individuals based on a graph genome built from high quality sequencing data. GraphInfer compares individual sequences against the linear representative path and infers the structure based on the whole genome graph and the mapping result. This strategy allows to take the advantages from both advancement for linear reference genome and genome graph.

2 GraphInfer

2.1 Features

GraphInfer is designed to infer structure of individuals having skimming sequencing data. The inference depends on a reference genome graph constructed from a population of individuals of the same species to predict the structure of the studied individuals. The longest linear representative path of the genome graph will be extracted so that skimming reads can be mapped and partial structure of the studied individuals can be obtained. To fill the gap, GraphInfer based on a scoring system to predict which path on the graph the studied individual should take. At the moment, the default scoring system defines that the output structure is the one containing the highest number of segment mapped by skimming reads. GraphInfer top up the weight value for those segments with a value equal to the number of nodes in the genome graph in order to assure these segments appearing in the output structure. The weight value of other segments remain the same as in the input graph. The algorithm to extract the longest path of BioGraph.jl were used here to infer the structure. The concept of GraphInfer is illustrated as in Figure 1.

2.2 Implementation

GraphInfer is a Python3 package that takes the following inputs: a genome graph in rGFA format, a binary matrix representing the presence/absence status of each stable DNA sequence in each individual constructing the genome graph, and a BED file containing the number of reads from the individual whose genomic structure will be inferred mapped against the longest representative path. The coordination indicated in the BED file is in accordance with the stable DNA segment in the rGFA file.

The genome graph in rGFA format can be constructed using minigraph [11]. The binary matrix and the longest path can be obtained by using PARROT (<https://github.com/nguyetdang/PARROT>) and BioGraph.jl (<https://github.com/nguyetdang/BioGraph.jl>). The BED file input can be obtained by using bedtool intersectbed [15] on the mapping result of skimming data on the linear representative path and its corresponding BED files.

GraphInfer returned 3 outputs: an array containing the predicted path of the skimming individual, a FASTA file having the prediction of the sequence and a BED file representing which stable segments are incorporated in the genomic structure of the studied individuals along with their certainty rating.

2.3 Validation

In order to validate the proposed strategy, we tested GraphInfer with the data from PARROT/BioGraph.jl on Asian rice genome (<https://github.com/nguyetdang/AsianRiceGraph>). We selected the sequencing data of *Azucena*, an individual used to construct the graph and we simulated structural variants and Illumina sequencing reads of 2x coverage from them by using ART [7].

Initially, we used bwa [10] to mapped simulated reads against *Nippobare* reference genome and then the longest path. However, mapping with bwa return a mapped percentage of 0% for both case.

When using minimap2 [9], mapping the simulated reads against the reference genome only gives a mapped percentage of 98.34% in which 97.56% are proper match. In case of using the longest path, we obtain the

percentage of read mapped and proper mapped of 98.77% and 98.06%, correspondingly, showing the importance of using a pangenome sequence in order to optimize mapping.

Through GraphInfer, we have found unmapped *Azucena* reads against the *Nipponbare* reference but that can be mapped in both longest path and *Azucena* in the graph. For instance, the read `ÀzucenaRS1_Chr1-585206ĉan` can be detected in the `longest_path_1` extracted from the graph, but is unmapped using the *Nipponbare* reference genome. It belongs to the stable sequence `s163337` indexed in the graph and belongs to *Azucena* but not to *Nipponbare*. This example illustrates the capacity of GraphInfer to identify and infer the global structural variation of samples sequences at low coverage.

NOTE: Further validation results will be updated after re-running the graph construction with the last new version of minigraph [11] that contains major changes.

3 Discussion

GraphInfer is the first tool to our knowledge that works on skimming data to infer the genome structure. GraphInfer offers an option to use BioGraph.jl based on mapping data beside the longest representative path. At the moment, GraphInfer is biased by the graph genome construction and completion, since it does not infer structures not included in the graph. In addition, the tool contains so far only one simple scoring system, which is top up the value of segments having mapped reads. Furthermore, based on the reported result, it is recommended to use minimap2 and the longest path for GraphInfer to obtain better structure inference.

For the future development, we will enhance the scoring system of GraphInfer so that it can integrate other parameters such as mapping quality, mapping score or customizable parameters provided by users.

Acknowledgement

Authors acknowledge the ISO 9001 certified IRD itrop HPC (member of the South Green Platform) at IRD Montpellier for providing HPC resources that have contributed to the research results reported within this paper. URLs: <https://bioinfo.ird.fr/> and <http://www.southgreen.fr>.

Funding

ND is supported by a France Excellence PhD Grant.

Conflicts of interest

The authors declare to have no conflict of interest.

References

- [1] C. E. G. Amorim, S. Vai, C. Posth, A. Modi, I. Koncz, S. Hakenbeck, M. C. L. Rocca, B. Mende, D. Bobo, W. Pohl, L. P. Baricco, E. Bedini, P. Francalacci, C. Giostra, T. Vida, D. Winger, U. von Freeden, S. Ghirotto, M. Lari, G. Barbujani, J. Krause, D. Caramelli, P. J. Geary, and K. R. Veeramah. Understanding 6th-century barbarian social organization and migration through paleogenomics. *Nature Communications* 2018 9:1, 9:1–11, 9 2018.
- [2] S. Chen, P. Krusche, E. Dolzhenko, R. M. Sherman, R. Petrovski, F. Schlesinger, M. Kirsche, D. R. Bentley, M. C. Schatz, F. J. Sedlazeck, and M. A. Eberle. Paragraph: A graph-based structural variant genotyper for short-read sequence data. *Genome Biology*, 20:1–13, 12 2019.
- [3] H. P. Eggertsson, H. Jonsson, S. Kristmundsdottir, E. Hjartarson, B. Kehr, G. Masson, F. Zink, K. E. Hjorleifsson, A. Jonasdottir, A. Jonasdottir, I. Jonsdottir, D. F. Gudbjartsson, P. Melsted, K. Stefansson, and B. V. Halldorsson. Graphyper enables population-scale genotyping using pangenome graphs. *Nature Genetics* 2017 49:11, 49:1654–1660, 9 2017.
- [4] H. P. Eggertsson, S. Kristmundsdottir, D. Beyter, H. Jonsson, A. Skuladottir, M. T. Hardarson, D. F. Gudbjartsson, K. Stefansson, B. V. Halldorsson, and P. Melsted. Graphyper2 enables population-scale genotyping of structural variation using pangenome graphs. *Nature Communications* 2019 10:1, 10:1–8, 11 2019.
- [5] A. P. Fuentes-Pardo and D. E. Ruzzante. Whole-genome sequencing approaches for conservation biology: Advantages, limitations and practical recommendations. *Molecular Ecology*, 26:5369–5406, 10 2017.
- [6] G. Hickey, D. Heller, J. Monlong, J. A. Sibbesen, J. Sirén, J. Eizenga, E. T. Dawson, E. Garrison, A. M. Novak, and B. Paten. Genotyping structural variants in pangenome graphs using the vg toolkit. *Genome Biology*, 21:1–17, 2 2020.
- [7] W. Huang, L. Li, J. R. Myers, and G. T. Marth. Art: a next-generation sequencing read simulator. *Bioinformatics*, 28:593–594, 2 2012.
- [8] B. Kehr, K. Trappe, M. Holtgrewe, and K. Reinert. Genome alignment with graph data structures: A comparison. *BMC Bioinformatics*, 15:1–20, 4 2014.
- [9] H. Li. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*, 34, 2018.
- [10] H. Li and R. Durbin. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, 25, 2009.
- [11] H. Li, X. Feng, and C. Chu. The design and construction of reference pangenome graphs with minigraph. *Genome Biology*, 21:1–19, 12 2020.
- [12] H. Li and J. Wren. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics*, 30:2843–2851, 10 2014.
- [13] R. N. Lou, A. Jacobs, A. P. Wilder, and N. O. Therkildsen. A beginner’s guide to low-coverage whole genome sequencing for population genomics. *Molecular Ecology*, 30:5966–5993, 12 2021.
- [14] R. Nielsen, J. S. Paul, A. Albrechtsen, and Y. S. Song. Genotype and snp calling from next-generation sequencing data. *Nature Reviews Genetics* 2011 12:6, 12:443–451, 5 2011.
- [15] A. R. Quinlan and I. M. Hall. Bedtools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26:841–842, 3 2010.
- [16] M. Rautiainen and T. Marschall. Graphaligner: rapid and versatile sequence-to-graph alignment. *Genome biology*, 21:253, 9 2020.
- [17] N. Rustagi, A. Zhou, W. S. Watkins, E. Gedvilaite, S. Wang, N. Ramesh, D. Muzny, R. A. Gibbs, L. B. Jorde, F. Yu, and J. Xing. Extremely low-coverage whole genome sequencing in south asians captures population genomics information. *BMC Genomics*, 18:1–12, 5 2017.

Chapter 5

Discussion and Perspectives

5.1 Discussion

During my PhD, I mainly focus on developing methods for structural variants. I studied different data models including linear reference genome and genome graph.

Linear reference genome approach is widely use in many genomic research. In case of structural variants detection, linear reference genome is used as a template and a coordinate system for the alignment of individual sequencing reads. As mentioned in section 1.2, a linear reference genome is even sometimes built specifically, based on a single individual dataset or on a consensus sequences of a group of individuals. Therefore, a linear reference genome is insufficient to represent the diversity of a species. As a result, reference bias occurs in genomic analysis, especially in variant calling step. To be specific, the reference bias can cause missing alternative alleles, mistaking heterozygous site as homozygous [33] or influencing allele frequencies [34]. Another possibility for alleviating the influence of the reference bias is to compare pairwise genome assembly of reads. However, the comparison will be complex when more and more alternative genome assemblies are integrated into the study.

In our study in Chapter 3, we applied both strategies (using a reference genome and comparing assemblies of individuals) to investigate the transposable elements insertions content and its dynamics. We worked on stable strains of wild-type *D. melanogaster* and *D. simulans* to describe the general landscape of TEIs. For TE dynamics, isogenic wild-type strain *D. melanogaster* and an unstable line (having a Piwi knockdown) with a succession of transpositions busts were incorporated in the study. We developed or applied methods to detect TE for both short-reads and long-reads data. In brief, for short-reads, they were aligned against a single linear reference genome of *D. melanogaster* to identify pair-end reads in which one end mapped against the reference and another mapped to TE database. The location of these pair-end reads indicate the potential position of TEIs. Next, we assemble long-read assemblies of each individual, compare them to the reference sequence, and use it as a template to remap raw long-reads against their corresponding assembly for structural variant calling. We then classified global variant (genome to genome) and minor insertion (between haplotypes within a single genome) variants. By comparing wild-type strains of *D. melanogaster* and *D. simulans*, we found that their TEs sequences are more similar than previous expectations. However, we considered only the detection of most recently active TE families, due to the limitation of sequence-to-sequence comparison and the missing haplotypes that are not present in the linear reference. While studying TE dynamics, we compared the TE content detected by using short-reads aligned against the reference genome and long-reads aligned against their corresponding assembly. The result mentioned in [1] highlighted the essential impact of long-read data and individual

assemblies in describing TE landscapes at an intra-genome scale. Based on the result of this publication, the limitations of using only a linear reference genome were pointed out in both TE content and dynamics study. At this point, it was important for me to understand about structural variant identification methods. Through the study, I experienced to curate and analyze different types of sequencing data including both short- and long-reads. Based on the result, I understood more about the limitations coming from conventional analysis approach using only a linear reference genome. Therefore, I assertively proposed the concept of pangenome and genome graph to continue my research.

As mentioned in subsection 1.2.1., the pangenome concept was first time proposed in bacteria [2]. After that, various terms and definitions were proposed [35, 36]. In my study, I mostly worked with a pangenome defined as the total of genomic content available in a population, in which the core genome is composed of shared sequences among (almost) all individuals and the remaining part being the dispensable genome. I represented this pangenome under a graph model, in which each node contains a stable sequence of nucleotides. The linkage between two nodes is called an edge, depicting how the two sequences are joined together. As I mentioned previously in the Limitations (subsection 1.3.4), in order to work with genome graph, biologists have to choose which GFA format and genome graph construction method to work with. In my case, I picked up rGFA and GFA1 formats, and I built genome graphs using minigraph.

However, although the file format and the construction method were clarified, it is still difficult for biologists to directly work with the genome graph. The next question was then how to implement a linearisation of a pangenome graph structure for a better all-day usage by biologist. In this case, I come up with the idea of linearising the pangenome by getting a representative path. At that point, I have access to the platinum quality sequencing data from 12 varieties belonging to 12 different subgroups of Asian rice [75]. My idea in this case was to construct a genome graph containing the commonly use reference *Oryza sativa Nipponbare* [74] as a template, and then to incorporate these 12 genomes to form a genome graph. The linear path extracted from the graph will share the most common content among individuals in the graph. I developed a Julia package for that purpose, namely BioGraph.jl (<https://github.com/nguyetdang/BioGraph.jl>). To facilitate the workflow with BioGraph.jl, PARROT (<https://github.com/nguyetdang/PARROT>) was developed for input/output transformation purposes. Based on the definition of the longest representative path, I came up with a way to calculate the weight value for each node in the genome graph. Each node will obtained the value equal to the number of individuals used to construct the genome graph. The longest path is the path going from a source node to a sink node with the highest number of weight value. This calculation is integrated in PARROT for easier implementation. Actually, based on the strategy to calculate the weight value, different graph can be obtained to help to answer various biological questions. For instance, the longest path on non-weighted graph can show the diversity of the population since it goes through highly diverse positions in the graph. Or as in GraphInfer mentioned in Chapter 5, weight value can help for structure variant inference on skimming data. Furthermore, BioGraph.jl provides an option to classified the subgraphes into different categories: lone node, lone cycle, simple graph. Lone nodes were contigs unique to individuals and do not belong to any subgraph. Avoiding lone cycles optimizes the running time while extracting the longest path. By applying this working pipeline on Asian rice data, we were able to construct a genome graph and its component. We also obtained a linear representative path having approximately 50 Mbp longer than the reference *Oryza sativa Nipponbare* [74]. This linear representative path provides a better linear template to work with than a sole reference genome. Moreover,

BioGraph.jl has a better run time and can work with higher number of nodes than any other graph tools for linearisation.

The recent release of minigraph, after the main writing of my PhD, has tremendous changes such as base-alignment while constructing genome graph. This change affects vastly the number of nodes and edges inside the genome graph. Therefore, we intend to re-run and re-perform the analysis to obtain the latest result before submitting our article for publication.

At the moment, the last question to target is how to perform SV inference for low-coverage data using the advantages proposed by genome graph. As mentioned in section 1.3 about Genome Graph, one of the hot topic to address is alignment problems. And I stated in the section 1.1 about structural variant that the inference result can be influenced by alignment and variant calling methods. Therefore, while tools for alignment against the graph is still in development, we come up with the idea of linearisation of the genome graph so that we can take advantage of well developed tools for alignment in linear reference genomes.

GraphInfer based on the mapping result of skimming data against the reference. We have tested the mapping result by bwa [82] and minimap2 [83]. We shown that for skimming data, minimap2 provides better mapping result thanks to its base calling alignment algorithm. When we compare the mapping between the alignment against *Nipponbare* and the longest path extracted from the graph, we are able to identify sequences that are not mapped against the *Nipponbare* but are in both the longest path and the graph. In the example, we confirmed that the sequence that can be mapped in the longest path truly belongs to the *Azucena* genome, that we used for simulating data. Further validation of GraphInfer is still on going and we will keep update the result after rerun all analyses with the new version of the Asian rice genome graph build with minigraph with its base-calling algorithm updates.

5.2 Perspectives

Overall, my PhD work provides some solutions for objectives indicated in the Chapter 2:

- Explore structure events in a common case, transposable elements, with the limit of the sequence-based approaches
- Implement a linearisation of the pangenome graph structure for a better all-day usage by biologists
- SV inference in low-coverage data using the linear pangenome

The first task were answered with the a publication about TE content and dynamics. A research tools, TrEMOLO, to detect TE movement was developed and released. We are looking forward in the development of the features to identify populational variations (and even somatic variations). We would like to determine TEs that are outsiders or represent in only a part of the population.

The second task was answered by proposing the idea of linearize the pangenome graph to obtain a linear path that still conserves all the coordinates of the graph, so that we can use tools developed for linear reference genomes while not compromising the benefits of the graph. There is a publication in preparation for this case and two research tools have been developed: BioGraph.jl for linearisation, and PARROT to handle the input/output transformation. There still some minor limitations in the two tools: for example, dealing with cycles inside the graph with BioGraph.jl is still challenging. Furthermore, at the

moment, PARROT only provides a presence/absence matrix and does not indicate the location of stable segments on each individual genome; this issue will be addressed in the next version. Finally, we only have an automated way to calculate weighted value for genome graph, but different case study should be taken into account such as longest linear path base on information from ancient recombination graph. We also schedule for the next releases of PARROT and BioGraph.jl to work with variation graph from vg, odgi and pggb.

The last task is partially answered by the current version of GraphInfer. Even if GraphInfer has the ability to predict structure of individuals having skimming sequencing data, the model based solely on presence/absence status is still basic. For the next version, we will work on a formula to calculate the weight value based on different parameters: mapping quality, local nodes presence/absence status corresponding to each individual and so on.

After finishing my PhD, I am thinking about the current problems that are still the obstacles for my biologist colleagues to work with genome graph. One significant issues is annotation. For example, in the Asian rice genome graph, there are 13 individuals with fully annotated information. However, there is no system to transfer the annotation from one individual to another. It provokes an essential question that must be answer, for example, "Where is my gene in *Azucena* can be found in *IR64* variety?". The annotation of the genome is not just simply consolidate a system to transfer the location of one gene to another, it is necessary to identify orthologs and incorporate those information into the graph. This is the work that need the collaboration between developers and biologists.

Another problem raises from the nature of genome graph. At the moment, genome graph constructed by minigraph does not access small variations, meanwhile, the high resolution from vg graph becomes a challenge for computing time when more and more individuals are added into the studied population. Therefore, it would be nice if there is a solution can overcome both obstacles. During the time studying knowledge graph with my collaborators in Vietnam for a side project (as mentioned in Annex A), I figure out the knowledge graph can contain more information within each node. Therefore, with my supervisor, we are thinking of a super graph in which a small version of variation graph can be incorporated in a node of the genome graph. Hence, it can increase the resolution of the graph while the number of nodes remains unchanged and do not interfere the computing time. Actually, we are still in the idea stage for this development and we hope to be able to have prototype for it.

5.3 Brief on my PhD

With the background of biotechnology and doing lab work for almost my study in Bachelor and Master level, doing a PhD in Genomics and Genetics is a really fresh experience.

I have chance to improve my programming skill by experience to many languages from Python, Julia, bash for analyzing the data to LaTeX and Markdown for writing and presentation. I also had chance to get used to with tool development process. One interesting thing that I recognize is that the actual work is not finish when we release a tool, it is the start of documentation, demo, and software maintenance.

Furthermore, my biology knowledge also enhance vastly since developing methods required the discussion with other researcher to understand their needs. I was able to work with rice and *Drosophila* models and have many cool discussions with researchers working with tomato, lettuce, roses, wheat.

Personally, through my PhD, half of it was "faire à la maison" due to the COVID-19 pandemics. There is the chance when I was depressed because of the lack of interaction.

However, when everything is almost over, I appreciate that time since it helps me to be mentally stronger. And I greatly appreciate any occasion to discuss and exchange my research idea. That's why I took advantages of many discussions when I went on IGGSy 2022, the only offline congress in my PhD, to share my work and to exchange with researchers from the same domain.

I don't know if PhD period can happen once or twice (may be thrice?) but for the moment, I am grateful for what I achieve here.

Bibliography

- [1] Mourdas Mohamed et al. “A Transposon Story: From TE Content to TE Dynamic Invasion of Drosophila Genomes Using the Single-Molecule Sequencing Technology from Oxford Nanopore”. In: *Cells* 9 (8 July 2020), p. 1776. ISSN: 20734409. DOI: 10.3390/CELLS9081776. URL: <https://www.mdpi.com/2073-4409/9/8/1776/htm%20https://www.mdpi.com/2073-4409/9/8/1776>.
- [2] Hervé Tettelin et al. “Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: Implications for the microbial ”pan-genome””. In: *Proceedings of the National Academy of Sciences of the United States of America* 102 (39 2005), pp. 13950–13955. ISSN: 00278424. DOI: 10.1073/pnas.0506758102.
- [3] Michael A. Brockhurst et al. “The Ecology and Evolution of Pangenomes”. In: *Current Biology* 29 (20 Oct. 2019), R1094–R1103. ISSN: 0960-9822. DOI: 10.1016/J.CUB.2019.08.012.
- [4] Philipp E. Bayer et al. “Plant pan-genomes are the new reference”. In: *Nature Plants* 2020 6:8 6 (8 July 2020), pp. 914–920. ISSN: 2055-0278. DOI: 10.1038/s41477-020-0733-0. URL: <https://www.nature.com/articles/s41477-020-0733-0>.
- [5] Nathan M. Springer et al. “Maize Inbreds Exhibit High Levels of Copy Number Variation (CNV) and Presence/Absence Variation (PAV) in Genome Content”. In: *PLOS Genetics* 5 (11 Nov. 2009), e1000734. ISSN: 1553-7404. DOI: 10.1371/JOURNAL.PGEN.1000734. URL: <https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1000734>.
- [6] Candice N. Hirsch et al. “Insights into the Maize Pan-Genome and Pan-Transcriptome”. In: *The Plant Cell* 26 (1 Feb. 2014), pp. 121–135. ISSN: 1040-4651. DOI: 10.1105/TPC.113.119982. URL: <https://academic.oup.com/plcell/article/26/1/121/6102300>.
- [7] Juan D. Montenegro et al. “The pangenome of hexaploid bread wheat”. In: *The Plant Journal* 90 (5 June 2017), pp. 1007–1013. ISSN: 1365-313X. DOI: 10.1111/TPJ.13515. URL: <https://onlinelibrary.wiley.com/doi/full/10.1111/tpj.13515%20https://onlinelibrary.wiley.com/doi/abs/10.1111/tpj.13515%20https://onlinelibrary.wiley.com/doi/10.1111/tpj.13515>.
- [8] John W. Belmont et al. “A haplotype map of the human genome”. In: *Nature* 2005 437:7063 437 (7063 Oct. 2005), pp. 1299–1320. ISSN: 1476-4687. DOI: 10.1038/nature04226. URL: <https://www.nature.com/articles/nature04226>.
- [9] Maren Wellenreuther et al. “Going beyond SNPs: The role of structural genomic variants in adaptive evolution and species diversification”. In: *Molecular Ecology* 28 (6 Mar. 2019), pp. 1203–1209. ISSN: 1365-294X. DOI: 10.1111/MEC.15066. URL: <https://onlinelibrary.wiley.com/doi/full/10.1111/mec.15066%20https://onlinelibrary.wiley.com/doi/abs/10.1111/mec.15066%20https://onlinelibrary.wiley.com/doi/10.1111/mec.15066>.

