



**HAL**  
open science

# Bioinformatic analysis of the apple genome and epigenome

Nicolas Daccord

► **To cite this version:**

Nicolas Daccord. Bioinformatic analysis of the apple genome and epigenome. Agricultural sciences. Université d'Angers, 2018. English. NNT : 2018ANGE0034 . tel-02129746

**HAL Id: tel-02129746**

**<https://theses.hal.science/tel-02129746>**

Submitted on 15 May 2019

**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.

# COUVERTURE TEMPORAIRE

## Thèse de doctorat de

L'UNIVERSITE D'ANGERS (1)

Comue Université Bretagne Loire

Ecole Doctorale n° 600

Ecole doctorale Ecologie, Géosciences, Agronomie et Alimentation

Spécialité : « (voir liste des spécialités) » (3)

« Analyse bioinformatique du génome et de l'épigénome  
du pommier » (5)

Thèse présentée et soutenue à « Angers », le « date » (6)

Unité de recherche : IRHS – INRA Angers(7)

Thèse N° : (8)

### Rapporteurs avant soutenance :

Thierry Lagrange DR1 Laboratoire Génome et Développement des plantes - CNRS  
Hélène Chiapello IR – Unité Mathématiques et Informatiques appliquées du Génome à l'Environnement (MaIAGE)

### Composition du Jury :

*Attention, en cas d'absence d'un des membres du Jury le jour de la soutenance, la composition du jury doit être revue pour s'assurer qu'elle est conforme et devra être répercutée sur la couverture de thèse*

|                    |                       |  |
|--------------------|-----------------------|--|
| Président :        | Prénom Nom            | Fonction et établissement d'exercice (9)(à préciser après la soutenance) |
| Examineurs :       | Clémentine Vitte      | Chercheur – Génétique quantitative et évolution – Le Moulon              |
|                    | Stéphanie Sidibe-Bocs | Chercheur – CIRAD - Montpellier  |
| Dir. de thèse :    | Etienne Bucher        | Directeur de recherches – IRHS INRA Angers                               |
| Co-dir. de thèse : | Claudine Landès       | Enseignant-chercheur – IRHS INRA Angers                                  |

### Invité(s)

Prénom Nom Fonction et établissement d'exercice

# Remerciements

Je voudrais tout d'abord remercier mon directeur de thèse Etienne Bucher pour sa disponibilité et son implication tout le long des trois années de ma thèse. Son optimisme et ses conseils m'auront permis de grandement améliorer la qualité des résultats que j'ai produit. Je remercie également ma co-directrice Claudine Landès pour ses conseils en Bioinformatique et sans qui je n'aurais pas engagé cette thèse.

Cette thèse n'aurait pas été possible sans le financement de l'université d'Angers et du projet GRIOTE. Je les remercie pour cette opportunité. Je voudrais également remercier l'IRHS ainsi que l'INRA pour m'avoir accueilli dans leurs locaux.

Je remercie également Mme Hélène Chiapello et Mr Thierry Lagrange pour avoir accepté de rapporter ce mémoire de thèse. J'associe à ces remerciements Mme Clémentine Vitte et Mme Stéphanie Sidibe-Bocs qui ont accepté d'être examinatrices.

J'adresse mes remerciements à Jean-Marc Celton avec qui j'ai longuement travaillé et échangé. Nos nombreuses discussions m'ont fait gagner beaucoup d'expérience afin de communiquer avec d'autres biologistes dans le futur, et m'auront permis d'améliorer les outils que j'ai développé durant ma thèse. Je voudrais également remercier Sébastien Aubourg pour son aide et son expertise qui non seulement m'ont permis de produire les résultats présentés dans ce manuscrit mais aussi de gagner beaucoup de connaissances théoriques. Merci également à Benjamin Istace pour ses conseils sur l'assemblage. Je remercie Sylvain Gaillard pour son aide technique et sa disponibilité qui auront été indispensables à ces travaux. Je remercie également le reste de l'équipe BioInfo pour leur accueil et les échanges que j'ai pu avoir avec eux.

Je tiens à remercier mes deux co-doctorants, Kay et Adrien, ainsi que le reste de l'équipe EpiCenter. Je remercie également l'ensemble des personnes avec lesquelles j'ai collaboré.

Plus personnellement, je remercie Erwan, Francois, Matthieu, Benjamin, David pour leur bonne humeur. Je remercie également mes parents pour leurs questions et leur curiosité incessante. Enfin, je remercie Mathilde pour ses encouragements.

# Contents

|   |           |
|---|-----------|
| <b>1 State of the art</b>                                 | <b>11</b> |
| 1.1 Genome sequencing                                     | 11        |
| 1.1.1 First generations of sequencing technologies        | 11        |
| 1.1.2 Last generation : long reads                        | 12        |
| 1.2 Genome assembly                                       | 13        |
| 1.2.1 Types of assembly algorithms                        | 13        |
| 1.2.2 Challenges in assembling repetitive elements        | 15        |
| 1.2.3 Handling the high error rates of long reads         | 17        |
| 1.2.4 Assembly scaffolding                                | 19        |
| 1.3 State of sequenced genomes                            | 21        |
| 1.3.1 Plant genomes                                       | 21        |
| 1.3.2 Rosaceae genomes                                    | 22        |
| 1.3.3 The apple ( <i>Malus Domestica</i> ) genome         | 23        |
| 1.3.4 Thesis objectives                                   | 24        |
| 1.4 Epigenetics   | 26        |
| 1.4.1 DNA Methylation                                     | 26        |
| 1.4.2 Methylation studies with NGS : bisulfite sequencing | 26        |
| 1.4.3 Differential methylation analysis                   | 28        |
| 1.4.4 Thesis objectives                                   | 31        |
| <b>2 Genome assembly and annotation</b>                   | <b>33</b> |
| 2.1 Introduction  | 33        |
| 2.2 Methods   | 34        |
| 2.2.1 Sequencing (done by collaborators)                  | 34        |
| 2.2.2 Genome assembly                                     | 37        |



|          |   |           |
|----------|---|-----------|
| 2.2.3    | Genome annotation   | 39        |
| 2.2.4    | Genome synteny  | 39        |
| 2.3      | Results   | 40        |
| 2.3.1    | GDDH13 V1.0   | 40        |
| 2.3.2    | GDDH13 V1.1   | 44        |
| 2.4      | Discussion  | 48        |
| 2.5      | Conclusion  | 50        |
| <b>3</b> | <b>DNA methylation in apple and DMRs</b>                    | <b>53</b> |
| 3.1      | Introduction  | 53        |
| 3.2      | Methods   | 53        |
| 3.3      | Results   | 55        |
| 3.3.1    | The apple methylome   | 55        |
| 3.3.2    | DNA methylation and fruit development                       | 55        |
| 3.3.3    | Correlations between methylation and expression             | 58        |
| 3.4      | Discussion  | 60        |
| 3.4.1    | The apple methylome   | 60        |
| 3.4.2    | DNA methylation and fruit development                       | 60        |
| 3.4.3    | Correlations between methylation and expression             | 61        |
| 3.5      | Conclusion  | 61        |
| <b>4</b> | <b>Differentially Methylated Regions detection pipeline</b> | <b>64</b> |
| 4.1      | Introduction  | 64        |
| 4.2      | Methods   | 64        |
| 4.2.1    | Programming language  | 64        |
| 4.2.2    | Pipeline global description                                 | 65        |
| 4.2.3    | Bisulfite reads mapping                                     | 65        |
| 4.2.4    | Global individual statistics and comparisons                | 65        |
| 4.2.5    | DMRs computing  | 65        |
| 4.2.6    | DMRs filtering  | 67        |
| 4.3      | Results   | 68        |
| 4.3.1    | Determining differential methylation on target regions      | 68        |
| 4.3.2    | DMRs between tissues in apple                               | 68        |
| 4.4      | Discussion  | 70        |

|          |   |           |
|----------|---|-----------|
| 4.4.1    | Biological validation of DMRs                       | 70        |
| 4.4.2    | Working with few biological replicates              | 70        |
| 4.4.3    | Selection of the regions to compare                 | 70        |
| 4.4.4    | DMRs filters  | 71        |
| 4.5      | Conclusion  | 71        |
| <b>5</b> | <b>Side projects</b>                                | <b>72</b> |
| 5.1      | <i>Tisochrysis lutea</i> genome assembly            | 72        |
| 5.1.1    | Introduction  | 72        |
| 5.1.2    | Methods   | 72        |
| 5.1.3    | Results   | 73        |
| 5.1.4    | Discussion and conclusion                           | 73        |
| 5.2      | Participation on the rose genome sequencing project | 74        |
| 5.2.1    | Introduction  | 74        |
| 5.2.2    | Methods   | 74        |
| 5.2.3    | Results   | 74        |
| 5.2.4    | Discussion and conclusion                           | 75        |