- [10] Lars Feuk, Andrew R. Carson, and Stephen W. Scherer. “Structural variation in the human genome”. In: *Nature Reviews Genetics* 2006 7:2 7 (2 Feb. 2006), pp. 85–97. ISSN: 1471-0064. DOI: 10 . 1038 / nrg1767. URL: <https://www.nature.com/articles/nrg1767>.
- [11] Jun Zhou et al. “Copy-number variation: the balance between gene dosage and expression in *Drosophila melanogaster*”. In: *Genome biology and evolution* 3 (1 2011), pp. 1014–1024. ISSN: 1759-6653. DOI: 10 . 1093 / GBE / EVR023. URL: <https://pubmed.ncbi.nlm.nih.gov/21979154/>.
- [12] Shaohua Fan and Axel Meyer. “Evolution of genomic structural variation and genomic architecture in the adaptive radiations of African cichlid fishes”. In: *Frontiers in Genetics* 5 (JUN 2014), p. 163. ISSN: 16648021. DOI: 10.3389/FGENE.2014.00163/ABSTRACT.
- [13] Frédéric J.J. Chain and Philine G.D. Feulner. “Ecological and evolutionary implications of genomic structural variations”. In: *Frontiers in Genetics* 5 (SEP 2014), p. 326. ISSN: 16648021. DOI: 10.3389/FGENE.2014.00326/BIBTEX.
- [14] Parithi Balachandran et al. “Structural variant identification and characterization”. In: *Chromosome Research* 2020 28:1 28 (1 Jan. 2020), pp. 31–47. ISSN: 1573-6849. DOI: 10 . 1007 / S10577 - 019 - 09623 - Z. URL: <https://link.springer.com/article/10.1007/s10577-019-09623-z>.
- [15] Christine R. Beck et al. “LINE-1 retrotransposition activity in human genomes”. In: *Cell* 141 (7 June 2010), pp. 1159–1170. ISSN: 00928674. DOI: 10.1016/J.CELL.2010.05.021/ATTACHMENT/EAD478C3-EAFA-4D3C-9633-F20D1E0FA2FA/MMC5.PDF. URL: [http://www.cell.com/article/S009286741000557X/fulltext%20http://www.cell.com/article/S009286741000557X/abstract%20https://www.cell.com/cell/abstract/S0092-8674\(10\)00557-X](http://www.cell.com/article/S009286741000557X/fulltext%20http://www.cell.com/article/S009286741000557X/abstract%20https://www.cell.com/cell/abstract/S0092-8674(10)00557-X).
- [16] Can Alkan, Bradley P. Coe, and Evan E. Eichler. “Genome structural variation discovery and genotyping”. In: *Nature Reviews Genetics* 2011 12:5 12 (5 Mar. 2011). Annotation 1, pp. 363–376. ISSN: 1471-0064. DOI: 10.1038/nrg2958. URL: <https://www.nature.com/articles/nrg2958>.
- [17] Aaron R. Quinlan and Ira M. Hall. “Characterizing complex structural variation in germline and somatic genomes”. In: *Trends in Genetics* 28 (1 Jan. 2012), pp. 43–53. ISSN: 0168-9525. DOI: 10.1016/J.TIG.2011.10.002. URL: [http://www.cell.com/article/S0168952511001685/fulltext%20http://www.cell.com/article/S0168952511001685/abstract%20https://www.cell.com/trends/genetics/abstract/S0168-9525\(11\)00168-5](http://www.cell.com/article/S0168952511001685/fulltext%20http://www.cell.com/article/S0168952511001685/abstract%20https://www.cell.com/trends/genetics/abstract/S0168-9525(11)00168-5).
- [18] Feng Zhang et al. “The DNA replication FoSTeS/MMBIR mechanism can generate genomic, genic and exonic complex rearrangements in humans”. In: *Nature Genetics* 2009 41:7 41 (7 June 2009), pp. 849–853. ISSN: 1546-1718. DOI: 10.1038/ng.399. URL: <https://www.nature.com/articles/ng.399>.
- [19] Andrew J. Holland and Don W. Cleveland. “Chromoanagenesis and cancer: mechanisms and consequences of localized, complex chromosomal rearrangements”. In: *Nature Medicine* 2012 18:11 18 (11 Nov. 2012), pp. 1630–1638. ISSN: 1546-170X. DOI: 10 . 1038 / nm . 2988. URL: <https://www.nature.com/articles/nm.2988>.
- [20] Franck Pellestor. “Chromoanagenesis: cataclysms behind complex chromosomal rearrangements”. In: *Molecular Cytogenetics* 2019 12:1 12 (1 Feb. 2019), pp. 1–12. ISSN: 1755-8166. DOI: 10 . 1186 / S13039 - 019 - 0415 - 7. URL: <https://molecularcytogenetics.biomedcentral.com/articles/10.1186/s13039-019-0415-7>.

- [21] T. Caspersson et al. “Chemical differentiation along metaphase chromosomes”. In: *Experimental Cell Research* 49 (1 Jan. 1968), pp. 219–222. ISSN: 0014-4827. DOI: 10.1016/0014-4827(68)90538-7.
- [22] Michael R. Speicher, Stephen Gwyn Ballard, and David C. Ward. “Karyotyping human chromosomes by combinatorial multi-fluor FISH”. In: *Nature Genetics* 12 (4 1996). ISSN: 10614036. DOI: 10.1038/ng0496-368.
- [23] Linping Hu et al. “Fluorescence in situ hybridization (FISH): An increasingly demanded tool for biomarker research and personalized medicine”. In: *Biomarker Research* 2 (1 Feb. 2014), pp. 1–13. ISSN: 20507771. DOI: 10.1186/2050-7771-2-3/TABLES/6. URL: <https://biomarkerres.biomedcentral.com/articles/10.1186/2050-7771-2-3>.
- [24] Elena V. Linardopoulou et al. “Human subtelomeres are hot spots of interchromosomal recombination and segmental duplication”. In: *Nature* 437:7055 437 (7055 Sept. 2005), pp. 94–100. ISSN: 1476-4687. DOI: 10.1038/nature04029. URL: <https://www.nature.com/articles/nature04029>.
- [25] Wulan Deng et al. “CASFISH: CRISPR/Cas9-mediated in situ labeling of genomic loci in fixed cells”. In: *Proceedings of the National Academy of Sciences of the United States of America* 112 (38 Sept. 2015), pp. 11870–11875. ISSN: 10916490. DOI: 10.1073/PNAS.1515692112/SUPPL_FILE/PNAS.201515692SI.PDF. URL: www.pnas.org/cgi/doi/10.1073/pnas.1515692112.
- [26] Chenghua Cui, Wei Shu, and Peining Li. “Fluorescence in situ hybridization: Cell-based genetic diagnostic and research applications”. In: *Frontiers in Cell and Developmental Biology* 4 (SEP Sept. 2016), p. 89. ISSN: 2296634X. DOI: 10.3389/FCCELL.2016.00089/BIBTEX.
- [27] Brian Teague et al. “High-resolution human genome structure by single-molecule analysis”. In: *Proceedings of the National Academy of Sciences of the United States of America* 107 (24 June 2010), pp. 10848–10853. ISSN: 00278424. DOI: 10.1073/PNAS.0914638107/SUPPL_FILE/ST04.XLS.
- [28] Saki Chan et al. “Structural variation detection and analysis using bionano optical mapping”. In: *Methods in Molecular Biology* 1833 (2018), pp. 193–203. ISSN: 10643745. DOI: 10.1007/978-1-4939-8666-8_16/COVER/. URL: https://link.springer.com/protocol/10.1007/978-1-4939-8666-8_16.
- [29] Shunichi Kosugi et al. “Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing”. In: *Genome Biology* 20 (1 June 2019), pp. 1–18. ISSN: 1474760X. DOI: 10.1186/s13059-019-1720-5/TABLES/1. URL: <https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1720-5>.
- [30] Peiyong Guan and Wing Kin Sung. “Structural variation detection using next-generation sequencing data: A comparative technical review”. In: *Methods* 102 (June 2016), pp. 36–49. ISSN: 1046-2023. DOI: 10.1016/J.YMETH.2016.01.020.
- [31] Søren M. Karst et al. “High-accuracy long-read amplicon sequences using unique molecular identifiers with Nanopore or PacBio sequencing”. In: *Nature Methods* 2021 18:2 18 (2 Jan. 2021), pp. 165–169. ISSN: 1548-7105. DOI: 10.1038/s41592-020-01041-y. URL: <https://www.nature.com/articles/s41592-020-01041-y>.
20<https://www.nature.com/articles/s41592-020-01041-y>.

- [32] Simon Ardui et al. “Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics”. In: *Nucleic acids research* 46 (5 Mar. 2018), pp. 2159–2168. ISSN: 1362-4962. DOI: 10.1093/NAR/GKY066. URL: <https://pubmed.ncbi.nlm.nih.gov/29401301/><https://pubmed.ncbi.nlm.nih.gov/29401301/?dopt=Abstract>.
- [33] Roger Ros-Freixedes et al. “Impact of index hopping and bias towards the reference allele on accuracy of genotype calls from low-coverage sequencing”. In: *Genetics Selection Evolution* 50 (1 Dec. 2018), pp. 1–14. ISSN: 12979686. DOI: 10.1186/S12711-018-0436-4/TABLES/9. URL: <https://gsejournal.biomedcentral.com/articles/10.1186/s12711-018-0436-4>.
- [34] Xiaowei Chen et al. “Biases and Errors on Allele Frequency Estimation and Disease Association Tests of Next-Generation Sequencing of Pooled Samples”. In: *Genetic Epidemiology* 36 (6 Sept. 2012), pp. 549–560. ISSN: 1098-2272. DOI: 10.1002/GEPI.21648. URL: <https://onlinelibrary.wiley.com/doi/full/10.1002/gepi.21648><https://onlinelibrary.wiley.com/doi/abs/10.1002/gepi.21648><https://onlinelibrary.wiley.com/doi/10.1002/gepi.21648>.
- [35] Pascal Lapierre and J. Peter Gogarten. “Estimating the size of the bacterial pan-genome”. In: *Trends in Genetics* 25 (3 Mar. 2009), pp. 107–110. ISSN: 0168-9525. DOI: 10.1016/J.TIG.2008.12.004. URL: <http://www.cell.com/article/S0168952509000055/fulltext><http://www.cell.com/article/S0168952509000055/abstract>[https://www.cell.com/trends/genetics/abstract/S0168-9525\(09\)00005-5](https://www.cell.com/trends/genetics/abstract/S0168-9525(09)00005-5).
- [36] Francisco Rodriguez-Valera and David W. Ussery. “Is the pan-genome also a pan-selectome?” In: *F1000Research* 2012 1:16 1 (Sept. 2012), p. 16. ISSN: 20461402. DOI: 10.12688/f1000research.1-16.v1. URL: <https://f1000research.com/articles/1-16>.
- [37] Christine Tranchant-Dubreuil, Mathieu Rouard, and Francois Sabot. “Plant Pangenome: Impacts on Phenotypes and Evolution”. In: *Annual Plant Reviews Online* 2 (2 May 2019), pp. 453–478. ISSN: 26393832. DOI: 10.1002/9781119312994.apr0664. URL: <https://onlinelibrary.wiley.com/doi/full/10.1002/9781119312994.apr0664><https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119312994.apr0664><https://onlinelibrary.wiley.com/doi/10.1002/9781119312994.apr0664>.
- [38] Fei Lu et al. “High-resolution genetic mapping of maize pan-genome sequence anchors”. In: *Nature Communications* 6 (2015). ISSN: 20411723. DOI: 10.1038/ncomms7914.
- [39] Wen Yao et al. “Exploring the rice dispensable genome using a metagenome-like assembly strategy”. In: *Genome Biology* 16 (1 2015). ISSN: 1474760X. DOI: 10.1186/s13059-015-0757-3.
- [40] Cécile Monat et al. “Comparison of two African rice species through a new pan-genomic approach on massive data”. In: *bioRxiv* (2018). DOI: 10.1101/245431. eprint: <https://www.biorxiv.org/content/early/2018/01/09/245431.full.pdf>. URL: <https://www.biorxiv.org/content/early/2018/01/09/245431>.
- [41] Habib Rijzaani et al. “The pangenome of banana highlights differences between genera and genomes”. In: *Plant Genome* 15 (1 2022). ISSN: 19403372. DOI: 10.1002/tpg2.20100.

- [42] Christine Tranchant-Dubreuil et al. “FrangiPANe, a tool for creating a panreference using left behind reads”. In: *bioRxiv* (2022). DOI: 10.1101/2022.07.14.499848. eprint: <https://www.biorxiv.org/content/early/2022/07/16/2022.07.14.499848.full.pdf>. URL: <https://www.biorxiv.org/content/early/2022/07/16/2022.07.14.499848>.
- [43] Jordan M. Eizenga et al. “Pangenome Graphs”. In: <https://doi.org/10.1146/annurev-genom-120219-080406> 21 (Sept. 2020), pp. 139–162. ISSN: 1545293X. DOI: 10.1146/ANNUREV-GENOM-120219-080406. URL: <https://www.annualreviews.org/doi/abs/10.1146/annurev-genom-120219-080406>.
- [44] Pavel A. Pevzner, Haixu Tang, and Michael S. Waterman. “An Eulerian path approach to DNA fragment assembly”. In: *Proceedings of the National Academy of Sciences of the United States of America* 98 (17 Aug. 2001), pp. 9748–9753. ISSN: 00278424. DOI: 10.1073/PNAS.171285098. URL: <https://www.pnas.org>.
- [45] Paul Medvedev and Michael Brudno. “Maximum likelihood genome assembly”. In: vol. 16. 2009. DOI: 10.1089/cmb.2009.0047.
- [46] Alexander Dilthey et al. “Improved genome inference in the MHC using a population reference graph”. In: *Nature Genetics* 47 (6 2015). ISSN: 15461718. DOI: 10.1038/ng.3257.
- [47] Glenn Hickey et al. “Genotyping structural variants in pangenome graphs using the vg toolkit”. In: *Genome Biology* 21 (1 Feb. 2020), pp. 1–17. ISSN: 1474760X. DOI: 10.1186/S13059-020-1941-7/TABLES/1. URL: <https://genomebiology.biomedcentral.com/articles/10.1186/s13059-020-1941-7>.
- [48] Heng Li, Xiaowen Feng, and Chong Chu. “The design and construction of reference pangenome graphs with minigraph”. In: *Genome Biology* 21 (1 Dec. 2020), pp. 1–19. ISSN: 1474760X. DOI: 10.1186/S13059-020-02168-Z/FIGURES/5. URL: <https://genomebiology.biomedcentral.com/articles/10.1186/s13059-020-02168-z>.
- [49] Wen-Wei Liao et al. “A Draft Human Pangenome Reference”. In: *bioRxiv* (July 2022), p. 2022.07.09.499321. DOI: 10.1101/2022.07.09.499321. URL: <https://www.biorxiv.org/content/10.1101/2022.07.09.499321v1%20https://www.biorxiv.org/content/10.1101/2022.07.09.499321v1.abstract>.
- [50] Joel Armstrong et al. “Progressive Cactus is a multiple-genome aligner for the thousand-genome era”. In: *Nature* 587 (7833 2020). ISSN: 14764687. DOI: 10.1038/s41586-020-2871-y.
- [51] Eric Dawson and Richard Durbin. “GFAKluge: A C++ library and command line utilities for the Graphical Fragment Assembly formats”. In: *Journal of Open Source Software* 4 (33 2019). DOI: 10.21105/joss.01083.
- [52] Giorgio Gonnella and Stefan Kurtz. “GfaPy: A flexible and extensible software library for handling sequence graphs in Python”. In: *Bioinformatics* 33 (19 2017). ISSN: 14602059. DOI: 10.1093/bioinformatics/btx398.
- [53] Chirag Jain et al. “Accelerating sequence alignment to graphs”. In: 2019. DOI: 10.1109/IPDPS.2019.00055.
- [54] Dmitry Antipov et al. “HybridSPAdes: An algorithm for hybrid assembly of short and long reads”. In: *Bioinformatics* 32 (7 2016). ISSN: 14602059. DOI: 10.1093/bioinformatics/btv688.

- [55] Mikko Rautiainen and Tobias Marschall. “GraphAligner: rapid and versatile sequence-to-graph alignment”. In: *Genome biology* 21 (1 Sept. 2020), p. 253. ISSN: 1474760X. DOI: 10.1186/s13059-020-02157-2.
- [56] Pesho Ivanov, Benjamin Bichsel, and Martin Vechev. “Fast and optimal sequence-to-graph alignment guided by seeds”. In: *bioRxiv* (2022). DOI: 10.1101/2021.11.05.467453. eprint: <https://www.biorxiv.org/content/early/2022/04/28/2021.11.05.467453.full.pdf>. URL: <https://www.biorxiv.org/content/early/2022/04/28/2021.11.05.467453>.
- [57] Hannes P. Eggertsson et al. “GraphTyper enables population-scale genotyping using pangenome graphs”. In: *Nature Genetics* 49 (11 Nov. 2017), pp. 1654–1660. ISSN: 15461718. DOI: 10.1038/ng.3964.
- [58] Hannes P. Eggertsson et al. “GraphTyper2 enables population-scale genotyping of structural variation using pangenome graphs”. In: *Nature Communications* 2019 10:1 10 (1 Nov. 2019), pp. 1–8. ISSN: 2041-1723. DOI: 10.1038/s41467-019-13341-9. URL: <https://www.nature.com/articles/s41467-019-13341-9>.
- [59] Jonas Andreas Sibbesen, Lasse Maretty, and Anders Krogh. “Accurate genotyping across variant classes and lengths using variant graphs”. In: *Nature Genetics* 50 (7 July 2018), pp. 1054–1059. ISSN: 15461718. DOI: 10.1038/s41588-018-0145-5.
- [60] Sai Chen et al. “Paragraph: A graph-based structural variant genotyper for short-read sequence data”. In: *Genome Biology* 20 (1 Dec. 2019). ISSN: 1474760X. DOI: 10.1186/s13059-019-1909-7.
- [61] S. Fukuoka et al. “Analysis of Vietnamese rice germplasm provides an insight into Japonica rice differentiation”. In: *Plant Breeding* 122 (6 2003). ISSN: 01799541. DOI: 10.1111/j.1439-0523.2003.00908.x.
- [62] Hien Thi Thu Vu et al. “Genetic diversity of Vietnamese lowland rice germplasms as revealed by SSR markers in relation to seedling vigour under submergence”. In: *Biotechnology and Biotechnological Equipment* 30 (1 2016). ISSN: 13102818. DOI: 10.1080/13102818.2015.1085330.
- [63] Nhung Thi Phuong Phung et al. “Characterization of a panel of Vietnamese rice varieties using DArT and SNP markers for association mapping purposes”. In: *BMC Plant Biology* 14 (1 2014). ISSN: 14712229. DOI: 10.1186/s12870-014-0371-7.
- [64] Benedict Paten et al. “Genome graphs and the evolution of genome inference”. In: *Genome Research* 27 (5 May 2017), pp. 665–676. ISSN: 1088-9051. DOI: 10.1101/GR.214155.116. URL: <https://genome.cshlp.org/content/27/5/665.full%20https://genome.cshlp.org/content/27/5/665%20https://genome.cshlp.org/content/27/5/665.abstract>.
- [65] Irina R. Arkhipova, Irina A. Yushenova, and Esther Angert. “Giant Transposons in Eukaryotes: Is Bigger Better?” In: *Genome Biology and Evolution* 11 (3 Mar. 2019), p. 906. ISSN: 17596653. DOI: 10.1093/GBE/EVZ041. URL: [/pmc/articles/PMC6431247/](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6431247/)?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6431247/.
- [66] Vladimir V. Kapitonov and Jerzy Jurka. “A universal classification of eukaryotic transposable elements implemented in Repbase”. In: *Nature Reviews Genetics* 9 (5 2008). ISSN: 14710056. DOI: 10.1038/nrg2165-c1.
- [67] Thomas Wicker et al. “A unified classification system for eukaryotic transposable elements”. In: *Nature Reviews Genetics* 8 (12 2007). ISSN: 14710064. DOI: 10.1038/nrg2165.

- [68] Rita Rebollo, Mark T. Romanish, and Dixie L. Mager. “Transposable elements: An abundant and natural source of regulatory sequences for host genes”. In: *Annual Review of Genetics* 46 (2012). ISSN: 00664197. DOI: 10.1146/annurev-genet-110711-155621.
- [69] Christian Biémont and Cristina Vieira. “Junk DNA as an evolutionary force”. In: *October* 443 (October 2006).
- [70] Christian Schlötterer et al. “Sequencing pools of individuals-mining genome-wide polymorphism data without big funding”. In: *Nature Reviews Genetics* 15 (11 2014). ISSN: 14710064. DOI: 10.1038/nrg3803.
- [71] Christoph D. Treiber and Scott Waddell. “Resolving the prevalence of somatic transposition in *Drosophila*”. In: *eLife* 6 (2017). ISSN: 2050084X. DOI: 10.7554/eLife.28297.
- [72] Edward Ryder and Steven Russell. “Transposable elements as tools for genomics and genetics in *Drosophila*”. In: *Briefings in Functional Genomics and Proteomics* 2 (1 2003). ISSN: 14739550. DOI: 10.1093/bfpg/2.1.57.
- [73] Casey M. Bergman et al. “Recurrent insertion and duplication generate networks of transposable element sequences in the *Drosophila melanogaster* genome.” In: *Genome biology* 7 (11 2006). ISSN: 14656914.
- [74] Yoshihiro Kawahara et al. “Improvement of the *Oryza sativa* nipponbare reference genome using next generation sequence and optical map data”. In: *Rice* 6 (1 2013). ISSN: 19398425. DOI: 10.1186/1939-8433-6-4.
- [75] Yong Zhou et al. “A platinum standard pan-genome resource that represents the population structure of Asian rice”. In: *Scientific Data* 7 (1 2020). ISSN: 20524463. DOI: 10.1038/s41597-020-0438-2.
- [76] Brian D. Ondov et al. “Mash: Fast genome and metagenome distance estimation using MinHash”. In: *Genome Biology* 17 (1 2016). ISSN: 1474760X. DOI: 10.1186/s13059-016-0997-x.
- [77] Éloi Durant et al. “Panache: A web browser-based viewer for linearized pangenomes”. In: *Bioinformatics* 37 (23 2021). ISSN: 14602059. DOI: 10.1093/bioinformatics/btab688.
- [78] Huong Thi Mai To et al. “Unraveling the Genetic Elements Involved in Shoot and Root Growth Regulation by Jasmonate in Rice Using a Genome-Wide Association Study”. In: *Rice* 12 (1 Dec. 2019), pp. 1–18. ISSN: 19398433. DOI: 10.1186/S12284-019-0327-5/TABLES/4. URL: <https://link.springer.com/articles/10.1186/s12284-019-0327-5>
<https://link.springer.com/article/10.1186/s12284-019-0327-5>.
- [79] Nhung Thi Phuong Phung et al. “Genome-wide association mapping for root traits in a panel of rice accessions from Vietnam”. In: *BMC Plant Biology* 16 (1 2016). ISSN: 14712229. DOI: 10.1186/s12870-016-0747-y.
- [80] Kim Nhung Ta et al. “A genome-wide association study using a Vietnamese landrace panel of rice (*Oryza sativa*) reveals new QTLs controlling panicle morphological traits”. In: *BMC Plant Biology* 18 (1 2018). ISSN: 14712229. DOI: 10.1186/s12870-018-1504-1.
- [81] Brian L. Browning, Ying Zhou, and Sharon R. Browning. “A One-Penny Imputed Genome from Next-Generation Reference Panels”. In: *American Journal of Human Genetics* 103 (3 2018). ISSN: 15376605. DOI: 10.1016/j.ajhg.2018.07.015.