# List of Figures

|  |    |
|--|----|
| 1.1 Evolution of the sequencing technologies   | 12 |
| 1.2 Principle of PacBio sequencing   | 13 |
| 1.3 Illustration of de Bruijn graph-based assembly   | 14 |
| 1.4 Illustration of an Overlap-Layout-Consensus assembly algorithm                                   | 16 |
| 1.5 Illustration of repeat resolving using short and long reads                                      | 17 |
| 1.6 PacBio reads error rate illustration   | 19 |
| 1.7 Illustration of contigs scaffolding  | 19 |
| 1.8 BioNano workflow illustration  | 20 |
| 1.9 Evolution of the volume of deposited genomic data  | 21 |
| 1.10 Summary of Rosaceae phylogeny and Rosaceae fruit morphologies                                   | 23 |
| 1.11 Evolutionary history of the cultivated apple  | 25 |
| 1.12 DNA methylation basics  | 27 |
| 1.13 Bismark's approach to bisulfite mapping and methylation calling                                 | 28 |
| 1.14 Methylation calling process illustration  | 29 |
| 1.15 Venn diagrams comparing the number of DMLs in DMRs detected by diffmer, BSsmooth and<br>RADMeth | 31 |
| 2.1 Assembly and validation of the GDDH13 doubled-haploid apple genome                               | 42 |
| 2.2 Synteny and distribution of genomic and epigenomic features of the apple genome                  | 45 |
| 2.3 Screenshots of a gene affected by genomic sequence errors in GDDH13 V1.0                         | 46 |
| 2.4 Results of the multiple rounds of Pilon polishing  | 47 |
| 2.5 Screenshot of a gene prediction error (gene splitting) in GDDH13 V1.1                            | 51 |
| 3.1 DNA methylation landscape of the GDDH13 genome   | 56 |
| 3.2 Differentially methylated regions between apple tree leaves and young fruits                     | 57 |
| 3.3 Results of average methylation analysis in gene promoters  | 59 |

|   |    |
|---|----|
| 3.4 Results of the methylation patterns analysis  | 63 |
| 4.1 Global schema of the bisulfite sequencing analysis pipeline   | 66 |
| 4.2 Schema of the DMRs identification process   | 67 |
| 4.3 Description of the biological replicates merging process  | 68 |
| 4.4 Methylation levels modifications observations in <i>Arabidopsis</i> with $\alpha$ -amanitine and zebularine treatments                              | 69 |
| 5.1 k-mer (19 bp) spectra of the <i>Tisochrysis lutea</i> genome  | 73 |
| 5.2 Illustration of the cutting of the contigs in case of genetic markers inconsistency   | 74 |
| 5.3 Graphical representation of the rose genome pseudo-molecules  | 76 |
| S1 GDDH13 obtention process   | 81 |
| S2 Comparison of the genetic and physical GDDH13 position of the SNP markers of the integrated genetic map on the new genome                            | 83 |
| S3 Comparison of the genetic and physical GDDH13 position of the SNP markers of the integrated genetic map on the "old" genome ([Velasco et al., 2010]) | 84 |
| S4 Histogram showing the number of mapped sRNAs of 21, 22, 23 and 24 nucleotides on genes and transposable elements of the GDDH13 genome                | 85 |
| S5 Genomic DNA methylation density in GDDH13  | 87 |

# List of Tables

|     |   |    |
|-----|---|----|
| 1.1 | Summary of sequenced rosaceae genomes   | 24 |
| 1.2 | List of recent methods to detect differentially methylated loci or regions  | 32 |
| 2.1 | Metrics of the different steps performed for the GDDH13 genome assembly   | 41 |
| 2.2 | Comparison of the GDDH13 genome with previously published assemblies of the apple genome                            | 43 |
| 2.3 | Integrated linkage map SNP markers mapping statistics on GDDH13 V1.0 and GDDH13 V1.1                                | 45 |
| 2.4 | Description of tested parameters in the different steps of the gene structural prediction and corresponding results | 52 |
| 4.1 | Number of DMRs found for each comparison between different apple tissues  | 69 |
| S1  | Summary of the genome assembly features and annotations of the apple GDDH13 genome                                  | 82 |
| S2  | List of CDS elements in which a genetic variant between GDDH13 and GDDH18 was found                                 | 86 |
| S3  | Differentially methylated genes in the comparison between GDDH13 and GDDH18   | 88 |

## List of abbreviations

**Kb, Mb, Gb** : Kilobases, Megabases, Gigabases

**GDDH13** : Golden Delicious Doubled Haploid n°13

**DBG** : De Bruijn Graph

**OLC** : Overlap-Layout-Consensus

**SMRT** : Single-Molecule Real Time

**PE** : Paired-End

**MP** : Mate-pair

**bp** : base pair

**TE** : Transposable Element

**LD** : Linkage Disequilibrium

**MD** : Malus Domestica Predicted Genes (Velasco et al., 2010)

**WGS** : Whole Genome Sequencing

**GO** : Gene Ontology

**WGBS** : Whole Genome Bisulfite Sequencing

**DMR** : Differentially Methylated Region

**DMC** : Differentially Methylated Cytosine

**Chr12** : Chromosome n°12

**DAP** : Day After Pollination

**[-100 : 100]** : a genomic region starting 100 bp before the specified element and ending 100 bp after

## Glossary

**N50** : size of the median-sized sequence in a set of sequences. Commonly used to assess the quality of a genome assembly

**L50** : number of the biggest sequences of a genome assembly necessary to obtain 50% of the assembly length

**contigs** : a set of sequences resulting from a genome assembly

**scaffold** : a sequence resulting from the bonding of two or more contigs by gaps of known length

**30X** : a sequencing depth of 30 x the genome size

# Introduction

Epigenetics refers to transcriptional changes occurring independently of any modification in the DNA sequence. Several epigenetic marks exist and one of the most commonly studied is DNA methylation, which refers to the covalent binding of a methyl group on a cytosine base. DNA methylation has been shown to be involved in gene expression modification, especially when the affected cytosines are located in gene promoter or body. A better understanding of this phenomenon could help to explain and predict some phenotypic changes and, in the context of plants and more particularly fruits, allow to exploit it to improve crops without resorting to GMOs. Apple is an interesting model to study epigenetics because it was subject to a relatively recent and quick domestication and underwent a lot of breeding events which could have produced several different genome wide methylation states.

The domesticated apple (*Malus domestica*) is one of the most cultivated and consumed fruit crop. About 80 millions tons of apple are produced each year in the world. In France, 1.737 million tons of apple were produced in 2013, which places it at the 7th rank of apple cultivating countries worldwide. Genomic knowledge of such a common fruit is therefore very important for geneticists and crop breeders to help create novel varieties optimizing important agronomic traits like fruit taste, size, colour and resistance to the diseases to which apple trees are subjected.

In order to study the apple methylome, an accurate reference genome is needed. A first version of the apple genome was published in 2010 [Velasco et al., 2010]. However, due to the limited sequencing technologies available at the time, the genome assembly was highly fragmented which had a negative impact on the quality of the gene annotation, as well on all the large scale genetic studies that depended on this reference. Therefore, we decided to produce a new apple reference genome using the latest sequencing technologies (PacBio, BioNano). The second chapter of this manuscript describes the genome assembly and the gene annotation processes. We generated an assembly of 643.2 MegaBases (Mb) with a N50 of 5,558 Kb. Most of it was oriented and anchored into 17 pseudo-molecules which represent the 17 chromosomes of apple. Using RNA-sequencing, public protein databases and *ab initio* prediction, we annotated 45,116 protein coding

genes. This structural annotation obtained a BUSCO [Simão et al., 2015], which quantifies the completeness of a gene annotation, of 96.8%.

Using whole genome bisulfite sequencing and this assembly as a reference sequence, we generated genome wide methylomes for two isogenic lines of apple, called GDDH13 and GDDH18, which are two haploid apple trees that produce apples of different sizes. By comparing their methylation states, we searched for Differentially Methylated Regions (DMRs) that could explain the fruit size difference by affecting gene expression. We found several DMRs associated to a list of candidate genes that could potentially be involved in determining fruit size. Moreover, we found general correlations between methylation in the gene putative promoters and body and gene expression. These results are described in the third chapter of this manuscript.

Given the several issues encountered during DMRs computing in the aforementioned part of this work, we decided to develop a complete and easy-to-use pipeline to compute DMRs using a low number of biological replicates. We produced a tool which can compute DMRs in a few hours, depending on the number of replicates and the size of the reference genome, and output comprehensive metrics in order to efficiently filter and interpret found DMRs. This work is developed in the fourth chapter of this manuscript.

Finally, the fifth chapter describes the few side projects that were conducted in parallel of the main work of this PhD thesis. We performed the genome assembly of *T. lutea* and obtained 193 contigs having a N50 of 853 Kb for a total assembly size of 82 Mb. We also participated in the last steps of the genome assembly of the rose (*R. chinensis*) genome, performing the genome polishing and anchoring on a genetic map. We anchored 90% of the assembly on 7 pseudo-molecules representing the 7 chromosomes of rose.



# Chapter 1

## State of the art

### 1.1 Genome sequencing

Sequencing a genome is crucial to conduct biological and genetic research on an organism. This part will first describe the evolution of genome sequencing and explain the improvements made with each new technology. Second, I will focus on the state of sequenced genomes of organisms more closely related to apple, which is the main organism of interest in this manuscript.

#### 1.1.1 First generations of sequencing technologies

##### Sanger sequencing

The Sanger sequencing, developed in 1977 [Sanger et al., 1977] was the first generation sequencing technologies. It uses a DNA primer, a DNA polymerase, a reference sequence, deoxynucleotides and di-deoxynucleotides. Di-deoxynucleotides stops the DNA strand elongation when used by the DNA-polymerase. By using the four types of di-deoxynucleotides separately, the method allows to visualize the last nucleotide of all possible aborted sequences on gel, thus reconstructing the original sequence. This produces reads shorter than 1 Kilobase (Kb) [Heather and Chain, 2016]. The first genome assembly techniques subsequently appeared [Staden, 1979] to assemble sequences, resulting from shotgun sequencing of overlapping DNA fragments, into contigs. This technology is still occasionally used in small sequencing projects for its accuracy.

##### Pyrosequencing and Solexa

This technique was developed with time thanks to new techniques such as PCR [Saiki et al., 1988] and conducted to the appearance of the second generation sequencing. One of the main technologies from this

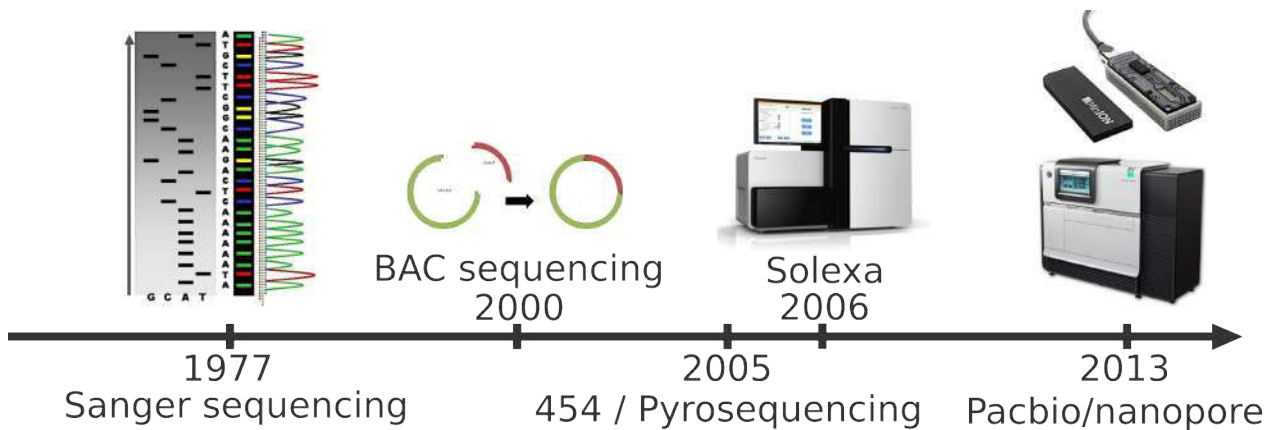


Figure 1.1 Evolution of the sequencing technologies. Left image created by Abizar Lakdawalla.

generation is the pyrosequencing [Ronaghi, 2001] which consists in synthesizing the complementary strand of a DNA fragment and detecting the activity of the DNA polymerase using a chemoluminescent enzyme. This was used to generate short reads of less than 300 nucleotides [Mashayekhi and Ronaghi, 2007]. Another technology of this generation is the Illumina/Solexa sequencing [Bennett, 2004] in which oligonucleotides are bound to flowcells before the PCR and base calling are performed. This produces short paired-end reads since both ends of each DNA fragment are sequenced, ranging from 150 nucleotides to 300 nucleotides in length for the last generation of machines. These reads have less than 1% error rate [Luo et al., 2012a]. This technology is still used in 2018 because of its high accuracy and low cost and was used during various projects this manuscript will describe.

### 1.1.2 Last generation : long reads

The last generation of sequencing technologies followed with the PacBio [Eid et al., 2009] and Nanopore [Clarke et al., 2015] technologies.

PacBio uses the principle of Single Molecule Real Time (SMRT) sequencing (Fig. 1.2). A single molecule of DNA is fixed at the bottom of a well. Each type of nucleotide is attached to a fluorescent dye. When the DNA polymerase uses a nucleotide to synthesize the complementary sequence to the fixed DNA molecule, the dye and the used nucleotide are separated. At this moment the dye emits fluorescence which is detected and interpreted into a base during the subsequent base-calling process. In contrast, Nanopore sequencing consists in driving DNA molecules through small biological channels. Each nucleic acid is determined by measuring the specific current change it provokes while passing through the nanopore.