- [82] Heng Li and Richard Durbin. “Fast and accurate short read alignment with Burrows-Wheeler transform”. In: *Bioinformatics* 25 (14 2009). ISSN: 13674803. DOI: 10.1093/bioinformatics/btp324.
- [83] Heng Li. “Minimap2: Pairwise alignment for nucleotide sequences”. In: *Bioinformatics* 34 (18 2018). ISSN: 14602059. DOI: 10.1093/bioinformatics/bty191.

Appendix A

Other publications

In this part, I mentioned manuscripts that were published and in preparation during my PhD.

A.1 Published paper

This is the work during my Master in 2018 and I am able to finished and published it with my colleagues in Vietnam in 2019.

ORIGINAL ARTICLE

Open Access



Unraveling the Genetic Elements Involved in Shoot and Root Growth Regulation by Jasmonate in Rice Using a Genome-Wide Association Study

Huong Thi Mai To^{1*}, Hieu Trang Nguyen^{1,2}, Nguyet Thi Minh Dang¹, Ngan Huyen Nguyen¹, Thai Xuan Bui¹, Jérémy Lavarenne², Nhung Thi Phuong Phung³, Pascal Gantet², Michel Lebrun^{1,2}, Stephane Bellafiore² and Antony Champion^{2*} 

Abstract

Background: Due to their sessile life style, plant survival is dependent on the ability to build up fast and highly adapted responses to environmental stresses by modulating defense response and organ growth. The phytohormone jasmonate plays an essential role in regulating these plant responses to stress.

Results: To assess variation of plant growth responses and identify genetic determinants associated to JA treatment, we conducted a genome-wide association study (GWAS) using an original panel of Vietnamese rice accessions. The phenotyping results showed a high natural genetic variability of the 155 tested rice accessions in response to JA for shoot and root growth. The level of growth inhibition by JA is different according to the rice varieties tested. We conducted genome-wide association study and identified 28 significant associations for root length (RTL), shoot length (SHL), root weight (RTW), shoot weight (SHW) and total weight (TTW) in response to JA treatment. Three common QTLs were found for RTL, RTW and SHL. Among a list of 560 candidate genes found to co-locate with the QTLs, a transcriptome analysis from public database for the JA response allows us to identify 232 regulated genes including several JA-responsive transcription factors known to play a role in stress response.

Conclusion: Our genome-wide association study shows that common and specific genetic elements are associated with inhibition of shoot and root growth under JA treatment suggesting the involvement of a complex JA-dependent genetic control of rice growth inhibition at the whole plant level. Besides, numerous candidate genes associated to stress and JA response are co-located with the association loci, providing useful information for future studies on genetics and breeding to optimize the growth-defense trade-off in rice.

Keywords: Jasmonate, Plant development, Growth inhibition, Genome-wide association studies, *Oryza sativa*

* Correspondence: to-thi-mai.huong@usth.edu.vn; antony.champion@ird.fr

¹University of Science and Technology of Hanoi (USTH), Vietnam Academy of Science and Technology (VAST), LMI-RICE2, 18 Hoang Quoc Viet, Cau Giay district, Hanoi, Vietnam

²Institut de Recherche pour le Développement (IRD), Université de Montpellier, UMR DIADE, UMR IPME, UMR LSTM, Montpellier, France
Full list of author information is available at the end of the article

Background

Rice is the main staple food for over half of the world's population. The demand on the quality and quantity of rice is now more and more strongly emphasized due to increasing global population as well as in the context of climate change.

In order to withstand both biotic and abiotic stresses, plants evolved efficient and sophisticated systems including an inducible defense system mediated by jasmonate (JA) (Nahar et al. 2012; Okada et al. 2015; Khan et al. 2016). This phyto-hormone synthesized from the fatty acid linolenic acid, plays important roles in plants, for example, in regulating development, growth and defense response to biotic stresses. The function of JA relies on the core, conserved signaling module, which constitutes the amino acid-conjugated bioactive compound JA-Ile, the JA receptor CORONATINE INSENSITIVE 1 (COI1) protein, the co-receptor and repressor JASMONATE-ZIM domain protein (JAZ), and the transcription factors (such as MYC2, MYC3, MYC4, MYC5...) (Chini et al. 2007; Katsir et al. 2008; Ye et al. 2009; Kazan and Manners 2013; Yan et al. 2016; Howe et al. 2018). In response to the developmental cues or stress signals, JA-Ile is accumulated and perceived by the COI1 receptor, the JAZ repressors is recruited for ubiquitination and degradation through the 26S proteasome manner, thereby relieving the repression of transcription factors and enabling the expression of JA-responsive genes and JA responses (Wasternack and Hause 2013). Besides these main components, the JAZ repressor also interact with other proteins, such as TOPLESS and NINJA (NOVEL INTERACTOR of JAZ), to repress the transcription factor MYC2 (Pauwels et al. 2010; Gasperini et al. 2015). The rice genome consists of three *OsCOI* genes (designated as *OsCOI1a*, *OsCOI1b*, and *OsCOI2*), in which only *OsCOI1a* and *OsCOI1b* could rescue the sterility phenotype in *Arabidopsis coi1* mutant (Lee et al. 2013). The number of *OsJAZ* and *OsNINJA* gene are also higher than that in *Arabidopsis* with 15 *OsJAZ* genes (designated as *OsJAZ1* to *OsJAZ15*) (Ye et al. 2009), and four *OsNINJA* (Kashihara et al. 2019), as compared to 13 JAZ and a single NINJA in *Arabidopsis*, respectively. Among these genes, the function of some of them have been investigated (Ye et al. 2009; Yamada et al. 2012; Toda et al. 2013; Cai et al. 2014; Horii et al. 2014; Wu et al. 2015; Hakata et al. 2017; Li et al. 2017; Kashihara et al. 2019). The functional homologous of MYC2, OsMYC2, was also characterized as the master regulator that involved in numerous response in rice (Uji et al. 2016; Ogawa et al. 2017a; Ogawa et al. 2017b; Uji et al. 2017). Although it is generally postulated that JA impacts the energy balance of plants between defense and development during stress condition, the mechanisms underlying this regulation are still not fully understood (Wasternack 2007; Nahar et al. 2011). When the JA-dependent defense system is continuously

activated, growth of the plant can be severely affected (Vos et al. 2013; Huot et al. 2014). As an example, repeated wounding of *Arabidopsis* cotyledons inhibits cell division and cell elongation causing a reduction in root length in a JA-dependent manner (Gasperini et al. 2015). In *Arabidopsis* and rice, interaction between JA and gibberellin signaling cascade regulate the growth-defense dynamics by which plant prioritize JA-dependent defense response over growth (Yang et al. 2012). Recently, Marcelo Campos et al demonstrated that growth-defence antagonism in *Arabidopsis* is regulated through a JA-dependent transcriptional network that restricts growth upon defense activation by JA (Campos et al. 2016). These contrasting activities of the hormone imply a broader role for the JA in regulating a compromise between growth and defense, thereby optimizing plant fitness in rapidly changing environments. Therefore, it is interesting to identify specific genes that could optimize the defense system of plants with less impact on growth.

Since JA is a central hormone in various stress signaling processes in plants, applying JA is a simplified way to mimic different types of stress. For example, the exogenous application of JA has been widely used to screen mutants collection affected in stress and development responses of plants in general and in rice in particular (Ye et al. 2009; Chan et al. 2010; Nahar et al. 2012). In *Arabidopsis*, the JA receptor COI1 protein has been identify following the phenotyping of mutagenized seedlings treated by coronatine, a structural homologue of the bioactive jasmonoyl-isoleucine hormone (Feys et al. 1994). In addition to COI1, other major JA-signaling proteins such as a JAZ transcriptional repressor and the transcription factor AtMYC2 have been identified based on JA treated mutant population (Lorenzo et al. 2004; Chini et al. 2007). Beside, exogenous JA application has also been widely used to study complex treats such as dynamic and architectures of JA regulatory network and hormones interaction involved in root architecture or leaf growth inhibition (Sun et al. 2009; Noir et al. 2013; Hickman et al. 2017).

In order to explore the diversity of rice growth inhibition in response to JA, 155 accessions of a rice collection were treated with exogenous JA. This collection represent a core panel of Vietnamese rice landraces that originate from different geographical locations and are adapted to different agrosystems including irrigated, rain-fed, lowland, upland or mangrove ecosystems. In accordance with their genetic nature, they can be divided into 3 types: *Indica* (89% - mostly lowland rice), *Japonica* (9.5% - mostly upland rice) and other (1.5%) (Bui 2010). They possess many precious traits including resistance to biotic stress, such as blast, blight, and brown plant hopper, and abiotic stress, such as tolerance to salinity or drought as well as submergence conditions (Bui and Nguyễn 2003; Bui 2010). Some accessions that

have good tolerance to various types of stress have been identified and used as sources of stress tolerance in breeding to create new varieties. However, even though the Vietnamese rice resources likely carry many valuable alleles that can be beneficial for agronomy, they remain insufficiently explored and utilized (Courtois et al. 1997; Bui 2010).

This collection has also already been genotyped by GBS with 21,613 SNP (Phung et al. 2014). Phung et al have successfully used this collection to identify new QTLs associated with root development in non-stress conditions using a genome-wide association study (GWAS) (Phung et al. 2016). Furthermore, GWAS has been conducted using this core panel of rice landraces to reveals new association loci controlling panicle morphological traits (Ta et al. 2018) and tolerance to water deficit during vegetative stage demonstrating the valuable genetic resource of this rice collection (Hoang et al. 2019).

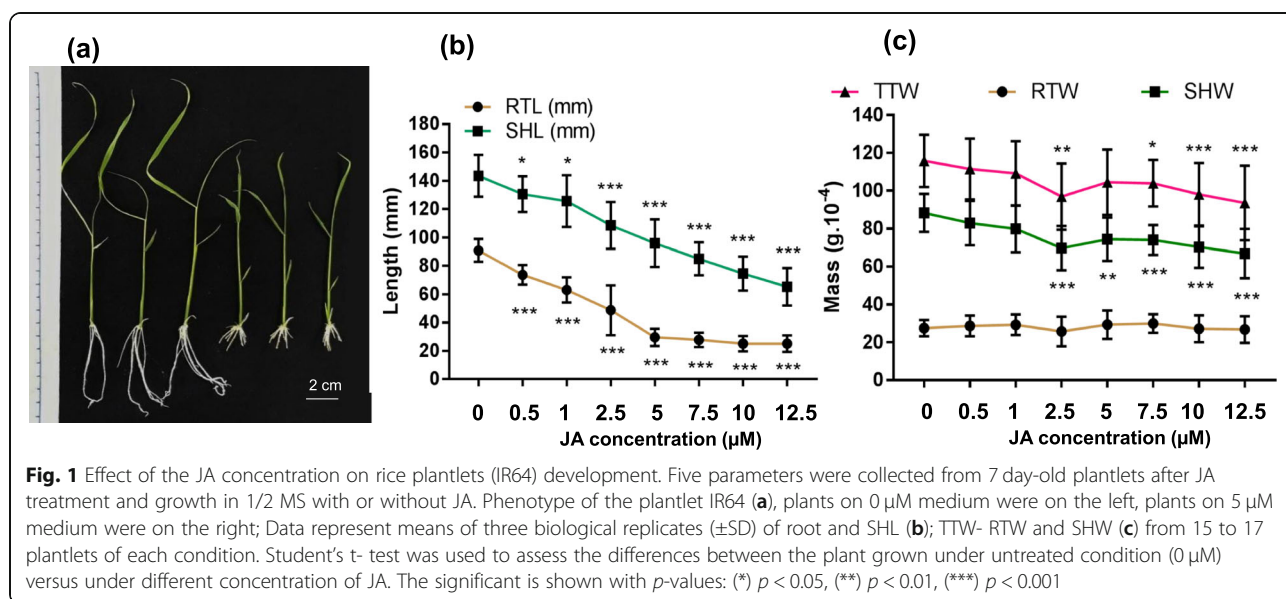
While responses to jasmonate have been studied considerably, it remains unclear how these responses are modulated by the genetic diversity. Here, we assess variation of root and shoot traits upon treatment with JA in the 155 rice accessions. We conducted genome-wide association studies and identified 28 significant associations with rice growth inhibition by JA, including three common QTLs associated to root and shoot growth inhibition traits. Using functional annotation and expression analyses we retrieved 560 genes linked to the most significant markers, and found 42% of the candidate genes to be responsive to jasmonate, indicating that among these genes are possible new players in jasmonate growth inhibition response pathway.

Results

Dose-Effect of JA on Rice Plantlet Growth

To study the dose-response to JA, in vitro experiments were carried out with *Oryza sativa indica* accession IR64 in order to define the optimal JA concentration for treatments. Eight concentrations of JA as 0 μ M (control), 0.5 μ M, 1 μ M, 2.5 μ M, 5 μ M, 7.5 μ M, 10.5 μ M and 12.5 μ M were chosen after several trials (data not shown). After 7 days of exposure to JA, root length (RTL), shoot length (SHL), root weight (RTW), shoot weight (SHW) and total weight (TTW) were evaluated as parameters to analyze the dose-effect of JA on rice growth.

As shown in Fig. 1a and b, treatment with 0.5 μ M of JA was sufficient to reduce RTL and SHL compared to the untreated condition ($P < 0.05$). This effect was highly significant ($P < 0.0001$) from the JA concentration of 2.5 μ M for these both parameters. At 2.5 μ M JA, a distinct difference was also observed for SHW ($P < 0.0001$) and TTW ($P < 0.0001$), as shown in Fig. 1c. However, we observed a remarkable reduction for SHL and RTL at 5 μ M JA of approximately 40% and 60%, respectively. Therefore, to ensure that we could observe the diversity of response within varieties, the phenotypes of these accessions after treatment with 5 μ M JA for a total 10 days from the germination stage, including 7 days of JA exposure, were recorded. A first trial of experiment with 10 Vietnamese rice accessions that represent the genetic diversity of the Vietnamese rice accessions were chosen based on the study of Phung et al (Phung et al. 2014) was conducted. The Vietnamese rice panel was divided into 2 main sub-panels: *Indica* and *Japonica*. The *indica* subpanel was further divided into six populations (I1 to I6) and the *japonica* subpanel was divided into 4 populations (J1 to J4). This screening with 6 *Indica* accessions



and 4 *Japonica* accessions was performed to evaluate the diversity of our rice accessions response to JA exogenous treatment. The phenotyping dataset indicated an obvious difference between treated and non-treated plants for all 10 varieties, and these representative accessions respond significantly differently to JA (5 μ M) (Additional file 1: Figures S1-S3 and Table S1). This preliminary study showed that 5 μ M of JA was suitable to highlight varied growth response with different genotypes. In addition, IR64 was used as an internal control in the full panel phenotyping,

Trait Heritability and Genetic Variability of 155 Rice Accessions

A list of the 155 accessions and their relevant information used in this study is presented in Additional file 2: Table S2. An analysis on the phenotypic variation and broad sense heritability is presented in Table 1. To confirm whether the variation exhibited in a certain trait could be attributed to genetic variation, the heritability coefficient was calculated according to the formula given by (Wray and Visscher 2008). From the result obtained, in both control and treatment conditions with the broad-sense heritability of all traits ranging from 0.90 to 0.96, the variations of the 5 studied traits were caused mostly by genetic polymorphisms, with well-controlled environmental factors. Hence, the quality of the phenotype dataset was sufficient for further study with genome-wide association mapping.

Phenotypic Diversity of the Vietnamese Rice Collection in Response to 5 μ M JA Treatment

Effects of 5 μ M JA on the phenotypic variation of the growth traits for 155 rice accessions were evaluated and illustrated in Fig. 2 and Additional file 3: Table S3.

The 155 rice accessions show considerable diversity in response to JA for all traits, predicting the diversity in response to environmental stresses. Except for RTW, exogenous JA treatment has a negative impact on all analyzed traits causing the limited growth of rice accessions. JA treatment has a significant negative effect on RTL and shoots length for all accessions. There are some rice accessions that are particularly affected by JA, such as accession G40 in which the RTL was inhibited up to 97% compared to the control condition, while the

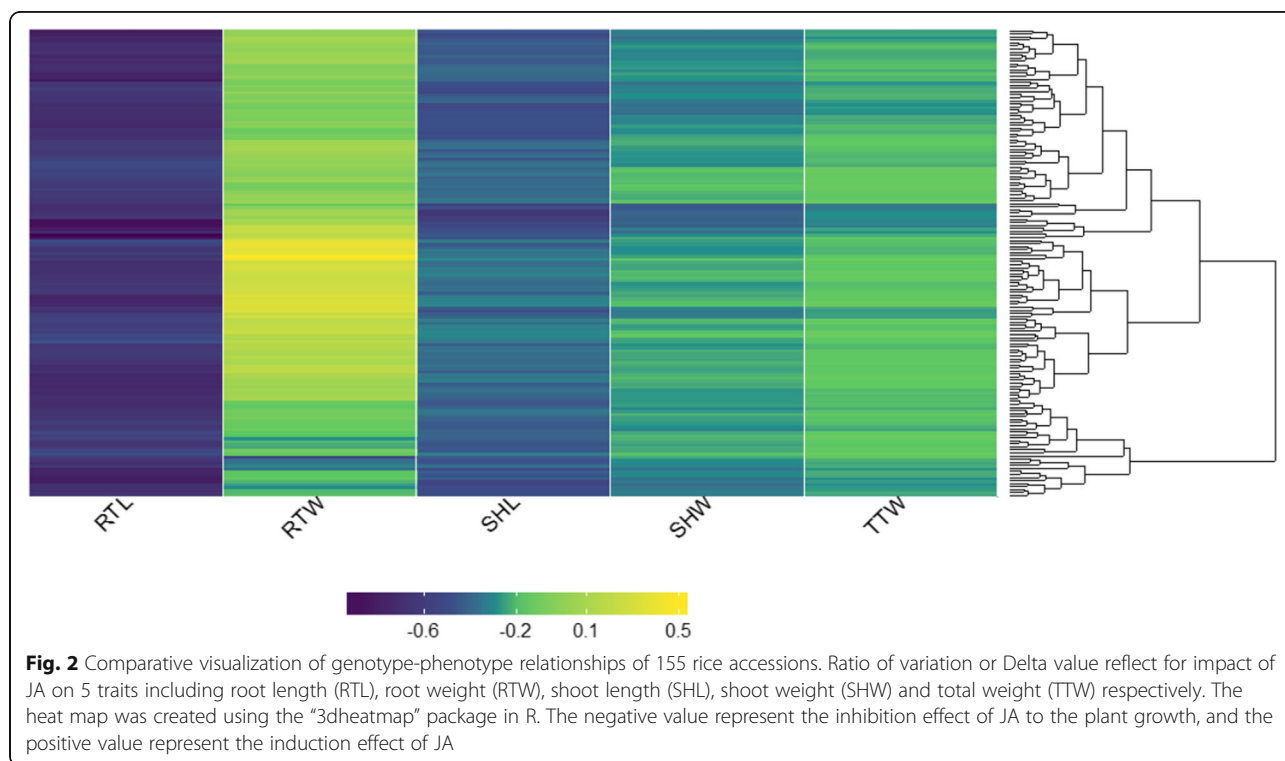
least susceptible variety, such as accession G52, showed RTL inhibition of 48% in response to JA. JA also inhibits shoot growth with a slightly lower ratio than that of the root, the most susceptible G132 with 66% SHL reduction and the least susceptible G146 with only 30% reduction, compared to plantlets in control condition. In contrast to the length parameters, the RTW and TTW of the plant are also affected but do not follow the same inhibition trend. Particularly, over half of accessions tend to increase the RTW when plants are stressed with exogenous JA. For example, with accession G65, although the root length is reduced by 68%, the dry weight of roots increased by nearly 50% when this variety was treated while for accession G223 both of RWT and RTL reduced by 55% and 62%, respectively. This can be explained by the increase in the number of crown roots as well as the root's diameter in these accessions (data not shown). Taken together, these results indicated a high phenotypic diversity among the rice core panel in response to JA.

Phenotypic Heterogeneity Among *Indica* and *Japonica* Sub-Groups and Among Ecosystem

The phenotypic heterogeneity among *Indica* and *Japonica* sub-groups (Fig. 3), among ecosystem (Fig. 4) and pair correlation of the 5 traits provide a descriptive overview of the data distribution for different rice subpopulations. In general, there are no significant differences within sub-groups, and the effects of JA response for all parameters in 2 sub-groups are quite homogenous. This means that the response to JA in this study is a universal response for *O. sativa L.* and is not affected by sub-group population. Apparently, *Indica* accessions tend to have higher values for RTW and SHL reduction by JA compared to *Japonica* accessions, which have a slightly higher value on RTL, SHW and TTW (Fig. 3). The correlation between the 5 studied traits was also evaluated and shown in the Fig. 3. There is a very strong correlation in term of inhibition effect between SHW, SHL and TTW, especially for SHW and TTW with r value of 0.99. Interestingly, there are a weak correlation in term of inhibition effect of JA among root traits; and almost no correlation between RTW and RTL, which was indicated by the r^2 value close to 0 (Fig. 3). Regards to their ecosystems, there are almost no clear differences between rice varieties which are Irrigated, Mangrove, and Rainfed lowland groups in term of response to

Table 1 phenotypic variation and broad sense heritability for each trait

Trait	Mean None treated	CV None treated	H^2 None treated	Mean JA treated	CV JA treated	H^2 JA treated
RTL (mm)	62.73986	7.857902	0.95	20.23247	7.858259	0.9
SHL (mm)	161.9466	28.58407	0.96	94.68829	24.91233	0.95
RTW (mg)	23.42427	5.550502	0.9	24.10098	6.765589	0.92
SHW (mg)	83.87246	20.16218	0.95	61.34949	16.19226	0.95
TTW (mg)	107.2967	25.03204	0.95	85.45047	21.83225	0.95



JA treatment except there are significant differences between Rainfed lowland groups and Upland rice groups for RTL, SHL and RTW in response to JA treatment (p -value < 0.01) (Fig. 4).