The reads produced by the PacBio and Nanopore technologies have two particularities compared to older generations' reads. First, they have a high error rate of around 15%. Specific assembly softwares were developed in order to handle this inconvenience and will be discussed later in this manuscript. Second,

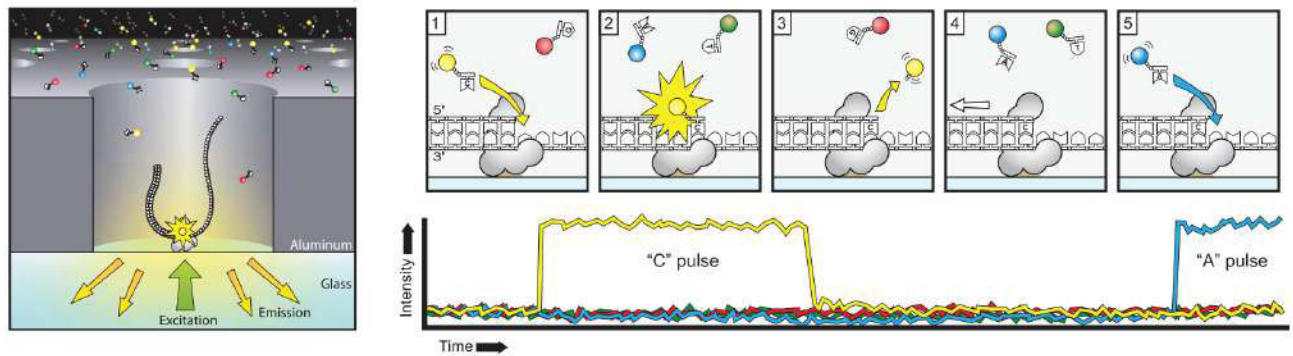


Figure 1.2 Principle of PacBio sequencing. (a) The DNA molecule is fixed at the bottom of the well. (b) Mechanism of nucleotide incorporation and fluorescence emission. The dye corresponding to the incorporated nucleotide emits fluorescence. Figure reproduced from [Eid et al., 2009](#).

these reads are longer, having a mean size of around 10 Kb depending on the library. This provides great advantages during genome assembly which made long reads mandatory to obtain a high-quality genome.

## 1.2 Genome assembly

DNA sequencing technologies produce very short random reads in comparison to the sizes of genomes. To overcome this limit, a number of nucleotides exceeding the genome size multiple times is sequenced. The aim of genome assembly softwares is to reconstruct the original genome from these short sequences. Once reads are produced, they have to be assembled into larger fragments called "contigs" in order to provide large enough sequences to perform gene annotation and genetic studies. Multiple assembly softwares were developed for this purpose and use different kinds of algorithms adapted to each type of reads. However there are multiple bottlenecks during the genome assembly step and, for now, it is challenging to obtain chromosome-scale contigs. Thus, the aim when assembling a genome is to obtain contigs as large as possible, which long reads allow to do.

### 1.2.1 Types of assembly algorithms

The aim of assembly algorithms is to merge overlapping reads in order to build sequences as long as possible. This implies an alignment or pseudo-alignment step which has to allow some differences between the sequences to account for sequencing errors but stringent enough to not produce chimeric contigs. Genome assembly algorithms are divided in two categories : De Bruijn Graph (DBG) algorithms which rely on a K-mer graph to find shared K-mers between reads, and Overlap-Layout-Consensus (OLC) algorithms which construct an overlap graph after multiple alignments of the reads.



## Overlap-Layout-Consensus algorithms

This kind of algorithm is separated in three phases (**Fig. 1.4** [Miller et al., 2010]). The first step, Overlap, consists in building an overlap graph from pair-wise reads comparison. During this step, each read is aligned against each other with the chosen alignment algorithm. When the alignment is performed, an overlap graph is constructed, in which nodes correspond to the reads and edges to the overlap found during the pairwise alignment. A stringent alignment leads to a less dense overlap graph, thus shorter contigs. On the contrary, a less stringent alignment leads to a dense graph, longer contigs at the cost of a specificity loss which is concretized by the risk of constructing chimeric contigs.

The second step, Layout, consists in rearranging the reads in the most consistent order in the overlap graph. Finally, during the consensus step, a multiple alignment between the remaining reads is performed and the most commonly found nucleotide at each position is chosen to construct the final consensus sequence.

Multiple Overlap-Layout-Consensus (OLC) assemblers are being used but the original one was the Celera assembler [Myers et al., 2000]. They differ by their alignment algorithm during the overlap step, their treatment and simplification of the overlap graph during the layout step, and sometimes during the read correction step for the most recent OLC assembler adapted for long reads.

During the overlap detection, an OLC algorithm can allow some mismatches between two reads and will still produce one node per read and the corresponding edges, while a DBG algorithm will create different nodes corresponding to the different K-mer found [Li et al., 2012]. Thus, OLC algorithms are, by nature, more adapted to erroneous long reads than DBG algorithms and are used in most long reads assemblers, like Canu [Koren et al., 2017], Falcon [Chin et al., 2016], or Cerulean [Deshpande et al., 2013].

### 1.2.2 Challenges in assembling repetitive elements

The main challenge when doing a *de novo* assembly of a genome is to correctly reconstruct its repetitive content [Treangen and Salzberg, 2012]. Repetitive sequences, which often correspond to transposable elements, represent a large portion of big genomes : 50% of the human genome [Schmid and Deininger, 1975], 35% of the rice genome [Takata et al., 2007] and 57% of the apple genome [Daccord et al., 2017]. Moreover, they play an essential role in epigenetic gene regulation [Waterland and Jirtle, 2003]. Typically, if one fails to correctly assemble repetitive elements, the assembly will consist in isolated gene islands surrounded by small collapsed (or falsely assembled together) fragments of transposable elements. Thus, a reference genome must have an exhaustive assembling of the transposable elements.

Technically, repetitive elements create ambiguities during the similarity searching between reads. In the case of DBG algorithms, finding two identical K-mer will result in one node in the graph, even if these two K-mers correspond to two different sequences in the genome. In the case of OLC algorithms, this will create