Association Mapping on Growth Traits in Response to JA Treatment

To explore whether specific SNPs markers in the genome are associated with the effects of exogenous JA on the phenotypic trait of rice growth, we performed GWAS on RTL, SHL, RTW, SHW and TTW using TASSEL v5.0. The rice collection has a bipolar structure with two main sub-groups, *Indica* and *Japonica* (Phung et al. 2014). Since GWAS results are greatly affected by the structure of the studied population, a GWAS was additionally performed with each sub-group to maintain a better population structure control and reduce the false positive rate. Therefore, the association mapping was examined for 155 accessions as whole panels, then separately, examining each independent sub-population including the *Indica* subpanel (95 accessions) and the *Japonica* sub-panel (60 accessions) (Figs. 5 and 6 and Additional file 4).

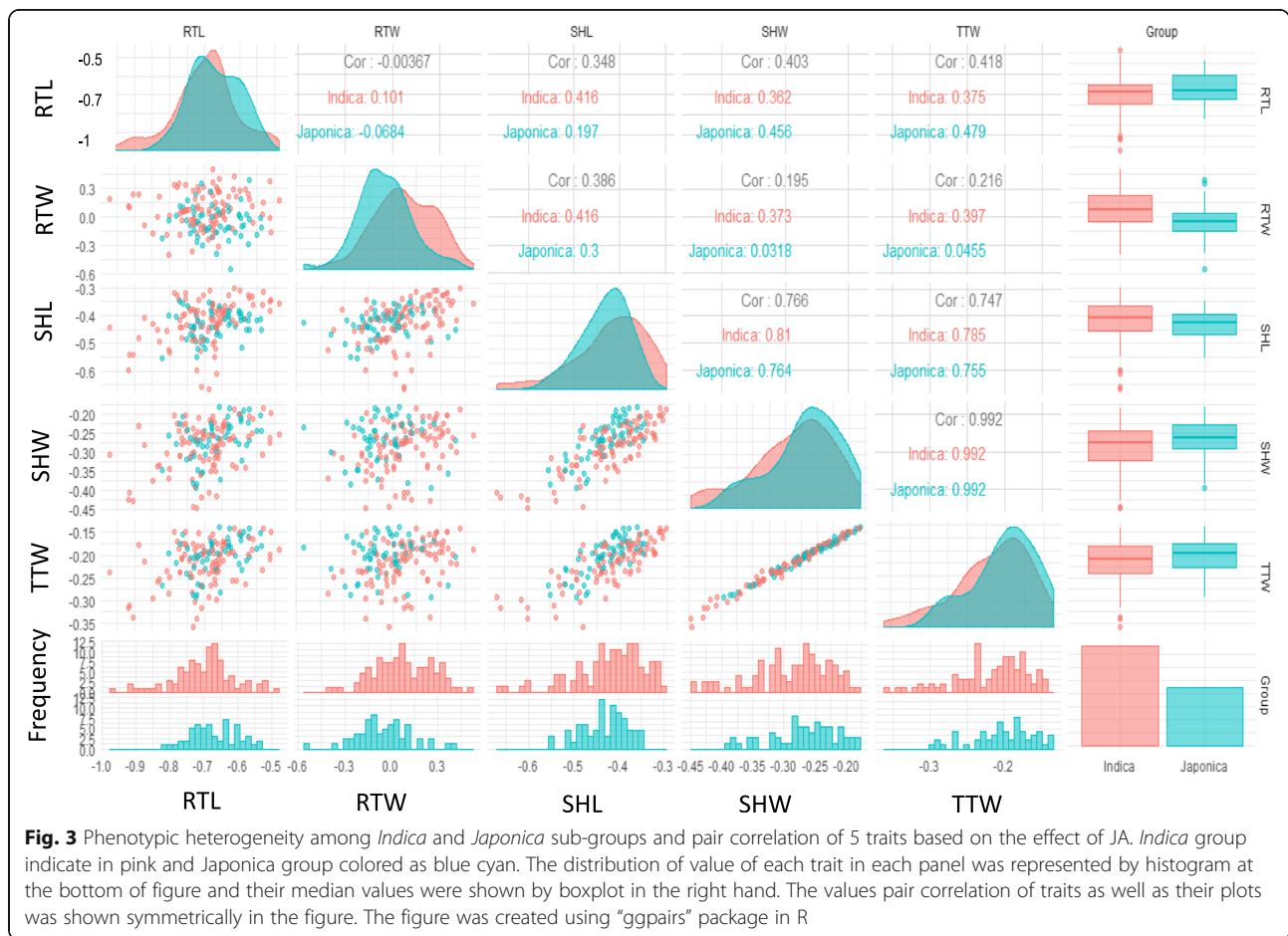
For all 5 traits in the whole panel, we applied the Mix Linear Model (MLM) that reconsiders both the population structure and kinship matrix. The results of genome-wide association mappings for effects of exogenous JA on SHL and RTW traits are illustrated in Figs. 5 and 6, respectively. Association mapping results

for RTL, SHW and TTW traits are presented in the Additional file 4: Figures S4, S5 and S6, respectively. In the sub-group panel, especially for a small size panel such as *Japonica* (60 accessions), the General Linear Model (GLM) was also used in conjunction with the MLM, and Quantile-quantile (Q-Q) plots showed a good fit between observed p -values and theoretical p -values. In this study, most Q-Q plots display good accumulative distributions of observed p -values, which were well fitted with the expected uniform distribution for the smallest $-\log_{10}(p$ -values).

Regarding the association of the effect of JA on SHL (Fig. 5), the Q-Q plot of all these plots for 3 panels were well fitted between observed p -values, and theoretical p -values. The statistical data pointed out 27 significant markers including 12 in the Whole panel (Fig. 5.1a), 8 significant markers in the *Indica* subpanel (Fig. 5.2a), and 7 significant markers in the *Japonica* subpanel (Fig. 5.3a). Chromosome 1 harbored 4 markers and chromosome 12 carried 4 markers for the whole panel, while chromosome 5 held 4 markers each for the *Japonica* subpanel.

The association mapping results describing the effect of JA on RTW is presented in Fig. 6. Q-Q plots showed well fitted compared to the expected line for all 3 panels. Thirteen significant markers were obtained. Data from MLM reported 5 significant markers to be shared between the Whole panel and *Indica* sub-panel.

Concerning the results on the RTL trait, presented in Additional file 4: Figure S4, there were 14 significant



markers found, in which the whole panel harbored 12 markers, the *Indica* sub-panel held 2 markers. There was only one common marker shared between the whole panel and the *Indica* sub-panel (Dj02_29022158F). Association mappings on SHW (Additional file 4: Figure S5) showed 21 significant markers, in which the whole panel holds 4 markers, the *Indica* sub-panel holds 9 markers and the *Japonica* sub-panel holds 8 markers. There were 5 significant markers appeared in chromosome 1 in *Indica* sub-panel. Twenty-three significant markers were obtained on TTW trait including 6 significant markers in Whole panel, 8 in *Indica* sub-panel and 9 in *Japonica* sub-panel (Additional file 4: Figure S6).

QTLs and Significant Markers Identification in the Whole Panel, *Indica* and *Japonica* Sub-Panel

Based on the results computed by TASSEL, 98 significant markers were identified at the threshold p -value $< 3.0E-04$. The linkage disequilibrium of each significant marker with their nearby markers was calculated to evaluate the possibility that these markers are associated within a population. Based on the calculation of linkage disequilibrium between SNPs computed using LD heatmap package, the list of

QTLs was generated and is presented in Additional file 5: Table S4. As shown in this Table 3 QTLs for the RTL trait, 3 QTLs for RTW, 7 QTLs for SHL, 7 QTLs for SHW and 8 QTLs for TTW were identified and associated with the most significant markers in this study. The mean size of these QTLs is around 290 kb, and the smallest QTL was qTTW1 measuring 50 kb in length and the largest QTL was qTTW4 with length of 1423 kb. In total, 28 QTLs were recorded with 98 considered significant markers for the effects of exogenous JA treatments on these 5 traits. There are 14 QTLs indicated in the Whole panel, 7 QTLs in *Indica* sub-panel and 4 QTLs in *Japonica* sub-panel.

TTW is a trait with a highest number of QTLs and significant markers, totaling 8 QTLs and 22 significant markers. Seven markers of the whole panel, 11 markers of the *Indica* subpanel and 6 markers of the *Japonica* sub-panel were obtained. For this parameter, 2 commons markers were shared between the Whole panel and *Japonica* sub-panel which were Sj04_32796870R and Sj04_32851563F in qTTW5. The qTTW3 with length of 532 kb in chromosome 1 contains 4 significant markers for the *Indica* sub-panel. The qTTW4 was also the longest QTL (1.4 Mbp), but it contained only 3 significant markers.

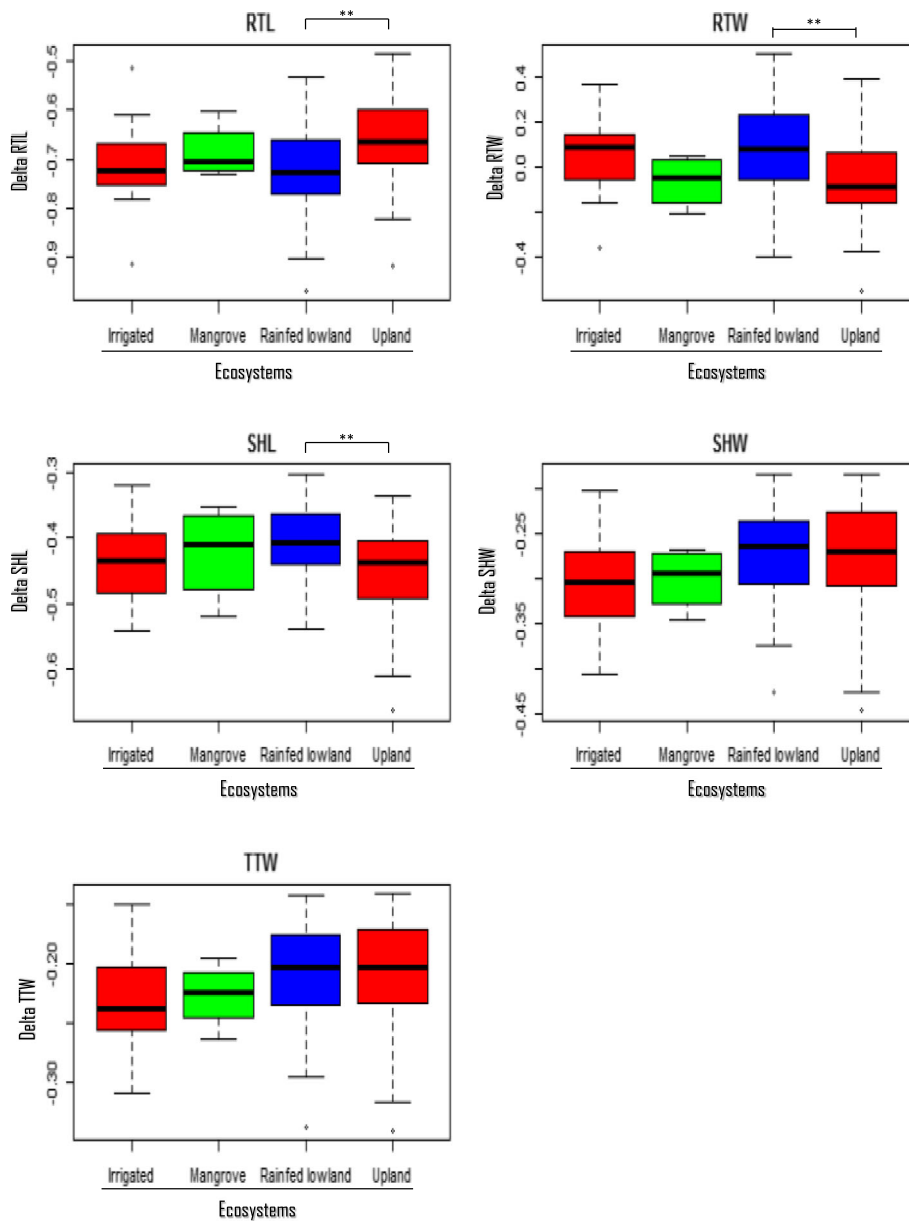


Fig. 4 Phenotypic heterogeneity among ecosystems. The response to JA of rice accessions among four main ecosystems were compared. Student's t- test was used to assess the differences between each couple of ecosystem. The (**) significant is shown with p - values < 0.01

For RTW, 9 markers for 3 QTLs were discovered after applying the LD heatmap. In addition, the Whole panel and *Indica* subpanel shared 9 markers together. The QTL named qRTW3, measuring 221 kb at chromosome 7, which holds 7 markers at p -values at least less than 5.1E-04.

Regarding the RTL trait, LD heatmaps were applied and the outcome was 3 QTLs with 5 significant markers. The whole panel contained 4 markers, doubled than the *Indica* which contained 2 markers. The *Japonica* sub-panel did not harbor any significant markers. There is only one significant marker that appeared in both the whole panel and *Indica* subpanel at chromosome 2 (Dj02_29022158F).

In the SHW trait, there were 7 QTLs and 14 significant markers. The *Indica* sub-panel had most significant markers, 6 markers were identified in the Whole panel; the *Japonica* sub-panel contained only 1 marker. qSHW1 of 532 kb was the longest QTL responsible for SHW, but only 3 significant markers were obtained.

The SHL trait held 7 QTLs and 16 significant markers including 9 markers for Whole panel, only 1 for *Indica* sub-panel and 6 for *Japonica* sub-panel.

The common QTLs across 5 studied traits were determined. Notably, we detected 3 pairs of overlap QTLs for 5 given traits, the qSHW1/qTTW3 located in chromosome 1

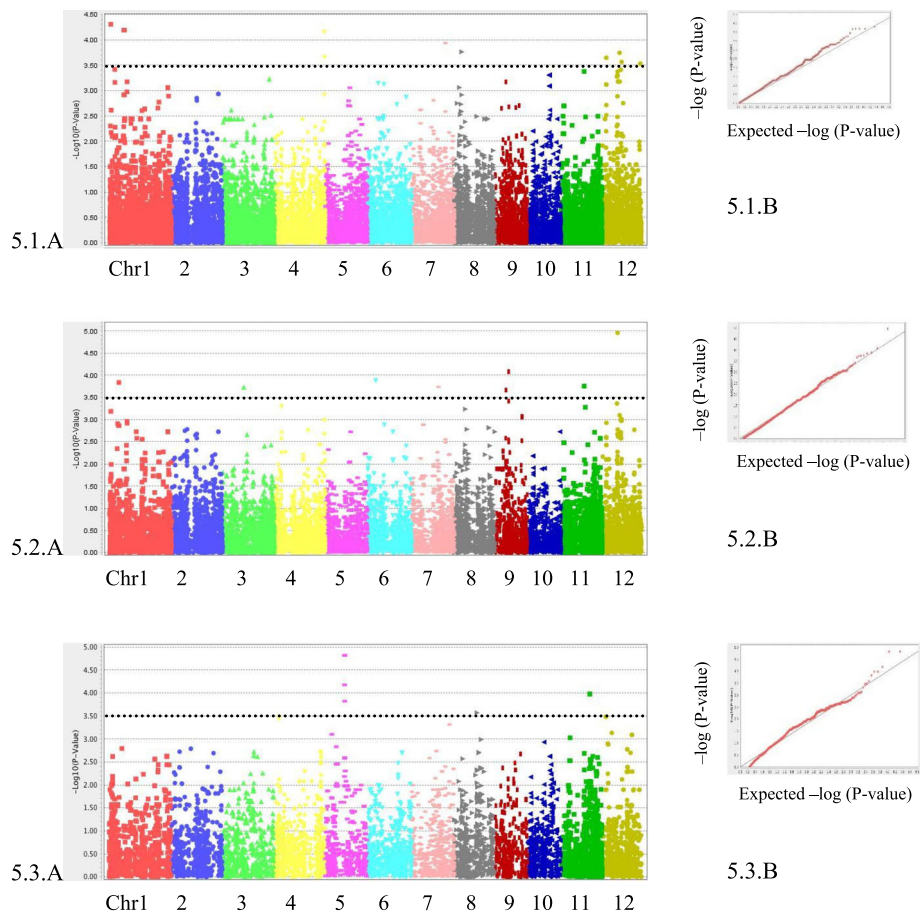


Fig. 5 GWAS for the effects of exogenous JA on SHL. Manhattan plot (a) and Quantile-quantile plot (b) for shoot length in a whole (5.1) panel or *Indica* (5.2) or *Japonica* (5.3) sub-panel. The black line indicates the suggestive significance threshold, $p = 3.0E-04$

including 4 significant markers (Dj01_38614134F, Dj01_38614137R, Sj01_38847771R and Sj01_38997056R); the qSHW7/qTTW8 located in chromosome 12 including 2 significant markers (Dj12_24056700R and Sj12_24257096R); the qSHW6/qSHL6 located in chromosome 12 included 3 lead markers (Sj12_02947152R, Sj12_02947164F and Dj12_2975990R). In short, SHW and TTW shared 2 common QTLs, SHL shared 1 common QTLs with SHW.

These results are the preliminary screening for the QTL associated with the growth of rice under stress condition. Further analysis should be conducted to functionally characterize these QTLs. The validated QTLs could help to dissect the underlying molecular mechanism controlling the defense/development trade off in rice.

Polymorphism Combination of Significant Markers Inside Some Selected QTLs

Using the significant SNPs located in each QTLs detected by GWAS (threshold p -value = $3.0 E-4$) to make a haplotype sequence, polymorphism combination analysis were shown in the Figs. 7 and 8 and Additional file 6: Figure S7

for 4 selected QTLs with at least 4 markers. Related to the overlap QTL named qSHW1/TTW3 that was detected by GWAS for the *Indica* subgroup (Fig. 7), the effect of JA stress is more important with accessions contained the ATAG haplotype than those contained the TAGT haplotype in the QTLs to the SHW (p -value = 0.006882) and TTW (p -value = 0.004772). For qRTW3, all the 6 markers were associated with RTW by presenting a strong linkage in a red block showed in the Fig. 8a and revealed 2 major haplotypes in Fig. 8b. The positive effect of JA on RTW were associated with AATAAAT haplotype and significantly higher than those contained the TTCTTTG (p -value = 0.0007959) (Fig. 8c). For the qTTW5 in *Japonica* panel that was detected after the GWAS analysis (Additional file 6: Figure S7a), 2 main haplotypes TAGT and CGAG were revealed (Additional file 6: Figure S7b), and haplotype 1 has higher value compare to haplotype 2 (p -value = 0.002629) (Additional file 6: Figure S7c). These results support that the sequence variation in each QTL region can contribute to the phenotypic difference of interested traits.