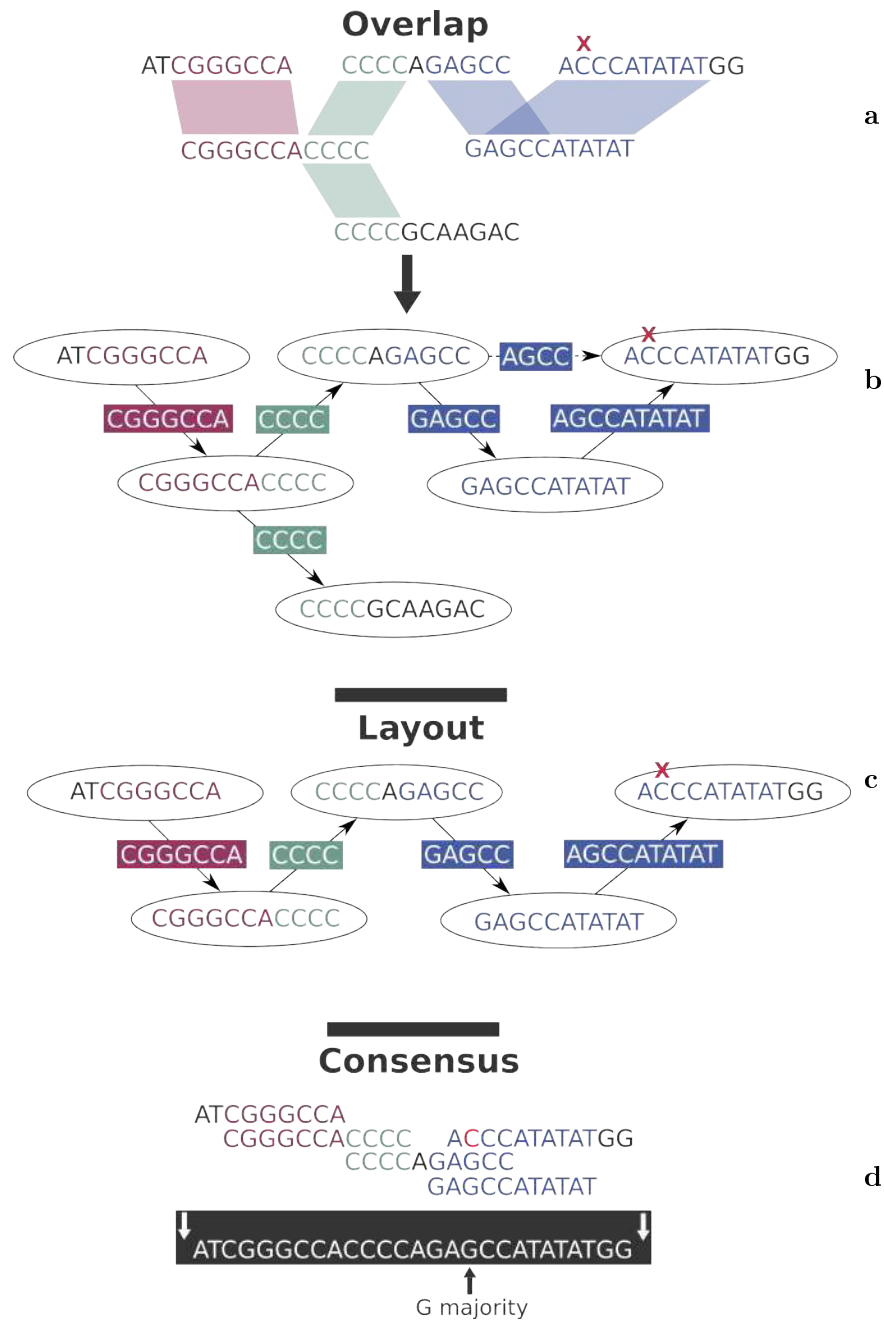


Figure 1.4 Illustration of an Overlap-Layout-Consensus assembly algorithm. (a) A multiple read alignment is performed during the overlap step. (b) The resulting overlap graph is constructed. (c) The overlap graph is filtered during the layout step. (d) The consensus sequence is reconstructed using the overlapping sequences parsed in the overlap graph.

ambiguity during the overlap step, in which the overlap graph will create ambiguous edges between the same nodes. In order to resolve a repetitive element, the read has to be longer than the repeat (**Fig. 1.5**). If no reads are longer than the repeat, it is impossible to make a distinction between reads resulting from the sequencing of multiple distinct repeats on the genome, thus collapsing all the repeats in one contig during the genome assembly and creating a break on each side of each unresolved repeats which leads to fragmented

assemblies. Long reads have the possibility of spanning longer repeats from edge to edge and allow to perform a better reconstruction of the repetitive content of a genome. For this reason, they are used in most of the modern genome sequencing (sometimes in combination with short reads), especially when the aim is to make a reference genome, despite their high error-rate which has to be handled by specialized assembly softwares.

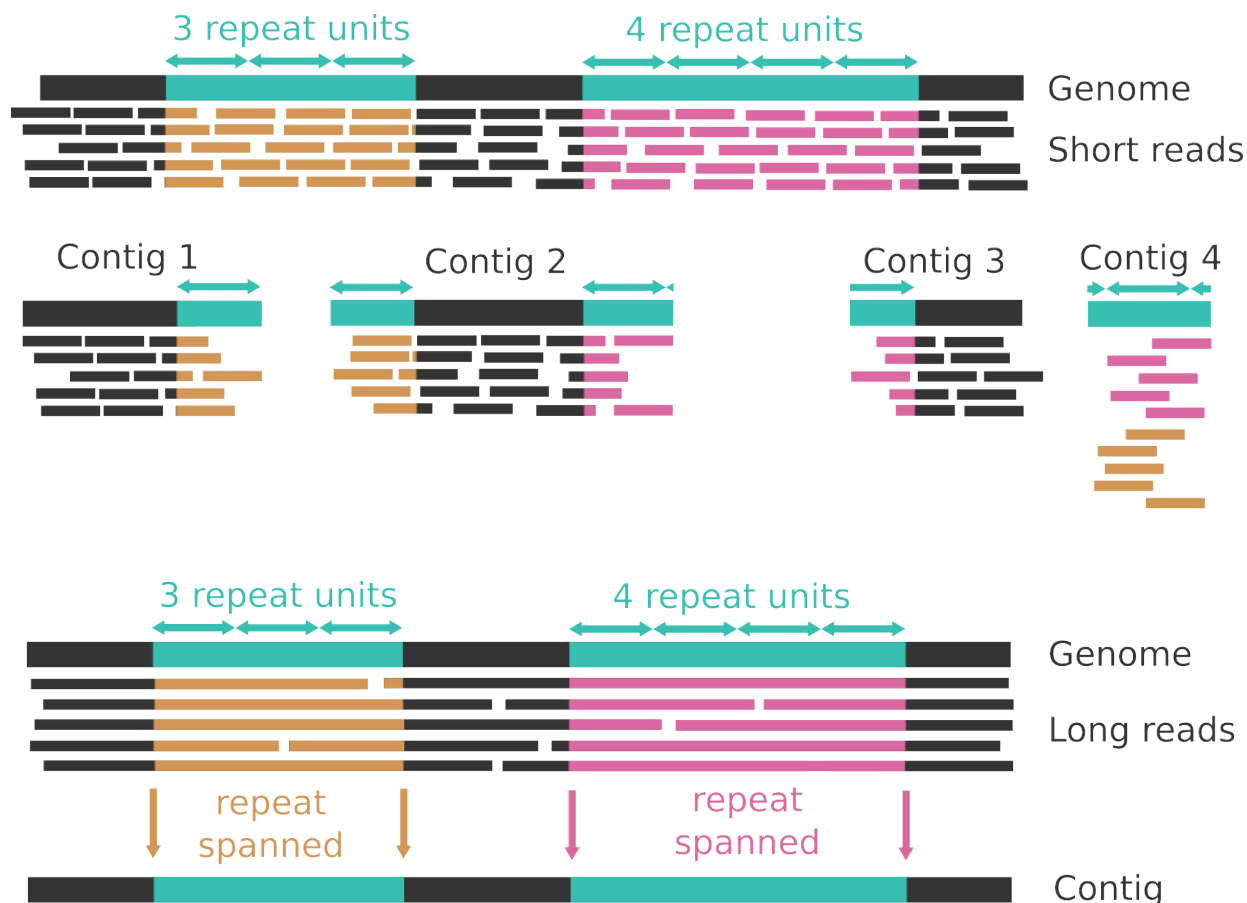


Figure 1.5 Illustration of repeat resolving using short and long reads. (a) Using short reads data. All the represented repeat units are identical. Repeats are drawn in light blue on the genomic sequence, and in orange and red on the reads to help differentiate the locus the reads come from. (b) same as (a) but using long reads.

### 1.2.3 Handling the high error rates of long reads

The major inconvenience of long reads technologies is their low nucleotide accuracy. PacBio and Nanopore reads have about 15% error-rate [Koren et al., 2012] [Judge et al., 2015], which complicates alignments between reads since the pairwise alignments between two reads will have twice the difference than these two reads individual error rates. During the assembly steps, if the alignments sensitivity is raised too high to counterbalance these innate differences, this can result in chimeric contigs assembly, especially because of the confusion of sequencing errors and differences between highly similar sequences. On the opposite, if

nothing is done to raise the sensitivity, sequencing errors will produce too many differences for the alignment software to demonstrate a similarity between two noisy reads of the same original sequences, and the resulting assembly will be highly fragmented.

Two main classes of approaches were developed to handle the high error rates of long reads : correction by self-alignment and hybrid error correction using a complementary high-accuracy sequencing technology.

### **Correction by self-alignment**

The read correction by self-alignment consists in performing a multiple alignment between reads prior to the assembly and generate consensus reads which will be used downstream. In this case, the sequencing depth helps to distinguish sequencing errors and real sequences differences. However, because the pairwise error rate is twice the individual reads error rates, a lot of too noisy reads are discarded during the correction. This results in a coverage of corrected reads inferior to the original coverage of sequenced reads [Lee et al., 2014]. In case of low-coverage sequencing, this can be a problem because a low amount of reads will be used in the subsequent assembly which can lead to a fragmented reconstruction of the genome. Moreover, the error correction is inefficient on regions with low coverage (Fig. 1.6). Specially designed aligners, like the BLASR software [Chaisson and Tesler, 2012], must be used in order to detect noisy overlaps between reads.