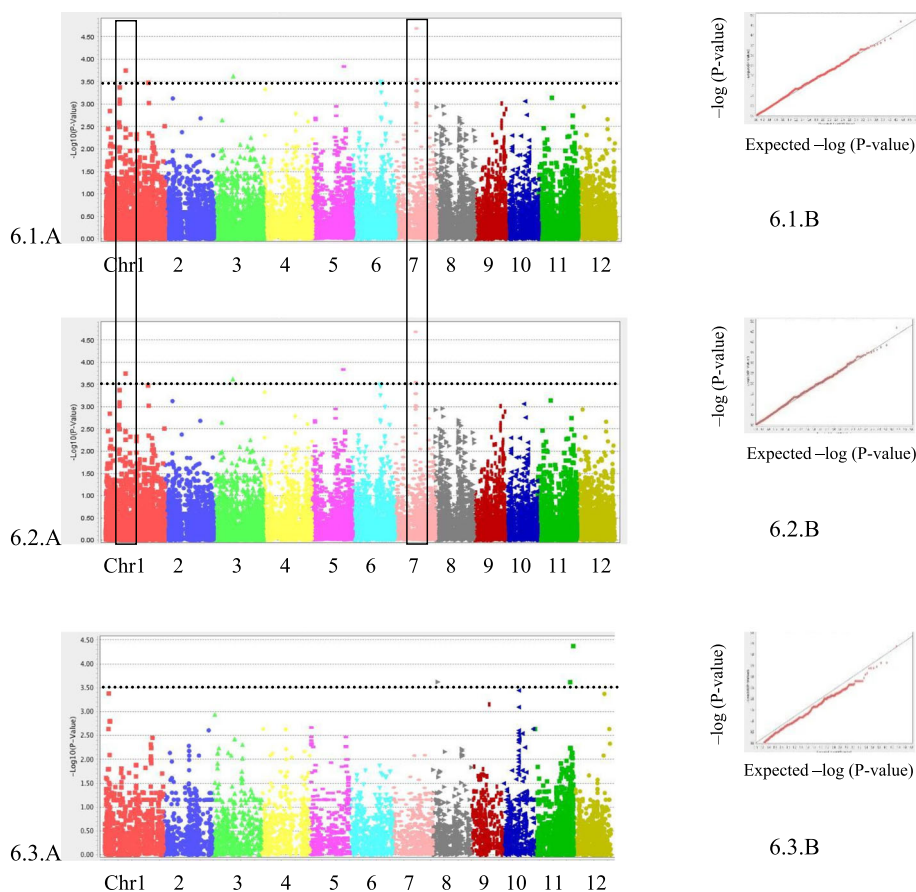


Fig. 6 GWAS for the effects of exogenous JA on RTW. Manhattan plot (a) and Quantile-quantile plot (b) for root weight in a whole (6.1) panel or *Indica* (6.2) or *Japonica* (6.3) subpanel. The black line indicates the suggestive significance threshold, $p = 3.0E-04$

Identification of Candidate Genes Co-Localized with the Significant Markers

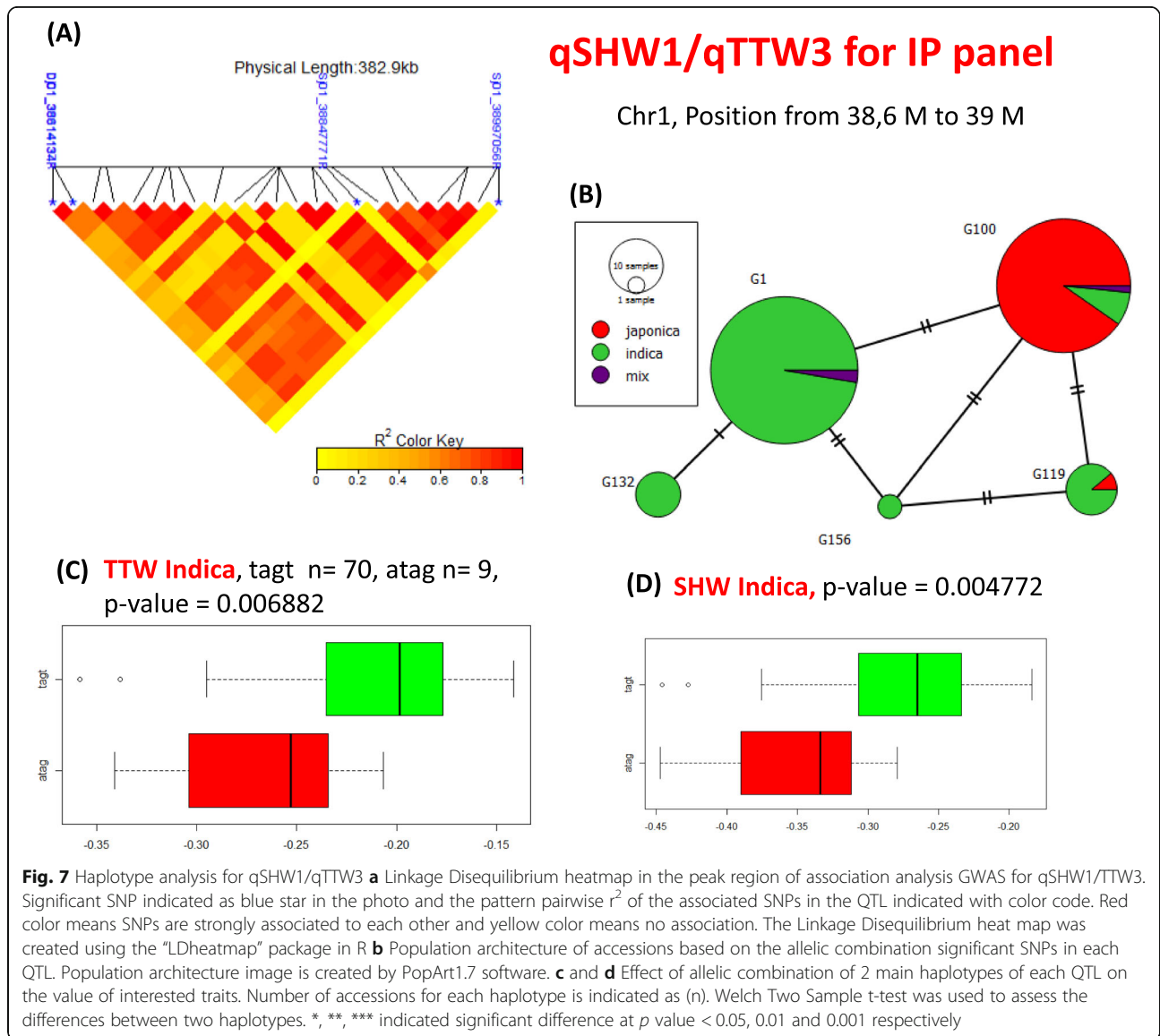
To identify candidate genes, annotations of the genes were first analyzed within the identified QTLs (25 kb up and downstream of the most significant SNP of the QTL) based on the MSU Rice Genome Annotation Project Database Release 7 (Kawahara et al. 2013). Among the significant markers found in the interval mapping of identified QTLs, there were 560 potential candidate genes found located on these QTLs with their annotation after removing all the transposons, hypothetical proteins and expressed proteins (Additional file 7: Table S5).

From this list, no known genes associated either with octadecanoid pathway (such as *OsAOS*, *OsAOC*, *OsJARI*, etc.) or with jasmonate perception (such as *OsCOII*, *OsJAZ* and *OsNINJA* genes) could be identified. We then conducted an annotation to functionally categorize the 560 genes using the second hierarchy level of the MapMan bins mapping. Detailed information and annotated function of these candidate genes is presented in Additional file 7: Table S5. The 10 most represented

MapMan bins that represent 44% of the annotated genes indicated functions in regulation of DNA transcription, receptor kinases and calcium signaling, protein post-translational modification, protein degradation, biotic and abiotic stresses (Table 2).

Identification of Candidate Genes Related to Jasmonate Signalling and Response in Shoot and Root

Exogenous application of jasmonate is known to lead in large changes in the transcriptome. In order to filter GWAS candidates, we integrated jasmonate RNAseq transcriptomes from the public TENOR database into the context of the GWAS analysis (Kawahara et al. 2016). We identified 232 genes showing up- and down-regulation under exogenous JA treatments with changes greater than 2-fold or lower - 2-fold during the time-course (6 times points from T0 to 24 h) with a FDR threshold set at 0,05. Among these genes, 137 and 95 are regulated by JA in root and shoot respectively (Additional file 8: Table S6). Interestingly, the annotated JA-responsive genes showed function mainly in regulation of DNA transcription (Additional file 9: Table S7). It

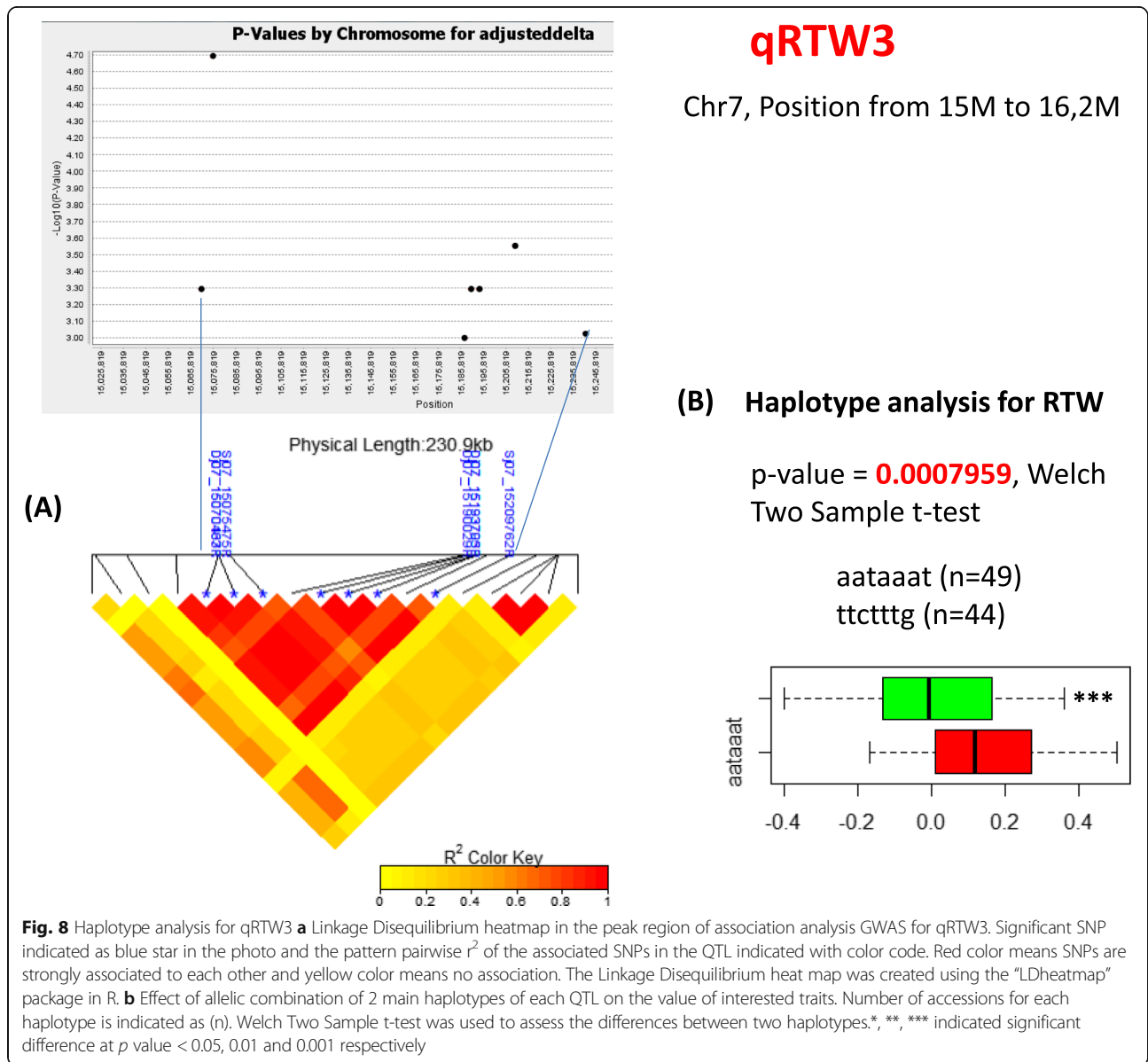


was previously reported that transcription factors (TFs) play important roles in regulating gene expression responses to jasmonate (Pré et al. 2008). To explore which types of TF were present in JA regulated genes, we used the PlantTFDB to classify TFs (Jin et al. 2017). We identify 12 and 7 TF in root and shoot respectively to be differentially regulated by JA belonging to 7 TF families (Tables 3 and 4). In addition, 5 TFs were JA-responsive both in root and shoot transcriptomes. Among 14 differentially expressed TFs, 5 C2H2, 1 BHLH, 1 ERF, 1 NAC, and 1 LBD were found up-regulated. Altogether these data suggest that several key signalling elements involved in the jasmonate response are associated with the different QTLs identified in this study.

Discussion

Natural Genetic Diversity of Rice Growth Inhibition in Response to Jasmonate

To better understand the role of JA in regulating complex growth response in rice and evaluate to which extent JA responses are modulated by genetic variability, we conducted a phenotyping analysis with 155 Vietnamese rice accessions based on a simple, robust and rapid phenotyping experiment in response to exogenous jasmonate. Jasmonate application has been widely used as a simple way to induce different types of defense response in plants, in general, and on rice, in particular. For example, Nahar and colleagues demonstrated that foliar application of exogenous JA triggered the systemic defense against the Root Knot Nematode (RKN) *M. graminicola* in rice



root (Nahar et al. 2011). In addition, exogenous treatment of JA induces PR genes such as *OsPR1a* and *OsPR1b* and protect the rice root from the RKN infection (Nahar et al. 2011, 2012). In this study, we applied 5 μ M of JA on plantlets growing in vitro during 7 days to score the effect on shoot and root growth. Seventeen plantlets per condition (with or without JA) for each genotype have been phenotyped for a total of 3094 rice plants showing significant repeatability and robustness. The use of 17 seedlings replicates per condition has improved the accuracy of the trait value (data not shown). Indeed, all traits showed a high heritability in our culture condition set up (Table 1). Recently, hormones treatments with auxin, cytokinin and abscisic acid have already been

used to evaluate natural variation that shapes root system responses in *Arabidopsis* (Ristova et al. 2018). As we observed from our root phenotyping experiments in response to JA, they identify common response patterns relating to particular hormone perturbation modulating by specific genotypes (this study Fig. 2 and (Ristova et al. 2018)). While many major jasmonate signaling elements have been identified as genes responsible for the jasmonate-related short-root phenotype in few *Arabidopsis* accessions, our results based on the haplotype analysis open the possibility to use unexplored rice accessions with contrasted phenotypes as parent lines to identify new alleles that control JA pathway involved in growth inhibition processes.

Table 2 Repartition of MapMan mapping of the ten over-represented gene ontologies

Second level mapman ontology code	Number of mapped genes	Name of ontology
27.3	51	DNA regulation of transcription
29.5	30	Protein degradation
29.4	26	Protein postranslational modification
30.2	20	Signalling receptor kinases
20.1	13	Stress biotic
29.2	12	Protein synthesis
33.99	12	Development unspecified
30.3	11	Signalling calcium
29.3	9	Protein targeting
20.2	8	Stress abiotic

GWAS and Candidate Genes

While many studies have analyzed jasmonate signaling and response, to the best of our knowledge no study has explored JA-dependent genetic mechanisms involved in these molecular processes taking into account natural genetic variation. In this study, we applied the GWAS approach to identify QTLs associated with plant growth modulation by JA treatment in an *Oryza sativa* rice collection. As a result, 28 significant associations, including 3 QTLs for RTL, 3 QTLs for RTW, 7 QTLs for SHL, 7 QTLs for SHW and 8 QTLs for TTW. The non-compressed MLM models, with an option of re-estimating the markers we used in this study, are well fitted and appropriate for analysis, as shown by the low rate of false positives in the LD linkage calculation we observed. Many common significant associations with high-log (*p*-value) were found for both the whole panel and *Indica* panel, such as qSHL1,

qRLT1, qTTW5, and qSHL3 (Additional file 5: Table S4). In addition, some QTLs were only found in the whole panel or in each sub-panel. It is likely that the limited size of the subpopulation, especially for the *Japonica* subpanel, which has only 55 accessions in the panel, could cause the decrease in the performance of GWAS analysis. Another hypothesis was already discussed in the study conducted by Phung et al in 2016. When dividing the panel into 2 subpanels, the number of polymorphic markers also decreased from 21,623 markers in the whole panel to 13,814 and 8821 markers for *Indica* and *Japonica*, respectively (Phung et al. 2016).

The GWAS analysis detected 28 QTLs with many genes annotated in the 25-kb window around a significant SNP. Based on MSU, RAPDB and MapMan annotation we were not able to identify genes already implicated in JA perception, biosynthesis and signalling.

Table 3 Expression pattern of rice transcription factors genes in response to jasmonate in root using RNAseq dataset from TENOR

RapDB ID	QTL	PlantTFDB annotation	Average of FC 1h	Average of FC 3h	Average of FC 6h	Average of FC 12h	Average of FC 1d
Os01g0129600	qTTW1	LBD	2,74	2,71	NS	3,07	10,14
Os01g0285300	qSHL1	MYB	NS	NS	-12,9	-36,08	-41,34
Os02g0705500	qRTL1	bHLH	2,3	NS	NS	NS	-2,14
Os02g0710300	qRTL1	bHLH	-2,64	-3,2	-83,06	-83,06	-22,88
Os03g0815100	qTTW4	NAC	28,21	33,74	44,38	31,21	32,8
Os03g0820300	qTTW4	C2H2	30,88	19,54	65,61	51,58	118,88
Os03g0820400	qTTW4	C2H2	10,81	7,13	23,08	12,67	15,56
Os03g0838800	qTTW4	C2H2	7,15	7,92	26,04	29,56	15,3
Os08g0160300	qSHL4	G2-like	NS	-7	-28,72	-7,8	-8,73
Os12g0582900	qSHW7/qTTW8	ERF	-11,91	-5,51	-4,11	-11,91	NS
Os12g0583700	qSHW7/qTTW8	C2H2	8,86	11,18	22,59	23,56	19,57
Os12g0586300	qSHW7/qTTW8	G2-like	-4,49	-28,66	-19,07	-43,08	-9,12

NS Not significant, FC Fold change

Table 4 Expression pattern of rice transcription factors genes in response to jasmonate in shoot using RNAseq dataset from TENOR

RapDB ID	QTL	PlantTFDB annotation	Average of FC 1h	Average of FC 3h	Average of FC 6h	Average of FC 12h	Average of FC 1d
Os01g0285300	qSHL1	MYB	NS	NS	-4,86	-11,17	-11,81
Os03g0815100	qTTW4	NAC	25,02	13,64	11,34	7,03	7,35
Os03g0820300	qTTW4	C2H2	5,38	2,71	NS	NS	NS
Os03g0838800	qTTW4	C2H2	NS	5,91	NS	NS	NS
Os08g0442400	qSHW4	ERF	NS	2,73	2,92	NS	7,21
Os12g0583700	qSHW7/qTTW8	C2H2	2,67	6,11	5,99	3,56	6,32
Os12g0586300	qSHW7/qTTW8	G2-like	-2,87	-3,71	-5,21	-5,47	-2,61

NS Not significant, FC Fold change

This observation could be explained by the long JA treatment we used to measure the growth parameter. Many JA biosynthesis genes as well as JAZs and TFs are known to play key function in the very early response and often play major roles where mutations can be detrimental in the wild. Another explanation could involve the presence of false negatives that are not identified by GWAS analysis.

Using available JA transcriptome we filtrated the 560 genes localized with the significant markers and found 42% of the candidate genes to be responsive to jasmonate, indicating that these genes are potential new players in jasmonate growth inhibition response pathway. Among this short JA-responsive gene list, we identified one Calmodulin-dependent protein kinases: OsSnRK1 β , encoded by *Os03g17980* located in Chromosome 3. In plants, energy depletion caused by stress is sensed and coordinated by an energy-sensing protein kinase, named sucrose non-fermenting-1-related protein kinase-1 (SnRK1). In a study conducted by Filipe et al. (2018), the overexpression of *OsSnRK1 α* confers broad-spectrum resistance in rice against various rice pathogens, including both hemibiotrophs (*Xoo PXO99*, *P. oryzae VT5M1*) and necrotrophs (*C. miyabeanus Cm988*, *R. solani AG1-1A* strain 16), while inhibiting the growth and development of rice. Interestingly, OsSnRK1 α activated the JA signaling and response pathway by boosting the expressions of a JA biosynthesis gene (*OsAOS2*) and JA-related genes such as *JIOsPR10* and *OsJAMYb* after inoculating the plant with *Pyricularia oryzae* (Filipe et al. 2018). Consequently, Filipe suggested that *OsSnRK1 α* plays a key role in the energy balance between plant defense and plant development (Filipe et al. 2018). Such a role should be researched for *OsSnRK1 β* , which we identified in a QTL associated with RTL inhibition.

In addition, other candidate genes that were found have a function related to secondary metabolism. For example, the 2 genes *Os01g18110* and *Os01g18120*, located in Chromosome 1, near the Sj01_10048970F marker, were annotated as Cinnamoyl CoA reductase (*OsCCR4*

and *OsCCR5* respectively), which are the first responsible enzymes specific to the pathway for lignin biosynthesis (Park et al. 2017). In *Cassia tora*, application of methyl jasmonate at 10 μ M promoted root sensitivity to aluminum-induced apoplastic peroxidase activity and H₂O₂ and lignin accumulation (Xue et al. 2008). In 2015, while studying the Scorpion peptide LqhIT2 in rice transgenic plants, Tianpei et al. (2015) found that the pathways downstream of LOX, such as the JA pathway, were activated and then lignin content was increased accordingly. They suggested that JA induce expression of genes involved in the phenylpropanoid pathway, leading to lignin accumulation in rice tissues (Taheri and Tarighi 2010). In 2011, while studying the induction of lignin biosynthesis induced by cell wall damage in *Arabidopsis*, Denness and his colleagues found that the plant regulates the biosynthesis of lignin through the interaction between the JA-dependent process and reactive oxygen species (Denness et al. 2011). A study on RICE SALT SENSITIVE3 protein encoded by *RSS3*- shown that this gene regulates root cell elongation under salinity condition (Toda et al. 2013). RSS3-JAZ-bHLH complex regulated jasmonate-induced gene expression involved in the cell wall metabolism including lignin and phenylpropanoids biosynthesis (Toda et al. 2013). *OsCCR4* and *OsCCR5* were found in relation to two root traits: the root mass and the RTL. Research using RiceXpro showed that *OsCCR5* was up-regulated after 6 h when roots of 7 day-old rice plantlets were treated with 100 μ M of JA. Consistent with the prediction of molecular process, this finding suggests that the effect of JA enhances the biosynthesis of some secondary macromolecules in roots such as lignin, making rice roots more resistant to adverse conditions. Indeed, we observed that the roots were not only shorter in response to JA but also less flexible compared to the untreated roots (data not shown). Diversity of root cell wall modification in response to JA will be explored further by metabolomics phenotyping of the rice collection.

The MapMan annotation of the candidate genes allowed to highlight a strong representation of genes involved in the

regulation of DNA transcription and signaling. The search for JA-responsive transcription factor via TFDB led to the identification of 12 TFs. Among them, several studies emphasized their regulation in response to stress like drought (*Os02g0705500*, (Minh-Thu et al. 2018)), excessive Fe in roots (*Os01g0129600*, (Bashir et al. 2014)), macronutrient deficiency (*Os12g0582900*, (Takehisa et al. 2015)), high temperature (*Os08g0442400*, (Endo et al. 2009)) and infections by *Xanthomonas oryzae* pv. *oryzae* (*Os12g0586300*, (Wang et al. 2019); *Os03g0820400*, (Yi et al. 2014)).

Functional analyzes were also conducted for the TFs *SNAC1* and *ZFP252* belonging respectively to the families NAC and C2H2. *SNAC1* is induced in guard cells by drought and enhances drought resistance in overexpression transgenic rice in the field under drought salt stress conditions (Hu et al. 2006). For *ZFP252*, it was also reported that this TF increased tolerance to salt and drought stresses by regulating the content of proline and soluble sugars under salt and drought stress (Xu et al. 2008). Unlike many TFs involved in stress tolerance responses, overexpression of *SNAC1* or *ZFP252*, does not lead to growth inhibition or yield penalty of transgenic plants (Hu et al. 2006; Xu et al. 2008). Transcriptome data sets from TENOR show that these two TFs are not only induced by JA in shoot but also in roots. Therefore, it would be interesting to determine whether the expression of these genes is differentially regulated in the contrasting genotypes for the JA response and to identify in roots their function in response to stresses.

Conclusion

The GWAS is a powerful tool to mine the valuable alleles hidden in the biodiversity richness of rice (Korte and Farlow 2013). This marks the first time that JA-related growth inhibition traits have been characterized in rice by GWAS. Identification of QTLs associated with the growth of rice treated with exogenous JA can help to dissect the underlying molecular mechanism controlling the defense/development trade off in rice. Further analysis should be conducted to functionally characterize these QTLs. We already develop a mapping segregating for four QTLs by generating a haplotype map. This analysis will be completed by using some accessions to re-sequencing at higher depth to confirm the results. Several promising candidate genes such as *OsSnRK1β* will be investigated for expression in contrasting accessions. The results of this research could help to create stress-tolerant rice accessions while maintaining plant growth.