The self-correction method was implemented in multiple softwares, among them the pre-assembly read corrector LoRDEC [Salmela and Rivals, 2014], or the first step of the assemblers Canu [Koren et al., 2017], PBCr self-correction [Koren et al., 2013] and HGAP [Chin et al., 2013]. It is suited for PacBio reads correction since they have uniformly distributed errors [Koren et al., 2012], provided the sequencing depth is sufficient. However, this is not the case concerning Nanopore reads for which the sequencing errors are not uniformly distributed [Judge et al., 2015] : these systematic errors will be incorporated into the consensus reads and remain in the assembly. In this case and in the case of low-depth PacBio sequencing, an alternate correction method is to use another complementary technology.

### **Hybrid error correction**

Hybrid error correction consists in the same process than the self-error correction previously described but using shorter high-quality reads, like Illumina reads, to build a corrected consensus for each long noisy read. This method was implemented several softwares, among them proovread [Hackl et al., 2014] and PBCr [Koren et al., 2012] which correct the reads prior to the assembly, and Pilon [Walker et al., 2014] which corrects the assembly itself, as a final step.

While dealing with Nanopore data, using Pilon at the end of the assembly step allows to correct systematic errors produced by the technology in order to further improve the assemblies [Istace et al., 2017] [Schmidt et al., 2017].



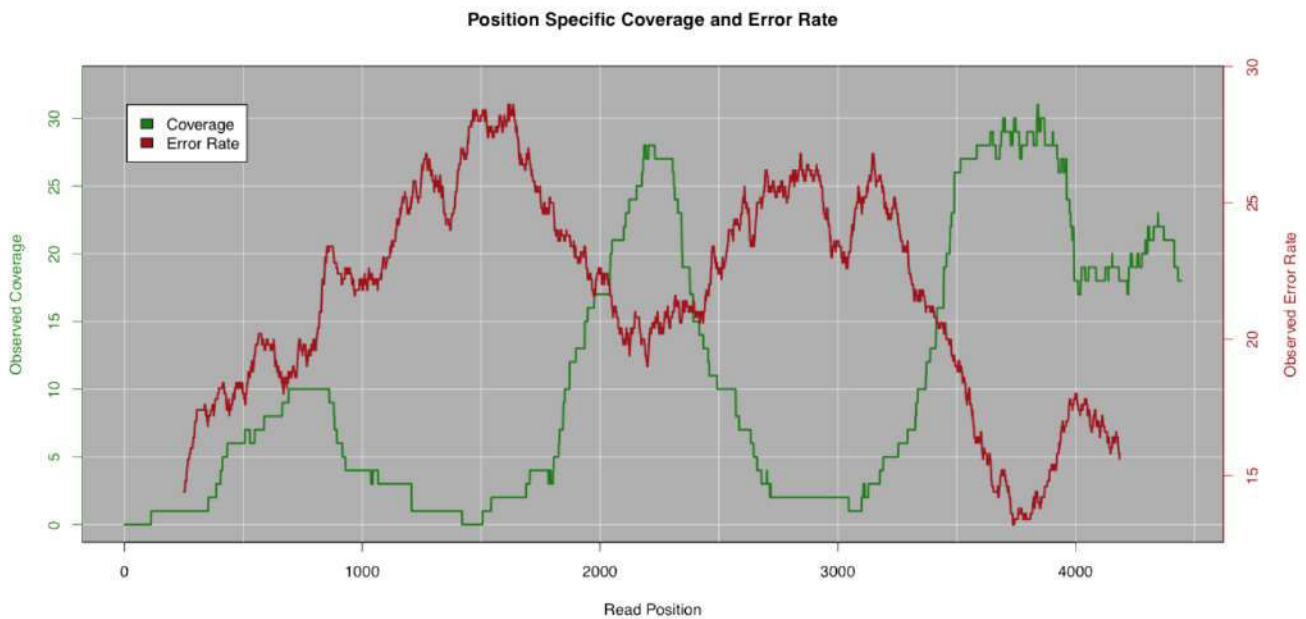


Figure 1.6 Error rate and PacBioToCA [Lee et al., 2014] coverage of an individual read. The plot shows the characteristics of an individual PacBio read in the pipeline. The red curve shows the local error rate relative to the reference genome computed by a 200bp sliding window and shows that the error rate can fluctuate from 15% to nearly 30%. The green curve shows the number of short reads that could be aligned by the PacBioToCA pipeline at each position in the read. The error rate and coverage levels are anti-correlated, which resulted in the read being split into multiple segments after correction. Figure reproduced from [Lee et al., 2014]

### 1.2.4 Assembly scaffolding

Once the assembly accomplished, a set of discontinuous sequences, the contigs, are obtained. Thanks to different technologies or a genetic map, it is possible to place and orient these contigs relatively to one another. This process is called scaffolding and allows to obtain longer sequences called scaffolds, which in the optimal case will be chromosome-scaled. Scaffolding usually does not add supplementary sequence information but only a determined number of unknown nucleotides (N) in between the anchored contigs (Fig. 1.7).



Figure 1.7 Illustration of contigs scaffolding. Three contigs are assembled into one scaffold. Known nucleotides are represented in color and unknown nucleotides (N) are represented in black.

Several different scaffolding technologies exist but the two most recent and commonly used to produce long

scaffolds are BioNano and Hi-C. BioNano [Persson and Tegenfeldt, 2010] optical mapping consists in labeling a seven nucleotides marker on the genome and producing long optical maps in which these markers will appear fluorescent (Fig. 1.8). Using the relative spacing between the markers, contigs are anchored on the BioNano maps and an adequate number of unknown nucleotides is placed in between the contigs. This technique usually produces scaffolds that span several megabases [Martin et al., 2016] [Hatakeyama et al., 2017].

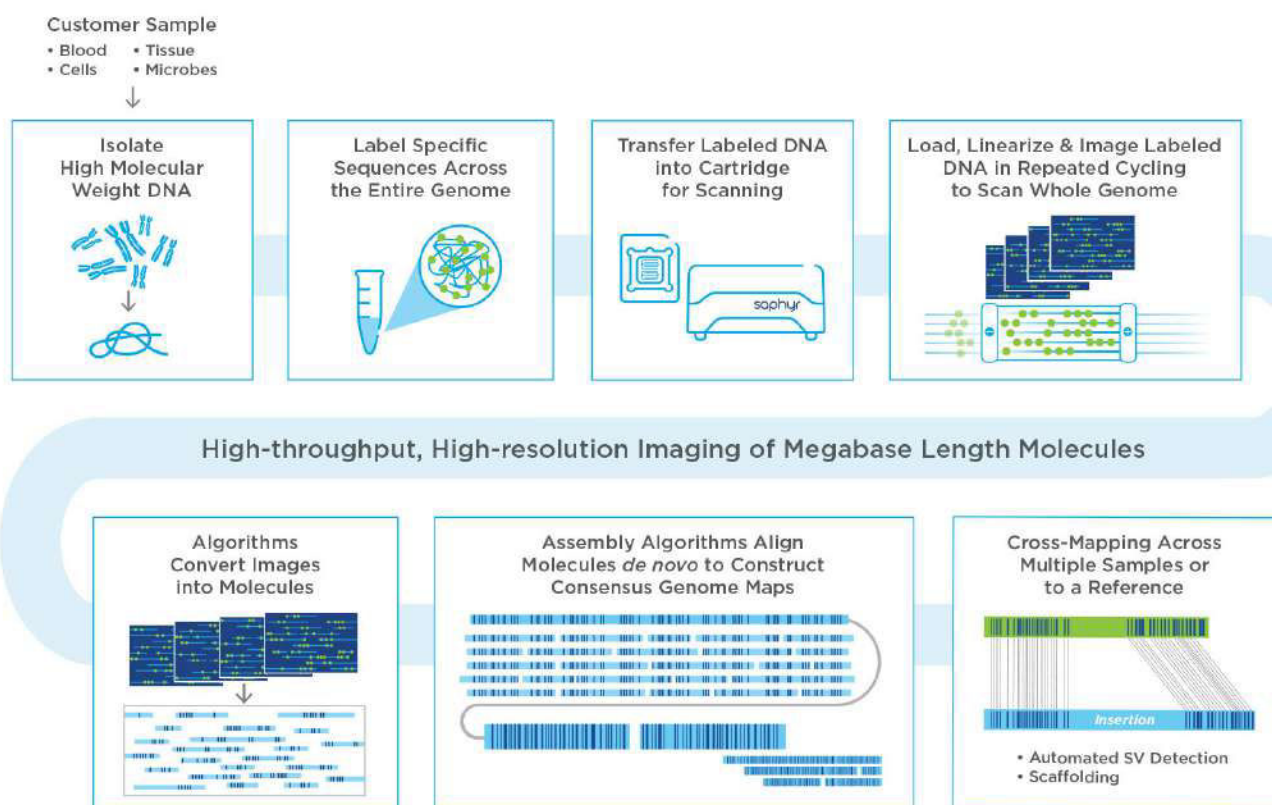


Figure 1.8 BioNano workflow illustration. Source : <https://bionanogenomics.com/technology/platform-technology/>

Hi-C [Kaplan and Dekker, 2013] is a technique used to show the spatial arrangement of the DNA inside the nucleus in the cell. By knowing all the relative positions of DNA sequences, it is possible to infer a relative position to every contig relatively to each other. This technique allows to obtain chromosome-scaled scaffolds [Dudchenko et al., 2017] [Putnam et al., 2016].

## 1.3 State of sequenced genomes

Modern sequencing technologies provide higher throughputs thus less expensive genomes. This leads to an increasingly high number of available genomes in public databases (Fig. 1.9). However, most of these genomes are draft assemblies, which are too fragmented to perform accurate advanced analysis.

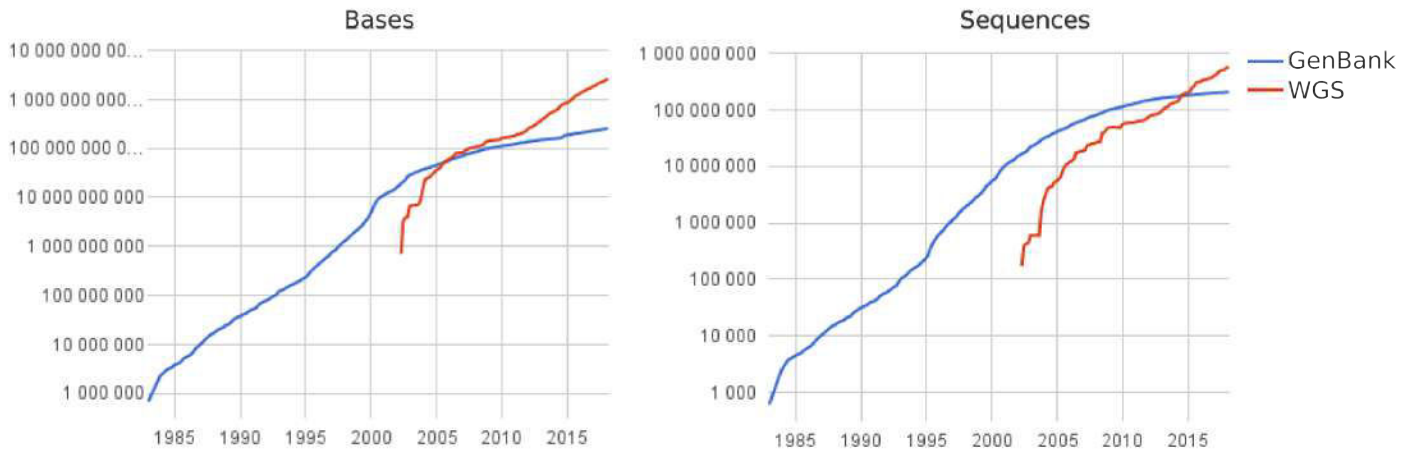


Figure 1.9 Evolution of the volume of deposited genomic data. Number of nucleotides (left) and sequences (right) stored in GenBank from 1982 to 2017. WGS and non WGS sequencing are represented in red and blue respectively. Data taken from [ncbi.nlm.nih.gov/genbank/statistics/](http://ncbi.nlm.nih.gov/genbank/statistics/).

Moreover, a large part of existing WGS concerns small-sized genomes like bacteria or fungi. Unlike plant and animal genomes, bacteria genomes are much smaller, hence have fewer repetitive sequences which make them easier to assemble.

### 1.3.1 Plant genomes

The first assembled plant genome was *A. thaliana* [Initiative et al., 2000] which has a relatively small size of 135 Mb and a low transposable element content of approximately 10%. Following this model plant, most of the produced high-quality plant genomes were for crops of agricultural importance, like rice [Yu et al., 2002], maize [Schnable et al., 2009], grapevine [Jaillon et al., 2007] and soybean [Schmutz et al., 2010]. Most of these were sequenced using BAC sequencing/capillary sequencing [Schatz et al., 2012]. Following these, new plant genomes were sequenced using second-generation (Illumina) sequencing technologies, like the first version of the apple genome [Velasco et al., 2010], tomato [Consortium et al., 2012b], or watermelon [Guo et al., 2013]. However most of the genomes of this generation remain highly fragmented because of several computational challenges caused by plant genomes specificities like genome duplications and the technical limit of read length.

First, they have a higher ploidy rate than other genomes [Meyers and Levin, 2006]. For example, some ex-

tremes like the bread wheat genome (*Triticum aestivum*) [Consortium et al., 2014] are hexaploid or higher, resulting of the merging of several ancient "subgenomes". Second, they have a higher heterozygosity rate [Gore et al., 2009], especially trees [Jaramillo-Correa et al., 2010]. These two particularities add more genotypic variations which complexify the overlap finding during the assembly if the aim is to assemble all the subgenomes together. Third, they tend to have a high transposable element content [Feschotte et al., 2002] which also complicates the assembly, especially if one doesn't have access to long reads (see 1.2.2). Long reads technologies unlocked a lot of genome sequencing opportunities and are now mandatory to obtain a good plant genome assembly. Several high-quality plant genomes were recently sequenced using Pacbio reads, including the rubber tree [Pootakham et al., 2017], as well as the quinoa [Zou et al., 2017] and the maize genome which was recently improved [Jiao et al., 2017]. The wild tomato genome was also recently published [Schmidt et al., 2017] using Nanopore reads. The contig N50 of these assemblies regularly surpass 1 Mb which is greatly superior to what could have been obtained using short reads and testifies to a good assembly of repetitive content.

### 1.3.2 Rosaceae genomes

Rosaceae is a plant family comprising about 3000 species [Xiang et al., 2016], including several commonly consumed fruits, such as apple or apricot, which went through domestication efforts. Rosaceae fruits can be very different : some are fleshy, like apple and strawberry, while others like almonds or chestnuts are dry (Fig. 1.10). Having a high diversity among a family is useful to study fruit evolution. This, added to the economical importance of these fruits, makes the need for high-quality rosaceae references genomes very high. For example, accessibility to a well-annotated fruit reference genome allows to identify which genes play a role in fruit taste, appearance or disease resistance and this knowledge can be exploited by breeders in order to produce new varieties.