Materials and Methods

Plant Materials

The collection used in the experiment was composed of 155 Vietnamese rice accessions whose seeds were provided by the Plant Resource Center in Hanoi, Vietnam. These accessions were mostly landrace lines collected from

various regions and ecosystems in Vietnam (Phung et al. 2014). The accession IR64 was added as internal controls in the phenotyping experiment. The GBS methods and the marker selection process, as described by Phung et al. (2014), were used to generate the genotypic data. The genotypic data consisted of 21,623 markers with no missing data that could be used for genome-wide association purposes.

Growth Condition

The rice collection was grown in vitro under axenic conditions in a programmable growth chamber (DAIHAN Scientific, Thermo Stable GC-450) under a 12 h daily light cycle at 80% humidity. The light intensity was 12,000 Lux. The temperature was adjusted to 26 °C and 28 °C during dark and light phases, respectively. Initially, the seeds were kept in the oven at 50 °C for 3 days to break dormancy. Then, they were decorticated manually and surface-sterilized by shaking with 70% ethanol for 2 min. After being rinsed with sterilized distilled water, the seeds were soaked in commercial bleach (3.8–4% sodium hypochlorite) with 2 drops of Tween 20 and were agitated for 25 min. Next, the seeds were washed 7–8 times with sterilized distilled water and left overnight in the dark at 26 °C to maximize water absorbance. Surface-sterilized seeds were sowed on an agar Petri dish at 6% w/v. After storing the Petri dish in a growth chamber for 1 day, each plantlet was transferred to a glass test tube with Murashige and Skoog (MS) medium at half concentration, pH = 5.8 (adjusted by potassium hydroxide), containing 2 g/L Phytigel (Sigma Aldrich), supplemented with or without jasmonic acid (TCI-Japan) at 5 μM for the phenotyping experiment. The tubes were subsequently transferred to a growth chamber, and the phenotyping process occurred 7 days after JA treatment.

Phenotyping Experiment

The experiment was conducted using an augmented randomized complete block design, with 7 blocks, each with 26 accessions such that the IR64 control was allocated in each of the 7 blocks and 25 others landraces accessions were allocated only once in the design (Boyle and Montgomery 1996). The experiment was performed in triplicates by dividing the experiment into three sub-blocks. Each sub-block had at least 5 to 6 replicates per condition and per accession. In each sub-block, the position of each replicate was randomized using IRRISTAT 4.0 software to avoid systematic effects due to positioning. For each plant, the length of the longest root (RTL) and the length of the shoot (SHL) were measured. After drying at 55 °C in the oven for 1 week, the weight of the whole plant (TTW) and of the shoot portion (SHW) were measured. The RTW was calculated by subtracting SHW from TTW. To ensure the homogenous germination rate, accessions having a low germination rate or slow germination time were discarded

from the study. In total, 155 rice accessions were assessed in this study.

Statistical Analysis

To analyze the phenotype data of the 155 Vietnamese rice accessions, analysis of variance (ANOVA) was performed to verify the effects of various blocks, genotypes and JA treatment. The block effect was calculated based on the phenotype of the IR64 controls. As the block effect was significant, the data were normalized by using the 'lme4' package in R. Next, the broad-sense heritability coefficient H^2 was calculated. Effects of JA on each trait were used as phenotype data in genome-wide association mapping. All statistical tests and analyses in studying the phenotype data of 155 Vietnamese rice accessions were run under R software version 3.4.3.

Genome-Wide Association Mapping

Association analyses were conducted on the whole panel having 155 rice accessions and a collection of 21,632 markers using Tassel v5.0 (Bradbury et al. 2007) to identify genetic variants associated with responses to exogenous JA treatment. From genotype data, a kinship matrix was generated by centered Identity by State (IBS) method proposed by TASSEL. To consider the structure of the collection derived from a previous study by Phung et al. (2014), a principle component analysis (PCA) with the top 6 components was also calculated from haplotype data, and an eigenvalue decomposition of the covariance matrix was executed, as recommended by Reich et al. (2008). Both the Generalized Linear Model (GLM) and non-compressed Mixed Linear Model (MLM) were applied, but we mainly used MLM with an option of re-evaluation of variance components for each marker to process the kinship matrix and a matrix combining both phenotype data and structure of the panel, producing a GWAS result. Then, the quantile-quantile plot (Q-Q plot) and Manhattan plot were drawn using TASSEL. The q -value corresponds to adjusted p -value after FDR analysis was computed to estimate the false discovery rate using package "qvalue" in R (Phung et al. 2016). However, in order to make the comparison within populations and across traits, we applied a less stringent p -value of $3.0E-4$ as suggestive threshold to declare that an association was significant.

Linkage Disequilibrium and QTL Selection

Subsequent to achieving a list of significant markers, a linkage disequilibrium heatmap for each marker was generated to confirm if this marker belonged to a quantitative-trait locus (QTL). The "LD heatmap" package allowed us to calculate pairwise linkage disequilibria between markers and to visualize the results (Shin et al. 2006). Only regions having at least one significant marker associated with nearby markers were considered as a QTL.

Screening for Annotated Genes and Transcriptome Analysis

The positions of found QTLs were then used to screen for candidate genes on the MSU Rice Genome Annotation Project Database, release 7.0 (Kawahara et al. 2013). Only expressed genes located around 25 kb before and after each significant marker were selected. A list of candidate genes with their annotation information was formulated.

For each of the 560 genes, we converted the MSU ID to RapDB ID using the RapDB ID converter (Sakai et al. 2013). These RapDB gene IDs allowed to map MapMan ontologies using the BinTree RAPDB-IRGSP1.0 version 1.0 (Thimm et al. 2004). Granularity of the bins was further reduced to the second level (e.g. 27.2: RNA.transcription) for readability. Also, RapDB gene IDs were used to import TF family annotation from PlantTFDB 4.0 mapping (Jin et al. 2017). In order to perform a cross analysis, the 560-gene list was used as a query into Tenor database (Kawahara et al. 2016) from which we pulled the FC and associated FDR values for all associated transcripts in response to JA in root and shoot. For each transcript and each time point, we retained the values presenting a FC $</> -2/2$ with FDR < 0.05 . For each gene and each time point, we then calculated the average FC and standard deviation, thus providing a list of genes both detected in the 50 kb interval of the QTLs of interest, and strongly and significantly differentially expressed in response to JA in root and/or shoot.

Haplotype Analysis

The significant markers of each LD block of interested QTL were used to define haplotype using alignment Nexus file. The 2 main haplotype was selected based on the geographic maps created by PopArt software version 1.7 (Population Analysis with Reticulate Trees software) (<http://popart.otago.ac.nz>) then we compare with the phenotype of each haplotype in order to confirm if the sequence variation in each QTL region can contribute to the phenotypic difference of interested traits.

Additional Files

Additional file 1: Figures S1, S2, S3 and Table S1. Growth inhibition of 10 representative accessions in response to JA. **Figure S1.** for histogram of distribution of each trait. **Figure S2.** presents the variation of 5 traits between 10 representative's accessions in non-treated and 5 μ M JA treated condition. **Figure S3.** illustrated the percentage reduction of each trait after JA treatment compare to the non-treatment. **Table S1.** expressed the heritability coefficient of each traits. (DOCX 969 kb)

Additional file 2: Table S2. List of 155 Vietnamese rice accessions used in this study with information on their gene bank number, their sub-populations groups as well as their ecosystems. (DOCX 52 kb)

Additional file 3: Table S3. Effects of 5 μ M JA on the phenotypic variation of the growth traits for 155 rice accessions. (CSV 12 kb)

Additional file 4: Figures S4, S5 and S6. GWAS for the effects of exogenous JA on RTL, SHW and TTW. Manhattan plot (A) and Quantile-

quantile plot (B) for RTL (**Figure S4**), SHW (**Figure S5**) and TTW (**Figure S6**) in a whole (S.x.1) panel or *Indica* (S.x.2) or *Japonica* (S.x.3) subpanel. The blue line indicates the suggestive significance threshold, $p = 3.0E-04$. Black rectangle represent common significant SNPs within panels. (DOCX 1060 kb)

Additional file 5: Table S4. List of signification markers of detected QTLs at p -value $< 3.0E-04$. The start (site 1) and end positions (site 2) of each QTL were estimated by expanded 25 kb to both terminals base on the LD linkage analysis. Chr for chromosome, RTW for root dry weight, SHW for shoot dry weight, RTL for root length, SHL for shoot length. (XLSX 24 kb)

Additional file 6: Figure S7. Haplotype analysis for qTTW5. (A) Linkage Disequilibrium heatmap in the peak region of association analysis GWAS for qRTW3. Significant SNP indicated as blue star in the photo and the pattern pairwise r^2 of the associated SNPs in the QTL indicated with color code. Red color means SNPs are strongly associated to each other and yellow color means no association. The Linkage Disequilibrium heatmap was created using the "LDheatmap" package in R (B) Population architecture of accessions based on the allelic combination significant SNPs in each QTL. Population architecture image is created by PopArt1.7 software. (C) Effect of allelic combination of 2 main haplotypes of each QTL on the value of interested traits. Number of accessions for each haplotype is indicated as (n). Welch Two Sample t-test was used to assess the differences between two haplotypes. *, **, *** indicated significant difference at p value < 0.05 , 0.01 and 0.001 respectively. (PPTX 116 kb)

Additional file 7: Table S5. List of candidate genes found ± 25 kb around the significant markers and MapMan and TFDB annotations. (XLSX 2416 kb)

Additional file 8: Table S6. Expression pattern of candidate genes in response to jasmonate in shoot and root using RNAseq dataset from TENOR. (XLSX 20 kb)

Additional file 9: Table S7. MapMan annotation of the JA-responsive candidate genes identified in the TENOR transcriptome database. (XLSX 384 kb)

Acknowledgements

The authors would like to thank "Rice Functional Genomics and Plant Biology" International Joint Laboratory (LMI-RICE2) for supporting us the processing charge. Our special thanks to CRP-RICE for supporting us the English proofreading service. We would like to thank Dr. Yves Vigouroux for insightful comments to the manuscript. We also thank to all colleagues: Dr. Hien TT Vu, Dr. Khanh D. Tran, Dr. Chung D. Mai, Hanh T. Kieu, Anh M. Ngo, Anh L. Nguyen, Thi Tho Nguyen and Toan V. Nguyen to provide significant contributions to the phenotyping experiments.

Authors' Contributions

HTMT and AC lead the research and designed the study. ML and PG contributed to the establishment of the rice collection. All authors performed the phenotyping experiments. AC and JL performed the transcriptome analysis. HTMT and AC analyzed the data. HTMT wrote the manuscript. All authors have read and approved the manuscript for publication.

Funding

This research is funded by the Vietnam National Foundation for Science and Technology Development (NAFOSTED) under grant number 106-NN.03-2016.15 to Huong TM To.

Availability of Data and Materials

The data sets supporting the results of this article are included within the article and its supporting files.

Ethics Approval and Consent to Participate

Not applicable.

Consent for Publication

Not applicable.

Competing Interests

The authors declare that they have no competing interests.

Author details

¹University of Science and Technology of Hanoi (USTH), Vietnam Academy of Science and Technology (VAST), LMI-RICE2, 18 Hoang Quoc Viet, Cau Giay district, Hanoi, Vietnam. ²Institut de Recherche pour le Développement (IRD), Université de Montpellier, UMR DIADE, UMR IPME, UMR LSTM, Montpellier, France. ³Agricultural Genetics Institute, LMI-RICE2, Hanoi, Vietnam.

Received: 9 May 2019 Accepted: 22 August 2019

Published online: 04 September 2019

References

- Bashir K, Hanada K, Shimizu M et al (2014) Transcriptomic analysis of rice in response to iron deficiency and excess. *Rice* 7:1–15. <https://doi.org/10.1186/s12284-014-0018-1>
- Boyle CR, Montgomery RD (1996) An application of the augmented randomized complete block design to poultry research. *Poult Sci* 75:601–607. <https://doi.org/10.3382/ps.0750601>
- Bradbury P, Zhang Z, Kroon D et al (2007) TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* 23(19):2633–2635
- Bui B (2010) Rice germplasm conservation in Vietnam. In: Vietnam, fifty years of rice research and development. Hanoi: Agriculture Publishing House, pp 167–178
- Bùi CB, Nguyễn TL (2003) The genetic basis of tolerance to environmental damage in rice. Agriculture Publishing House, TP Hồ Chí Minh
- Cai Q, Yuan Z, Chen M et al (2014) Jasmonic acid regulates spikelet development in rice. *Nat Commun* 5:3476. <https://doi.org/10.1038/ncomms4476>
- Campos ML, Yoshida Y, Major IT et al (2016) Rewiring of jasmonate and phytochrome B signalling uncouples plant growth-defense tradeoffs. *Nat Commun* 7:1–10. <https://doi.org/10.1038/ncomms12570>
- Chan EKF, Rowe HC, Kliebenstein DJ (2010) Understanding the evolution of defense metabolites in *Arabidopsis thaliana* using genome-wide association mapping. *Genetics* 185:991–1007. <https://doi.org/10.1534/genetics.109.108522>
- Chini A, Fonseca S, Fernández G et al (2007) The JAZ family of repressors is the missing link in jasmonate signalling. *Nature* 448:666–671. <https://doi.org/10.1038/nature06006>
- Courtois B, Hong NH, Pham VH et al (1997) Genetic diversity of traditional varieties of upland rice from Vietnam and prospects offered by improved varieties. *Agric D ev* 15:163–167
- Denness L, McKenna JF, Segonzac C et al (2011) Cell wall damage-induced lignin biosynthesis is regulated by a reactive oxygen species- and jasmonic acid-dependent process in *Arabidopsis*. *Plant Physiol* 156:1364–1374. <https://doi.org/10.1104/pp.111.175737>
- Endo M, Tsuchiya T, Hamada K et al (2009) High temperatures cause male sterility in rice plants with transcriptional alterations during pollen development. *Plant Cell Physiol* 50:1911–1922. <https://doi.org/10.1093/pcp/pcp135>
- Feyls B, Benedetti CE, Penfold CN, Turner JG (1994) *Arabidopsis* mutants selected for resistance to the phytotoxin coronatine are male sterile, insensitive. *Society* 6:751–759
- Filipe O, De Vleeschauwer D, Haec A et al (2018) The energy sensor OsSnRK1a confers broad-spectrum disease resistance in rice. *Sci Rep* 8:1–12. <https://doi.org/10.1038/s41598-018-22101-6>
- Gasperini D, Ch etelat A, Acosta IF et al (2015) Multilayered organization of jasmonate signalling in the regulation of root growth. *PLoS Genet* 11:1–27. <https://doi.org/10.1371/journal.pgen.1005300>
- Hakata M, Muramatsu M, Nakamura H et al (2017) Overexpression of TIFY genes promotes plant growth in rice through jasmonate signaling. *Biosci Biotechnol Biochem* 81:906–913. <https://doi.org/10.1080/09168451.2016.1274638>
- Hickman R, Van Verk MC, Van Dijken AJH et al (2017) Architecture and dynamics of the jasmonic acid gene regulatory network. *Plant Cell* 29:2086–2105. <https://doi.org/10.1105/tpc.16.00958>
- Hoang GT, Van Dinh L, Nguyen TT et al (2019) Genome-wide association study of a panel of Vietnamese rice landraces reveals new QTLs for tolerance to water deficit during the vegetative phase. *Rice* 12. <https://doi.org/10.1186/s12284-018-0258-6>
- Hori Y, Kurotani KI, Toda Y et al (2014) Overexpression of the JAZ factors with mutated jas domains causes pleiotropic defects in rice spikelet development. *Plant Signal Behav* 9:1–7. <https://doi.org/10.4161/15592316.2014.970414>
- Howe GA, Major IT, Koo AJ (2018) Modularity in jasmonate signaling for multistress resilience. *Annu Rev Plant Biol* 69:1–29. <https://doi.org/10.1146/annurev-arplant-042817-040047>

- Hu H, Dai M, Yao J et al (2006) Overexpressing a NAM, ATAF, and CUC (NAC) transcription factor enhances drought resistance and salt tolerance in rice. *Proc Natl Acad Sci* 103:12987–12992. <https://doi.org/10.1073/pnas.0604882103>
- Huot B, Yao J, Montgomery BL, Yang S (2014) Growth – defense tradeoffs in plants : a balancing act to optimize fitness. *Mol Plant*:1267–1287. <https://doi.org/10.1093/mp/ssu049>
- Jin J, Tian F, Yang DC et al (2017) PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. *Nucleic Acids Res* 45:D1040–D1045. <https://doi.org/10.1093/nar/gkw982>
- Kashihara K, Onohata T, Okamoto Y et al (2019) Overexpression of OsNINJA1 negatively affects a part of OsMYC2-mediated abiotic and biotic responses in rice. *J Plant Physiol* 232:180–187. <https://doi.org/10.1016/j.jplph.2018.11.009>
- Katsir L, Schilmiller AL, Staswick PE et al (2008) COI1 is a critical component of a receptor for jasmonate and the bacterial virulence factor coronatine. *Proc Natl Acad Sci* 105:7100–7105. <https://doi.org/10.1073/pnas.0802332105>
- Kawahara Y, De Bastide M, Hamilton JP et al (2013) Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data, pp 1–10
- Kawahara Y, Oono Y, Wakimoto H et al (2016) TENOR: database for comprehensive mRNA-Seq experiments in rice. *Plant Cell Physiol* 57:e7. <https://doi.org/10.1093/pcp/pcv179>
- Kazan K, Manners JM (2013) MYC2: the master in action. *Mol Plant* 6:686–703. <https://doi.org/10.1093/mp/sss128>
- Khan GA, Vogiatzaki E, Glauser G, Poirier Y (2016) Phosphate deficiency induces the Jasmonate pathway and enhances resistance to insect herbivory. *Plant Physiol* 171:632–644. <https://doi.org/10.1104/pp.16.00278>
- Korte A, Farlow A (2013) The advantages and limitations of trait analysis with GWAS: a review. *Plant Methods* 9:29. <https://doi.org/10.1186/1746-4811-9-29>
- Lee HY, Seo J, Cho JH, et al (2013) *Oryza sativa* COI Homologues Restore Jasmonate Signal Transduction in *Arabidopsis* coi-1 Mutants. 8:1–9. <https://doi.org/10.1371/journal.pone.0052802>.
- Li X, Guo Z, Lv Y et al (2017) Genetic control of the root system in rice under normal and drought stress conditions by genome-wide association study. *PLoS Genet* 13:1–24. <https://doi.org/10.1371/journal.pgen.1006889>
- Lorenzo O, Chico J, Sánchez-Serran J, Solano R (2004) JASMONATE-INSENSITIVE1 encodes a MYC transcription factor essential to discriminate between different jasmonate-regulated defense responses in arabidopsis. Author(s): Oscar Lorenzo, Jose M. Chico, Jose J. Sánchez-Serrano and Roberto Solano Published. *Plant Cell* 16:1938–1950. <https://doi.org/10.1105/tpc.022319> with
- Minh-Thu P-T, Kim JS, Chae S et al (2018) A WUSCHEL homeobox transcription factor, osWOX13, enhances drought tolerance and triggers early flowering in rice. *Mol Cells* 41:781–798. <https://doi.org/10.14348/molcells.2018.0203>
- Nahar K, Kyndt T, De Vleeschouwer D et al (2011) The jasmonate pathway is a key player in systemically induced defense against root knot nematodes in rice. *Plant Physiol* 157:305–316. <https://doi.org/10.1104/pp.111.177576>
- Nahar K, Kyndt T, Hause B et al (2012) Brassinosteroids suppress rice defense against root-knot nematodes through antagonism with the jasmonate pathway. *Mol Plant-Microbe Interact* 26:106–115. <https://doi.org/10.1094/mpmi-05-12-0108-fi>
- Noir S, Bomer M, Takahashi N et al (2013) Jasmonate controls leaf growth by repressing cell proliferation and the onset of endoreduplication while maintaining a potential stand-by mode. *Plant Physiol* 161:1930–1951. <https://doi.org/10.1104/pp.113.214908>
- Ogawa S, Kawahara-Miki R, Miyamoto K et al (2017a) OsMYC2 mediates numerous defence-related transcriptional changes via jasmonic acid signalling in rice. *Biochem Biophys Res Commun* 486:796–803. <https://doi.org/10.1016/j.bbrc.2017.03.125>
- Ogawa S, Miyamoto K, Nemoto K et al (2017b) OsMYC2, an essential factor for JA-inductive sakuranetin production in rice, interacts with MYC2-like proteins that enhance its transactivation ability. *Sci Rep* 7:1–11. <https://doi.org/10.1038/srep40175>
- Okada K, Abe H, Arimura GI (2015) Jasmonates induce both defense responses and communication in monocotyledonous and dicotyledonous plants. *Plant Cell Physiol* 56:16–27. <https://doi.org/10.1093/pcp/pcu158>
- Park HL, Bhoo SH, Kwon M et al (2017) Biochemical and expression analyses of the Rice cinnamoyl-CoA reductase gene family. *Front Plant Sci* 8:1–14. <https://doi.org/10.3389/fpls.2017.02099>
- Pauwels L, Gemma FB, Jan G et al (2010) NINJA connects the co-repressor TOPLESS to jasmonate signalling. *Nature* 464:788–791. <https://doi.org/10.1038/nature08854>
- Phung NTP, Mai CD, Mournet P et al (2014) Characterization of a panel of Vietnamese rice varieties using DArT and SNP markers for association mapping purposes. *BMC Plant Biol* 14:1–16. <https://doi.org/10.1186/s12870-014-0371-7>
- Phung TPN, Mai CD, Hoang GT et al (2016) Genome-wide association mapping for root traits in a panel of rice accessions from Vietnam. *BMC Plant Biol*. <https://doi.org/10.1186/s12870-016-0747-y>
- Pré M, Atallah M, Champion A et al (2008) The AP2/ERF domain transcription factor ORA59 integrates Jasmonic acid and ethylene signals in plant defense. *Plant Physiol* 147:1347–1357. <https://doi.org/10.1104/pp.108.117523>
- Reich D, Price AL, Patterson N (2008) Principal component analysis of genetic data. *Nat Genet* 40:491–492
- Ristova D, Giovannetti M, Metesch K, Busch W (2018) Natural genetic variation shapes root system responses to phytohormones in *Arabidopsis*. *Plant J* 96:468–481. <https://doi.org/10.1111/tpj.14034>
- Sakai H, Lee SS, Tanaka T et al (2013) Rice annotation project database (RAP-DB): an integrative and interactive database for rice genomics. *Plant Cell Physiol* 54. <https://doi.org/10.1093/pcp/pcs183>
- Shin J-H, Blay S, Graham J, McNeney B (2006) LDheatmap: an R function for graphical display of pairwise linkage disequilibria between single nucleotide polymorphisms. *J Stat Softw* 16. <https://doi.org/10.18637/jss.v016.c03>
- Sun J, Xu Y, Ye S et al (2009) *Arabidopsis* ASA1 is important for jasmonate-mediated regulation of auxin biosynthesis and transport during lateral root formation. *Plant Cell Online* 21:1495–1511. <https://doi.org/10.1105/tpc.108.064303>
- Ta KN, Khong NG, Ha TL et al (2018) A genome-wide association study using a Vietnamese landrace panel of rice (*Oryza sativa*) reveals new QTLs controlling panicle morphological traits. *BMC Plant Biol* 18:1–15. <https://doi.org/10.1186/s12870-018-1504-1>
- Taheri P, Tarighi S (2010) Riboflavin induces resistance in rice against *Rhizoctonia solani* via jasmonate-mediated priming of phenylpropanoid pathway. *J Plant Physiol* 167:201–208. <https://doi.org/10.1016/j.jplph.2009.08.003>
- Takehisa H, Sato Y, Antonio BA, Nagamura Y (2015) Global transcriptome profile of rice root in response to essential macronutrient deficiency global transcriptome profile of rice root in response to essential macronutrient deficiency © 2013 Landes Bioscience. Do Not Distribute 2324:4–10. <https://doi.org/10.4161/psb.24409>
- Thimm O, Blasing O, Gibon Y et al (2004) MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *Plant J* 37:914–939. <https://doi.org/10.1111/j.1365-313X.2004.02016.x>
- Tianpei X, Li D, Qiu P, et al (2015) Scorpion peptide LqhIT2 activates phenylpropanoid pathways via jasmonate to increase rice resistance to rice leafrollers. *Plant Sci* 230:1–11. <https://doi.org/10.1016/j.plantsci.2014.10.005>.
- Toda Y, Tanaka M, Ogawa D et al (2013) RICE SALT SENSITIVE3 forms a ternary complex with JAZ and class-C bHLH factors and regulates jasmonate-induced gene expression and root cell elongation. *Plant Cell* 25:1709–1725. <https://doi.org/10.1105/tpc.113.112052>
- Uji Y, Akimitsu K, Gomi K (2017) Identification of OsMYC2-regulated senescence-associated genes in rice. *Planta* 245:1241–1246. <https://doi.org/10.1007/s00425-017-2697-5>
- Uji Y, Taniguchi S, Tamaoki D, et al (2016) Overexpression of OsMYC2 results in the upregulation of early JA-responsive genes and bacterial blight resistance in rice. *Plant Cell Physiol* 57:1814–1827. <https://doi.org/10.1093/pcp/pcw101>.
- Vos IA, Pieterse CMJ, Van Wees SCM (2013) Costs and benefits of hormone-regulated plant defences. *Plant Pathol* 62:43–55. <https://doi.org/10.1111/ppa.12105>
- Wang C, Tariq R, Ji Z et al (2019) Transcriptome analysis of a rice cultivar reveals the differentially expressed genes in response to wild and mutant strains of *Xanthomonas oryzae* pv. *oryzae*. *Sci Rep* 9:1–13. <https://doi.org/10.1038/s41598-019-39928-2>
- Wasternack C (2007) Jasmonates: an update on biosynthesis, signal transduction and action in plant stress response, growth and development. *Ann Bot* 100:681–697. <https://doi.org/10.1093/aob/mcm079>
- Wasternack C, Hause B (2013) Jasmonates: biosynthesis, perception, signal transduction and action in plant stress response, growth and development. An update to the 2007 review in annals of botany. *Ann Bot* 111:1021–1058. <https://doi.org/10.1093/aob/mct067>
- Wray N, Visscher P (2008) Estimating trait heritability. *Nat Educ* 1:29
- Wu H, Ye H, Yao R et al (2015) OsJAZ9 acts as a transcriptional regulator in jasmonate signaling and modulates salt stress tolerance in rice. *Plant Sci* 232:1–12. <https://doi.org/10.1016/j.plantsci.2014.12.010>

- Xu DQ, Huang J, Guo SQ et al (2008) Overexpression of a TFIIIA-type zinc finger protein gene ZFP252 enhances drought and salt tolerance in rice (*Oryza sativa* L.). *FEBS Lett* 582:1037–1043. <https://doi.org/10.1016/j.febslet.2008.02.052>
- Xue YJ, Tao L, Yang ZM (2008) Aluminum-induced cell wall peroxidase activity and lignin synthesis are differentially regulated by jasmonate and nitric oxide aluminum-induced cell wall peroxidase activity and lignin synthesis are differentially regulated by jasmonate and nitric oxide. *Society*:9676–9684. <https://doi.org/10.1021/jf802001v>
- Yamada S, Kano A, Tamaoki D et al (2012) Involvement of OsJAZ8 in jasmonate-induced resistance to bacterial blight in rice. *Plant Cell Physiol* 53:2060–2072. <https://doi.org/10.1093/pcp/pcs145>
- Yan J, Li S, Gu M et al (2016) Endogenous bioactive jasmonate is composed of a set of (+)-7- iso- JA-amino acid conjugates. *Plant Physiol* 172:2154–2164. <https://doi.org/10.1104/pp.16.00906>
- Yang D-L, Yao J, Mei C-S et al (2012) Plant hormone jasmonate prioritizes defense over growth by interfering with gibberellin signaling cascade. *Proc Natl Acad Sci U S A* 109:E1192–E1200. <https://doi.org/10.1073/pnas.1201616109>
- Ye H, Du H, Tang N et al (2009) Identification and expression profiling analysis of TIFY family genes involved in stress and phytohormone responses in rice. *Plant Mol Biol* 71:291–305. <https://doi.org/10.1007/s11103-009-9524-8>
- Yi SY, Lee HY, Kim HA et al (2014) Microarray analysis of bacterial blight resistance 1 mutant rice infected with *Xanthomonas oryzae* pv. *oryzae*. *Plant Breed Biotechnol* 1:354–365. <https://doi.org/10.9787/pbb.2013.1.4.354>

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Submit your manuscript to a SpringerOpen[®] journal and benefit from:

- Convenient online submission
- Rigorous peer review
- Open access: articles freely available online
- High visibility within the field
- Retaining the copyright to your article

Submit your next manuscript at ► [springeropen.com](https://www.springeropen.com)

A.2 Paper in submission

This is the work during my Bachelor year in 2016. We are able to submit this long experimental work after years of preparing transgenic materials and performing experiment.

Paper submitted to Journal of Experimental Botany on July 11th, 2022.

Title: Silencing of Plant Defensin Type 1 reduces sensitivity to pathogens and zinc toxicity in *Arabidopsis thaliana*

Nga Ngoc, Thi Nguyen, Olivier Lamotte, Mohanad Alsulaiman, Sandrine Ruffel, Gabriel Krouk, **Dang Nguyet**, Sébastien Aimé, Pierre Berthomieu, Christian Dubos, David Wendehenne, Denis Vile, and Françoise Gosti

A.3 Paper in preparation

In the 2nd year of my PhD, I am learning new concept related to graph. Nothing is better than "learning by doing". Therefore, I am working with my collaborators from Phenikaa University and Hanoi University of Pharmacy to apply my graph knowledge into studying network pharmacology. In this research, I work mostly to construct a knowledge graph containing information from different aspects such as: traditional plant medicine, bioactive compounds, disease targets and so on. The manuscript is in preparation and we will soon submit in the 3rd quarter of the year.

Title: Toward multi-component, multi-target therapeutic of Chronic obstructive pulmonary disease: *In silico* study of *Iris domestica* and *Mimosa pudica* combination using molecular docking, molecular dynamics simulation, binding free energy calculation and network pharmacology

Authors: **Nguyet Thi-Minh Dang**, Tuan Ngoc Do, Nhung Thi-Hong Duong, Anh Tuan Pham, Chi Quynh Nguyen

Appendix B

Conferences

B.1 Oral presentation

International Genome Graph Symposium 2022, Switzerland

Title: PARROT: PAngenome gRaph Related Output Transmutator



PARROT

Pangenome gRaph Related Output Transmutator

Nguyet DANG

International Genome Graph Symposium

Monte Verità, July 2022

1

OR

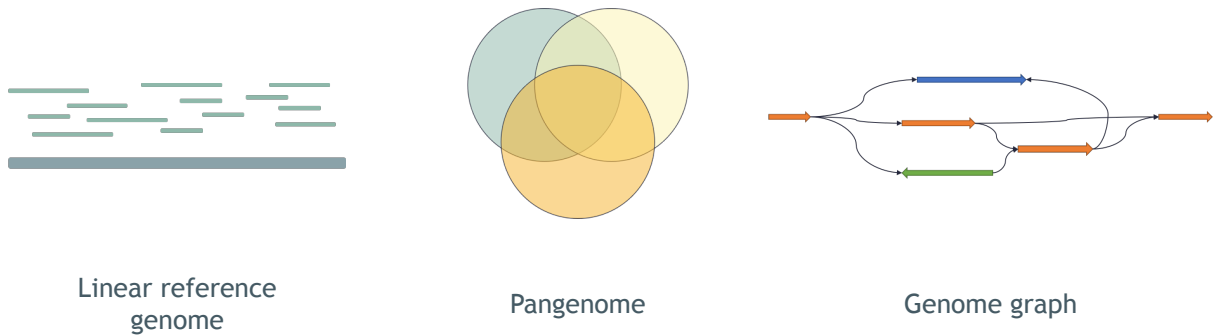
The story of how I end up in developing tools for genome graph and come to the symposium

- *PARROT*: Pangenome gRaph Related Output Transmutator
- *BioGraph.jl*: Julia package to extract the longest path
- *GraphInfer*: Structural variation inference based on graph genome for skimming data

1

2

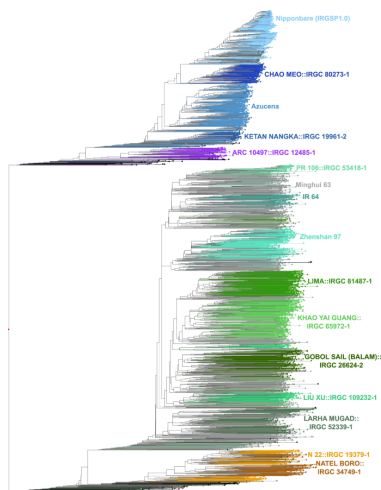
Trend in genome research



2

3

!!! Question from our biologist colleagues



Phylogenetic tree of Asian rice accessions
(Zhou et al., 2020)

- Where are the position of my gene of interest in different varieties?
- Graph is a mess, I want a linear reference!
- I don't have sequencing at high coverage, can I benefit from the pangenome and the graph genome?

3

4

Strategies

Step 1: Genome graph construction

Step 2: Longest representative path identification

Step 3: Individual skimming sequence mapping with currently available tools

Step 4: Individual genome structure inference

The shape of the graph was inspired from sevenbridges.com

4

5

Graph genome construction - minigraph

Graph 1 (asm 1)

Align asm 2 to graph 1

Construct graph 1/2

Coarse graph 1/2

Align asm 3 to graph 1/2

Construct graph 1/2/3

Coarse graph 1/2/3

Genome graph construction with minigraph

(Heng Li, minigraph github)

Input:

- Oryza sativa Nipponbare reference
- 12 nearly gap free reference genomes of Asian rice (Zhou et al., 2020)

Output: A genome graph in rGFA format

```

S s1 CTGAA SN:Z:chr1 SO:i:0 SR:i:0
S s2 ACG SN:Z:chr1 SO:i:5 SR:i:0
S s3 TGGC SN:Z:chr1 SO:i:8 SR:i:0
S s4 TGTGA SN:Z:chr1 SO:i:12 SR:i:0
S s5 TTTC SN:Z:foo SO:i:0 SR:i:11
S s6 CTGA SN:Z:foo SO:i:12 SR:i:11
S s7 GTTAC SN:Z:bar SO:i:5 SR:i:2
L s1 + s2 + 0M
L s2 + s3 + 0M
L s3 + s4 + 0M
L s2 + s5 + 0M
L s5 + s6 + 0M
L s6 + s4 + 0M
L s1 + s7 - 0M
L s7 - s6 + 0M
                    
```

rgfa format example

(Heng Li, gfatools github)

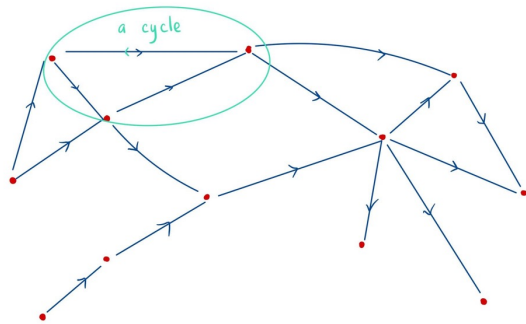
Statistical information provided by gfatools:

- Number of segments: 252540
- Number of links: 357504
- Number of arcs: 715008
- Max rank: 12
- Total segment length: 548979827
- Average segment length: 2173.833
- Sum of rank-0 segment lengths: 374422835
- Max degree: 10
- Average degree: 1.416

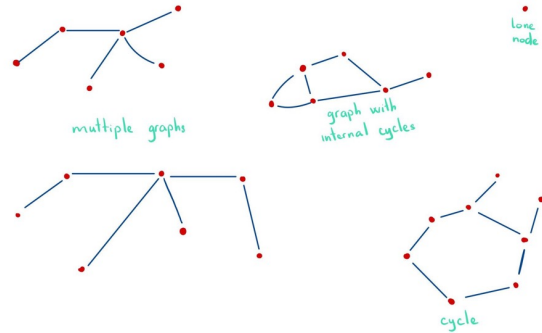
There is somethings missing... 5

6

Issues 1 - Multiple graph form



We are still dealing with this issue



We separate the information in rGFA into smaller simple graphs

6

7

BioGraph.jl - Graph component identification

- Lone Node: only have one vertices
- Lone Cycle: has no source vertices or sink vertices
- Simple Graph: others sub graph induced from rGFA

Algorithm 1: Graph Components of GFA

Result: Graph Components of GFA

init GFA input *GFA*

init the *start_nodes*

init the *end_nodes*

Find all *source_nodes*

Find all *sink_nodes*

Find all weakly connected components *g_com* of GFA

for *com* in *g_com* **do**

if *length(com)* = 1 **then**
 └ Classify as Lone Node

else
 Make sub graph induced from GFA
 Find sub graph *sg_source_nodes* and *sg_sink_nodes*
 if *sg_source_node* = [] or *sg_sink_nodes* = [] **then**
 └ Classify as Lone Cycle

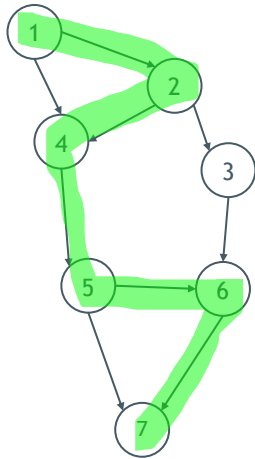
else
 └ Classify as Simple Graph

return *g_com* with classification

7

8

Issues 2 - How the longest path is extracted?

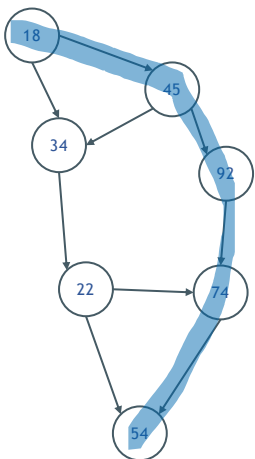


Longest path = the path from starting point to the ending point through the maximal number of nodes
=> Non - weighted graph

8

9

Issues 2 - How the longest path is extracted?



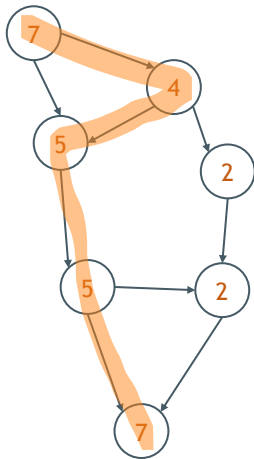
Longest path = the path from starting point to the ending point through the maximal number of nodes
=> Non - weighted graph

Longest path = the path containing the highest number of base pairs
=> Weighted graph (weight value = the length of sequences in each node)

8

10

Issues 2 - How the longest path is extracted?



Longest path = the path from starting point to the ending point through the maximal number of nodes
=> Non - weighted graph

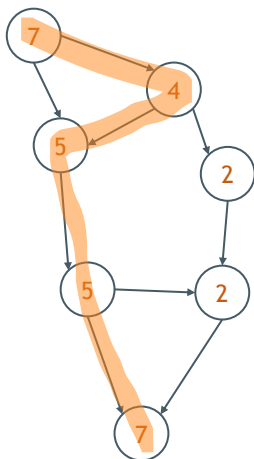
Longest path = the path containing the highest number of base pairs
=> Weighted graph (weight value = the length of sequences in each node)

Longest representative path = the path containing nodes appearing in the most number of individuals
=> Weighted graph (weight value = the number of individuals having that node)

8

11

Issues 2 - How the longest path is extracted?



Longest path = the path from starting point to the ending point through the maximal number of nodes
=> Non - weighted graph

Longest path = the path containing the highest number of base pairs
=> Weighted graph (weight value = the length of sequences in each node)

Longest representative path = the path containing nodes appearing in the most number of individuals
=> Weighted graph (weight value = the number of individuals having that node)

Based on the the strategy to set weight value, differnt paths can be obtained.

8

12

PARROT - Weight value calculation

GAF format when remap reference genomes against obtained graph

```
Os128077RS1_Ctg40      50227  44305  47115  +      >s104335>s104336>s104337>s104339>s104340
Os128077RS1_Ctg40      50227  1103    6038   +      >s147586>s21792>s21793>s164408
```

Presence/absence matrix

SegName	Start	End	SegID	AzucenaRS1	IRGSP-1	Os117425RS1	Os125619RS1	Os125827RS1	Os127518RS1
IRGSP-1.0_Chr1	0	27974	>s1	1	1	1	3	1	
IRGSP-1.0_Chr1	27974	37449	>s2	1	1	1	2	1	
IRGSP-1.0_Chr1	37449	37505	>s3	1	1	1	2	1	
IRGSP-1.0_Chr1	37505	44126	>s4	1	1	1	2	1	
IRGSP-1.0_Chr1	44126	44282	>s5	1	1	1	2	1	
IRGSP-1.0_Chr1	44282	54574	>s6	1	1	1	2	1	
IRGSP-1.0_Chr1	54574	57838	>s7	1	1	1	1	2	
IRGSP-1.0_Chr1	57838	79333	>s8	1	1	1	2	1	
IRGSP-1.0_Chr1	79333	79362	>s9	1	1	1	1	1	

Answer: Is a segment of interest in an individual appearing in another one?

Weight value of each stable sequence corresponding to the number of individuals having that sequence

```
s146039 12
s146041 12
s146043 5
s146045 5
s163242 9
s146048 8
s163243 2
s146050 10
s163244 4
s146052 10
s163245 3
s146054 8
s163246 5
```

<https://github.com/nguyetdang/PARROT>

9

13

BioGraph.jl - Longest path extraction

$$A = \begin{bmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,n_v} & a_{1,n_v+1} \\ a_{2,1} & a_{2,2} & \dots & a_{2,n_v} & a_{2,n_v+1} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{n_v,1} & a_{n_v,2} & \dots & a_{n_v,n_v} & a_{n_v,n_v+1} \\ a_{n_v+1,1} & a_{n_v+1,2} & \dots & a_{n_v+1,n_v} & a_{n_v+1,n_v+1} \end{bmatrix}$$

$$W = \begin{bmatrix} w_{1,1} & w_{1,2} & \dots & w_{1,n_v} & w_{1,n_v+1} \\ w_{2,1} & w_{2,2} & \dots & w_{2,n_v} & w_{2,n_v+1} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ w_{n_v,1} & w_{n_v,2} & \dots & w_{n_v,n_v} & w_{n_v,n_v+1} \\ w_{n_v+1,1} & w_{n_v+1,2} & \dots & w_{n_v+1,n_v} & w_{n_v+1,n_v+1} \end{bmatrix}$$

$$\begin{aligned} \text{find } & \mathbf{A} & (1) \\ \text{maximize } & \sum a_{i,j} w_{i,j} & (2) \\ \text{subject to } & a_{i,j} = 0 \text{ if } a_{i,j} \notin E & (3) \\ & a_{ij} = 0 \text{ or } 1 \text{ if } a_{i,j} \in E & (4) \\ & \sum_{i \in \mathcal{S}_1} \sum_{j \notin \mathcal{S}_1} a_{i,j} = 1 & (5) \\ & \sum_{i \notin \mathcal{S}_2} \sum_{j \in \mathcal{S}_2} a_{i,j} = 1 & (6) \\ & \sum_{i \in \mathcal{S}_2} a_{i,n_v+1} = 1 & (7) \\ & \sum_{i \notin \mathcal{S}_1} a_{i,j} \leq 1 \text{ for all } j \notin \mathcal{S}_1 \cup n_{v+1} & (8) \\ & \sum_{i \notin \mathcal{S}_1} a_{i,j} - \sum_{k \notin \mathcal{S}_1} a_{j,k} = 0 \text{ for all } j \notin \mathcal{S}_1 \cup n_{v+1} & (9) \end{aligned}$$

10

14

Application on Asian rice genomes

BioGraph.jl-Get_summary (whole rGFA)

- Simple graphs: 509
- Lone cycles: 0
- Lone nodes: 45

BioGraph.jl-Get_summary (simple_graph_1)

- Vertices: 28771
- Edges: 40682
- Source_node: 8
- Sink_node: 8
- Path: 0

BioGraph.jl-Extract longest path (simple_graph_509)

```
["s247120\t-", "s119008\t-"]
```

```
>Graph509
ATCGTCTAATTTTATGGGCAATAGAATTTTGTTCGGCCCAAGTTCAGCC
GCCCACTTTTATTTCTGGTCCAACATATTTTCGTTTATTTTGTTCCTGG
CCCGAAGCCGATGAAATTACACAACACAACACTTTTTTCCCTGGAGAATAT
CAAATTAACAGATGGAAATCTATGTATCAGAGACAAGACATCCTTTAACTTT
ATAATTTACGTTAAACTGCAGAAATATATCATGTATATCATATAATTGGTAAT
AAATATTTCTCTCTCTCCTGAAGAGCTTGGCTCAATCAGATTTAAATATTT
TTCTAAGCTCTACATATTTTCTGAATTTAACTGAACCTAACACATGCAAAATC
```

Header	Start	End	Segment	Direction
Graph509	1	93	s247120	-
Graph509	94	7330	s119008	-

Size of *Oryza sativa* Nipponbarre: ~ 362M

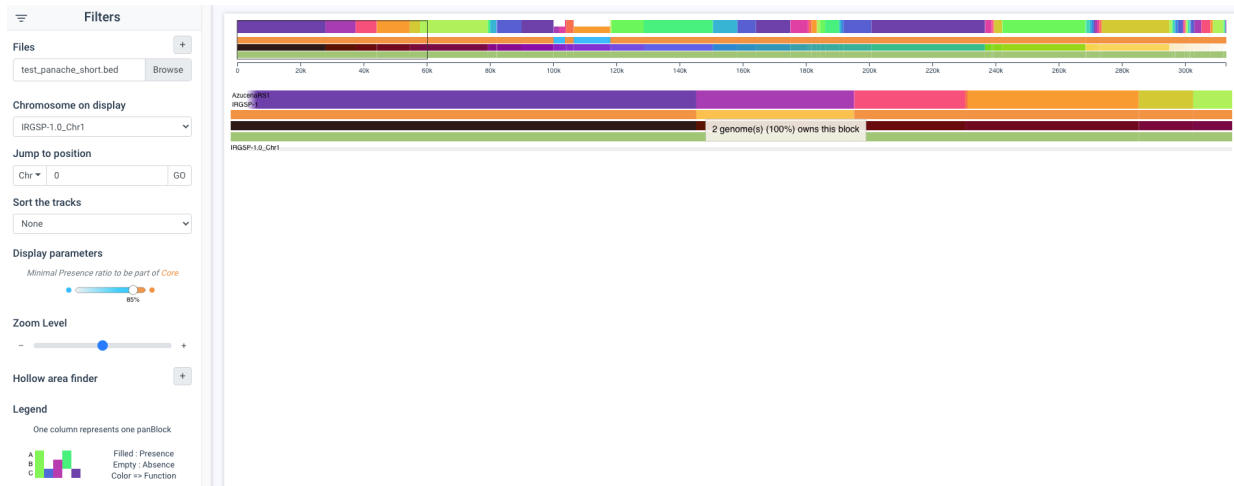
Size of the longest representative path size: ~ 413M

<https://github.com/nguyetdang/BioGraph.jl>

11

15

PARROT + PANACHE - Visualization



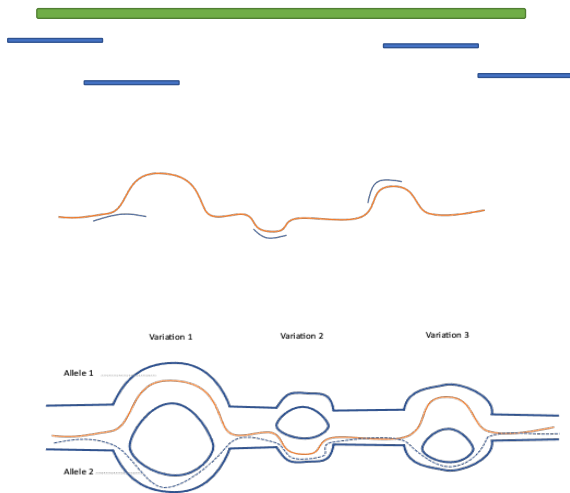
<https://github.com/SouthGreenPlatform/panache>

Developed by Éloi Durant and colleagues

12

16

Genome structure inference



Skimming data + *Oryza sativa*
Nipponbare reference genome: only SNPs
 data were in retrieved

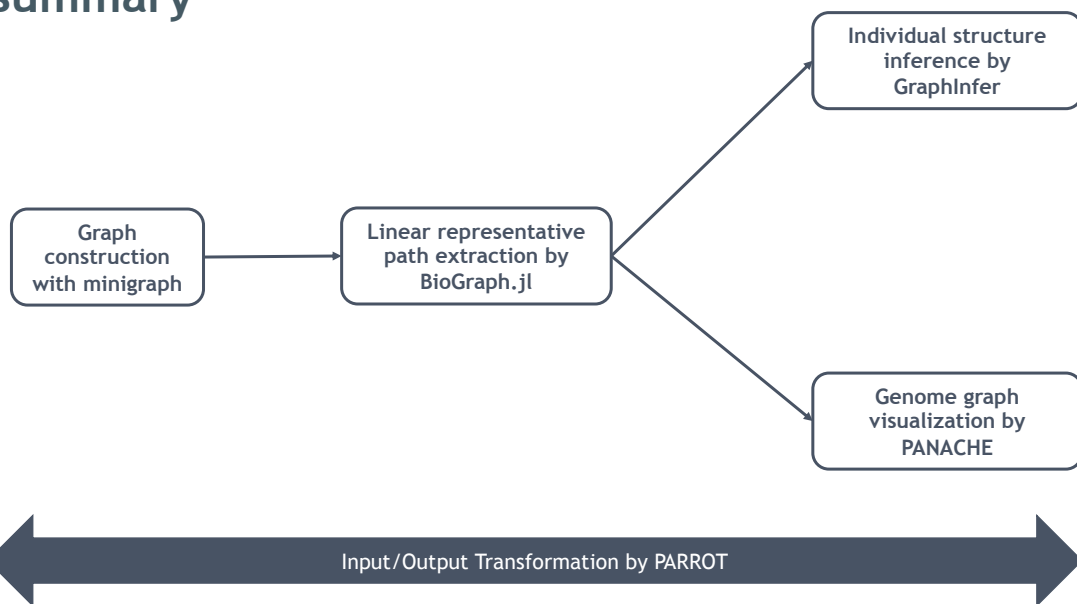
Map skimming data against longest
 representative path
 Segment having read mapped will obtain
 weight value

Using weight value to finding the longest
 path on the graph containing those
 mapped segments

13

17

In summary



14

18

How to access our study?

Tools:

<https://github.com/nguyetdang/BioGraph.jl>

<https://github.com/nguyetdang/PARROT>

GraphInfer: in development

Manual: Prepare to be released

Final release date : September 20th, 2022

2:00 PM CET

IRD Montpellier

Yes! It is the date & time of my thesis defense! You are all invited if you would like a more details and updates!



In case you cannot come to Montpellier, Zoom session will be available.

If you have questions related to the visualization, the guy developing PANACHE will present his work on the 29th September

15

19

Acknowledgement

Funding



Research unit



Development team



Francois SABOT

Éloi Durant



Tuan Do


16


20

B.2 Poster presentation


JOBIM 2020, Montpellier


Title: Artificial intelligence approach for predicting variations in low-coverage NGS sequences





Biodiversité
Agriculture
Alimentation
Environnement
Terre
Eau






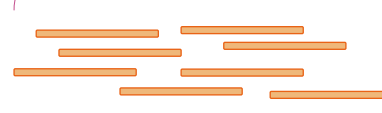
Artificial intelligence approach for predicting variations in low-coverage NGS sequences

Nguyet DANG and François SABOT


How to identify variation in low-coverage sequences?



High-coverage sequences



Low-coverage sequences



Vietnamese rice collection: 172 samples mostly at 2X

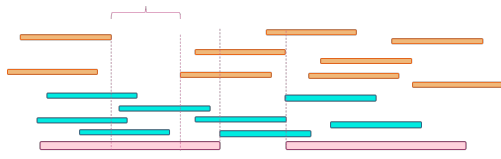
Difficult to identify structural variations

- Real variants?
- Or not yet be sequenced?

1

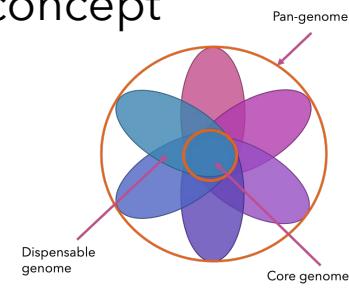
Single reference and pangenome concept

Variants identification



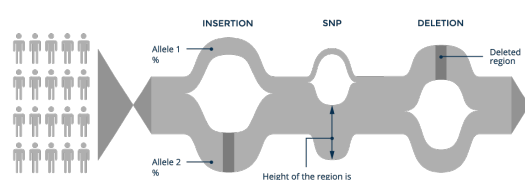
█ Reference
█ Individual 1
█ Individual 2

Reads for genomic part not available in reference genome become unmapped



Components of a pangenome

Using reference genome proposing advantages and inconvenients



Height of the region is proportional to allele frequency <https://www.sevenbridges.com/graph/>

Genome graph concept

The use of only a reference genome as a template for structural variation analysis is not sufficient, especially for low-coverage sequencing data. Hence, we would like to propose analysis approaches using pangenome and genome graph concepts with the help of machine learning algorithms for structural variation analysis.

2

Map-then-assemble approach

	ind 1	ind 2	ind 3	ind 4
Pos 1	1	1	0	0
Pos 2	0	1	0	0
Pos 3	1	1	1	0
Pos 4	1	1	0	1

	ind X
Pos 1	1
Pos 2	NA
Pos 3	1
Pos 4	1

Presence/absence matrix of a pangenome from high-coverage data

- Subcollection from 3K Rice Genome
- 12 rice reference genomes

2X samples from Vietnamese rice collection

Low-coverage sequenced individual's profile

Classification / Neural network + Single matrix decomposition methods

	ind X
Pos 1	1
Pos 2	0
Pos 3	1
Pos 4	1

Estimated null value

Mechanism for obtaining pangenome from map-then-assemble approach

Current progress: Testing models with PAV matrix of Chromosome 1 of Nipponbarre reference genome

3

Study structural variations with genome graph

Graph 1 (asm 1):

Align asm 2 to graph 1

Construct graph 1/2

Coarse graph 1/2:

Align asm 3 to graph 1/2

Construct graph 1/2/3

Coarse graph 1/2/3:

Mechanism to obtain genome graph from minigraph

<https://github.com/lh3/minigraph>

2 test cases of the longest path:

Longest path = the path containing the highest number of basepairs

Longest path = the path going from starting point to the ending point through the maximal number of nodes

Longest path

Low-coverage sample reads

Low coverage sample graph

The longest path shows the mutual path of all individuals in the studied population. Hence, in case the low coverage sample reads in the longest path, they are the same among individuals.

Reference graph

Reads not mapped against the longest path will be compared with the reference graph. Classification algorithm is used to identify which path the studied individual should follow.

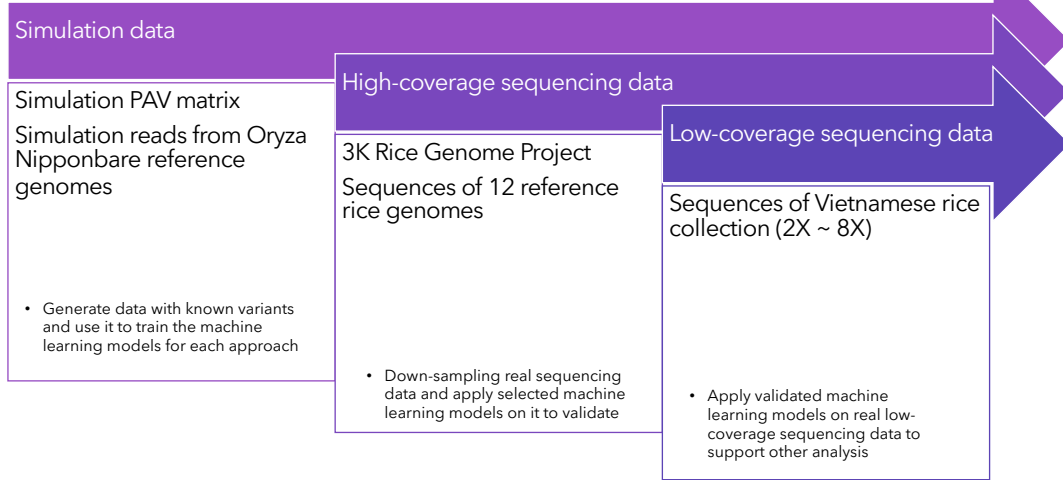
Combine information from two previous steps to obtain full graph of studied samples

Final graph is obtained.

Current progress: Testing different methods for the longest path

4

Data use in the study



Current progress: Working on simulation data and preparing high-coverage sequencing data for validation.



I'm working hard to learn !!!

See you in JOBIM 2021 for more updates !!!

JOBIM 2021, Paris

Title: BioGraph: A Julia package to extract the longest representative path from a pangenome graph

BioGraph: A Julia package to extract the longest representative path from a pangenome graph

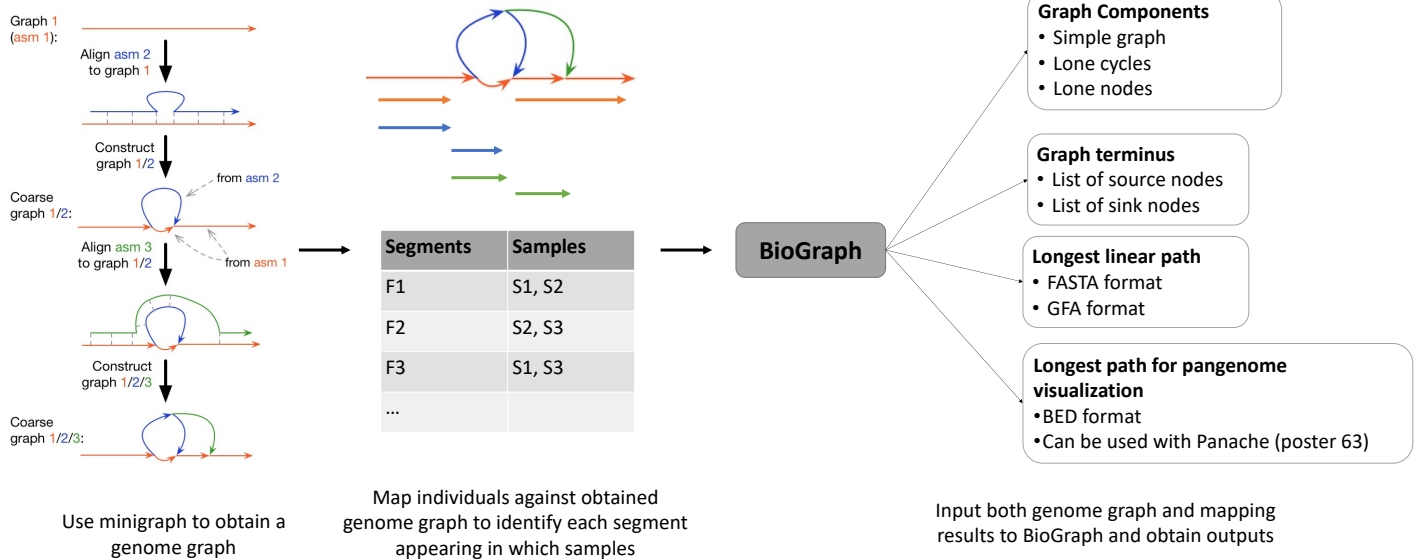
Nguyet DANG¹, Tuan DO² and François SABOT¹ | Contact: thi-minh-nguyet.dang@ird.fr

¹DIADÉ, Univ Montpellier, IRD, Montpellier, France, ²N2TP Technology Solutions JSC., Hanoi, Vietnam

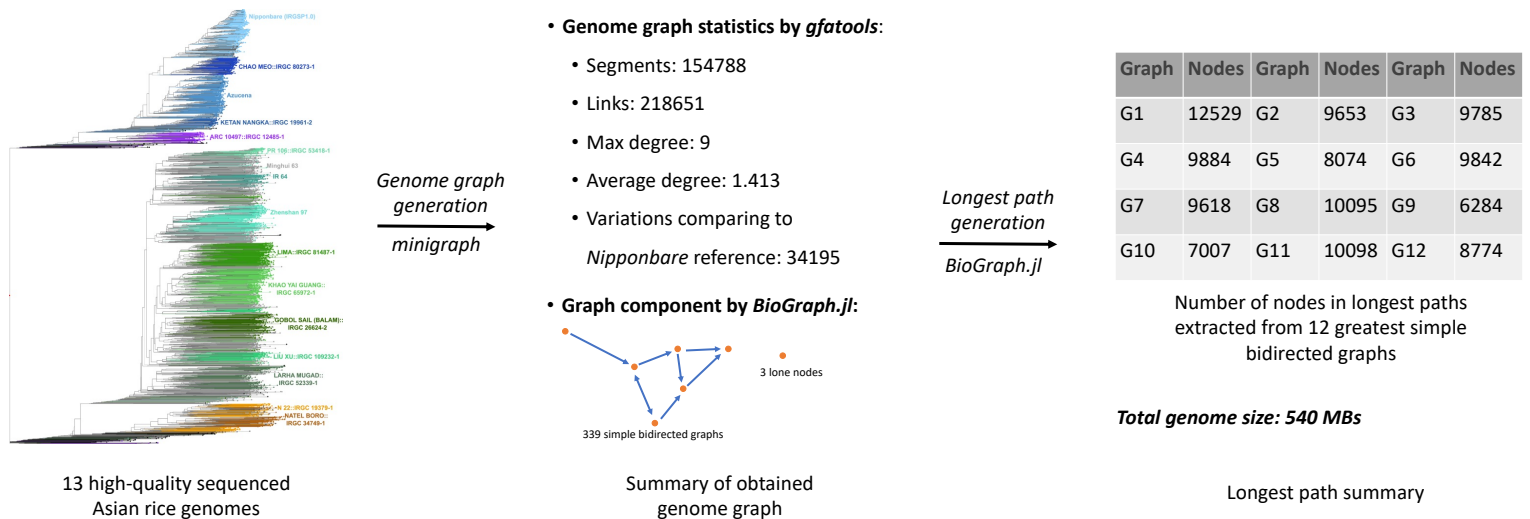
Introduction

To compare multiple genomes, a linear reference genome was often used as a coordination system to describe genes, variations and other functional annotations across individuals. However, this single reference was shown not to be sufficient to grasp every existing genomic variation such as copy number variations (CNV), presence/absence variations (PAV) or more general structural variations (SV) [1]. To overcome this limitation, the concept pangenome composing a core-genome and a dispensable genome was applied to investigate a group of genomes [2]. Graph-based data model generated by incrementally incorporating genome-to-graph alignment information was one of the novel approaches to represent pangenome information [3]. Here, we propose a Julia package to extract the longest representative path from a pangenome graph that are usable for available tools working with linear reference genome while conserving the additional information provided by a graph.

Workflow of BioGraph



Application on Asian rice genomes



Short computing time

Being able to run in parallel thanks to optimizers

Integration of weighted values

Presence/absence variations can be used in analysis

Visualization ability

Output can be used in Panache visualization tool

Dependencies

- | | |
|--|---|
| Optimizer | Julia packages |
| <ul style="list-style-type: none"> • CPLEX – IBM Optimizer • Gurobi Optimizer • CBC Optimizer | <ul style="list-style-type: none"> • LightGraph • GraphIO • ParserCombinator • JuMP |

Perspectives

BioGraph.jl still have some limitations that need to overcome. Firstly, the package can only show correct results with input data from minigraph. For graph from other sources, it is necessary to assure that there is no overlap among graph segments. Secondly, cycles inside inputted graph are eliminated by restricting each nodes appeared only once in the graph. We are working to find a practical approach to identify and investigate cycles in the graph. At the moment, we are working on integrating input from other genome graph generation tools. We are trying to develop a CLI version of the package so that it can be more user-friendly. In the next releases, we are finding an approach to identify the longest path obtained from a P-line of the GFA file.

References

- [1] Xiaofei Yang, Wan-Ping Lee, Kai Ye, and Charles Lee. One reference genome is not enough. *Genome Biology*, 20(1):104, 2019.
- [2] Christine Tranchant-Dubreuil, Mathieu Rouard, and François Sabot. Plant Pangenome: Impacts On Phe- notypes And Evolution. *Annual Plant Reviews*, May 2019.
- [3] Heng Li, Xiaowen Feng, and Chong Chu. The design and construction of reference pangenome graphs with minigraph. *Genome Biology*, 21(1):265, 2020.
- [4] Yong Zhou, Dmytro Chebotarov, Dave Kudrna, Victor Llaca, Seunghee Lee, Shanmugam Rajasekar, Nahed Mohammed, Noor Al-Bader, Chandler Sobel-Sorenson, Praveena Parakkal, Lady Johanna Arbelaez, Natalia Franco, Nickolai Alexandrov, N. Ruairadh Sackville Hamilton, Hei Leung, Ramil Mauleon, Mathias Lorieux, Andrea Zuccolo, Kenneth McNally, Jianwei Zhang, and Rod A. Wing. A platinum standard pan-genome resource that represents the population structure of asian rice. *Scientific Data*, 7(1):113, 2020.



Biodiversité
Agriculture
Alimentation
Environnement
Terre
Eau



Appendix C

Trainings

RÉCAPITULATIF DE PARTICIPATION AUX FORMATIONS

THI MINH NGUYET DANG

Doctorat : Génétique et génomique

Ecole Doctorale : GAIA - Biodiversité, Agriculture, Alimentation, Environnement, Terre, Eau

Etablissement : Université de Montpellier

Date de la 1ere inscription en thèse : 1 juillet 2019 (3 A en 2021)

Directeur de thèse : François SABOT

Sujet de thèse : Méthodes d'inférence des variations structurelles à l'échelle du génome dans les données séquençage basse profondeur à l'aide du graphe de pangéome

Formations suivies

Catégorie : Communication

- ▣ Delivering presentations to a non-scientific audience (16 septembre 2020 - 17 septembre 2020) EN LIGNE - EN DISTANCIEL
14 heures
- ▣ Exploit the scientific literature (12 octobre 2020) Université de Montpellier
14 heures
- ▣ How to effectively communicate with non-specialists (21 octobre 2020) Université de Montpellier
21 heures

Total du nombre d'heures pour la catégorie Communication : 49 h

Catégorie : Enseignement à distance

- ▣ MOOC Ethique de la recherche (17 septembre 2019 - 28 novembre 2019) En ligne
15 heures

Total du nombre d'heures pour la catégorie Enseignement à distance : 15 h

Catégorie : Langues vivantes

- ▣ FLE - Français Langue Étrangère (session février 2020) (03 février 2020 - 29 avril 2020) Université de Montpellier, Département des Langues Bâtiment 5
26 heures
- ▣ FLE - Français Langue Étrangère (session septembre 2019) (30 septembre 2019) Université de Montpellier, Département des Langues Bâtiment 5
18 heures Note : B1
- ▣ FLE - Français Langue Étrangère (session septembre 2020)/French as a Foreign Language (28 septembre 2020) Université de Montpellier, Département des Langues Bâtiment 5
26 heures

Total du nombre d'heures pour la catégorie Langues vivantes : 70 h

Catégorie : Outils et méthodes

- ▣ Writing your thesis effectively : from unstructured ideas to an organised text (01 décembre 2020 - 3 décembre 2020) Université de Montpellier
20 heures

Total du nombre d'heures pour la catégorie : 20 h

Total participation : 154 heures / 8 modules