Rosaceae genomes are usually diploid and small to medium-sized compared to other plant genomes, ranging from around 200 Mb for strawberry [Shulaev et al., 2011] to more than 600 Mb for apple [Velasco et al., 2010]. Some of them (apple and pear) recently underwent a whole genome duplication event which results in a higher number of chromosomes than other rosaceae [Velasco et al., 2010] [Wu et al., 2013]. A few rosaceae genomes of economical interest have been sequenced (Table 1.1) but none of them were made with the latest generation of sequencing technologies. Thus, most of these genomes are highly fragmented and have a small N50 rarely surpassing 1 Mb. However, some of these genomes are currently being resequenced, or were very recently resequenced using long reads, like apple (this work), pear, strawberry [Edger et al., 2017] and rosa (this work), resulting in chromosome-scaled assemblies.

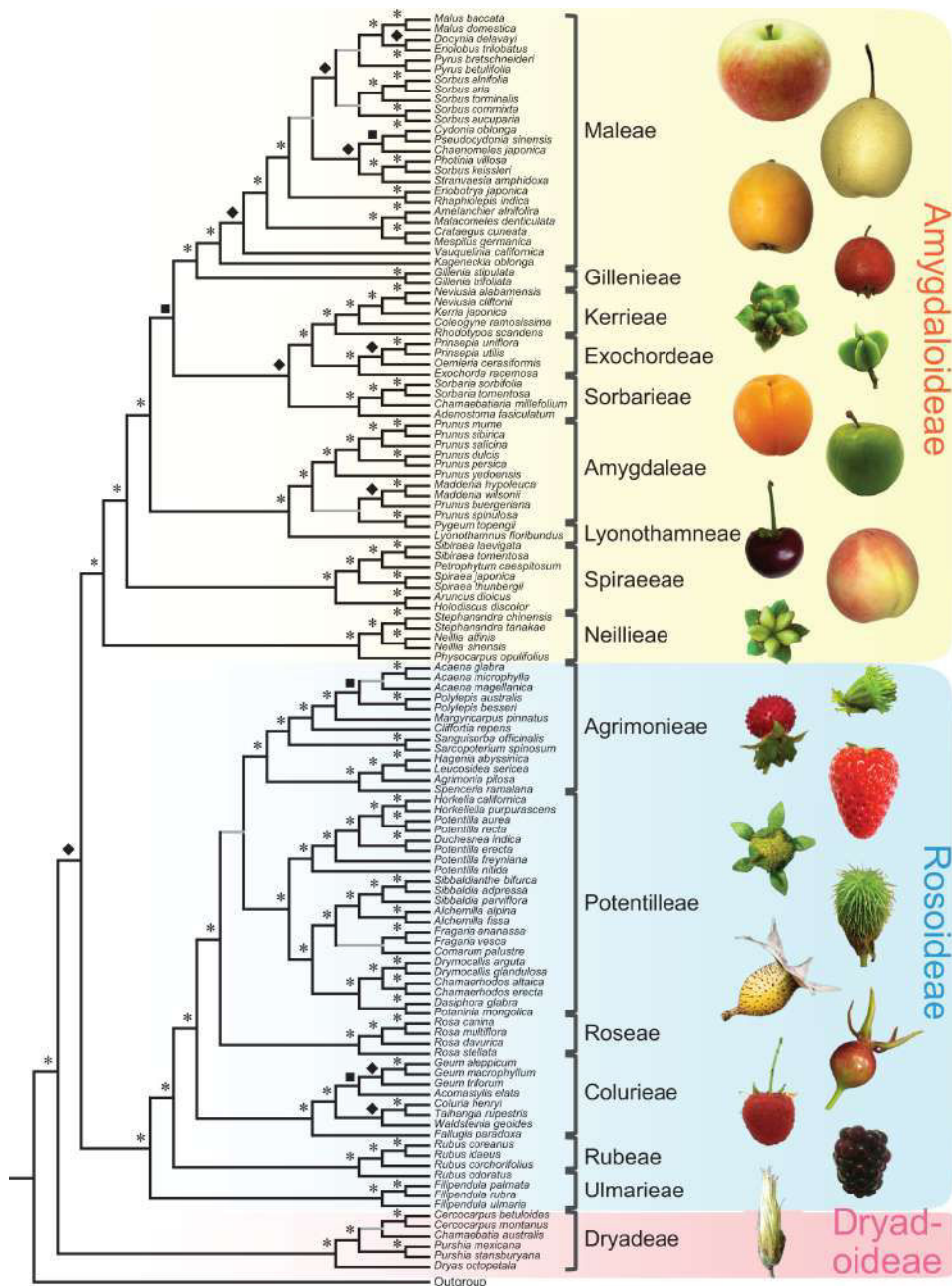


Figure 1.10 Summary of Rosaceae phylogeny and Rosaceae fruit morphologies. On the left is a summary tree with results from five coalescence analyses of 882, 571, 444, 256, and 113 gene sets, respectively, and a concatenation analysis using the 113-gene supermatrix. Topologies consistent in all six trees are drawn in black lines. Grey lines show uncertain relationships, with some trees support the topology. Figure taken from [Xiang et al., 2016](#)

### 1.3.3 The apple (*Malus Domestica*) genome

The domesticated apple (*Malus domestica*) is a economically important crop cultivated in almost all the northern hemisphere, but also in some countries of the southern hemisphere like New Zealand or Brazil. It results from a domestication of *Malus sieversii*, which originates from the Tian Shian mountains in Asia

| Species                              | N50 (Kb) | Size (Mb) | Chromosomes | Gene number | Reference            |
|--------------------------------------|----------|-----------|-------------|-------------|----------------------|
| Apple ( <i>Malus domestica</i> )     | 16       | 603       | 17          | 57386       | Velasco et al., 2010 |
| Pear ( <i>Pyrus bretschneideri</i> ) | 540      | 512       | 17          | 42812       | Wu et al., 2013      |
| Peach ( <i>Prunus persica</i> )      | 4000     | 224       | 8           | 27852       | Verde et al., 2013   |
| Apricot ( <i>Prunus mume</i> )       | 577      | 237       | 8           | 31390       | Zhang et al., 2012   |
| Strawberry ( <i>Fragaria vesca</i> ) | 1300     | 210       | 7           | 25050       | Shulaev et al., 2011 |

Table 1.1 Summary of sequenced rosaceae genomes.

(Fig. 1.11). It was hybridized with *Malus silvestris* and *Malus orientalis* along the silk road to result in the modern commercialized apple [Cornille et al., 2014].

Apple is one the most consumed fruit : about 1.5 million tons of apple are consumed in France each year. Apple is a self-incompatible plant, and, like most other crops, is subject to several diseases, like apple scab, fire blight or crown gall. To prevent these, usually, an important quantity of pesticides is used by farmer on apple corps. In this context, the knowledge of the apple genome and the genetic studies are very valuable to produce apples more efficiently and in a safer way for public health.

The apple genome was already sequenced in 2010 [Velasco et al., 2010]. It has an estimated size of 742 Mb, is diploid and possesses 17 chromosomes. The authors found 57,386 genes and a transposable element content of 42%. The genome has been shown to be at least partially duplicated before the work presented in this manuscript, and genetically very close to pear. However, only short Illumina reads were used to sequence the genome. This can be problematic especially in the case of apple for which the genome is potentially completely duplicated and heterozygous, resulting in very similar, but different sequences being present four times in the genome. The consequence is that the genome assembly is very fragmented (122,146 contigs ; contig N50 = 16,171b). It has been shown that a good genome assembly will have a huge impact on the gene annotation [Florea et al., 2011] and thus on all the downstream biological analysis. In this context, we decided to produce a new apple genome using long reads to provide a high-quality work base to the apple scientific community.

### 1.3.4 Thesis objectives

The first objective of this thesis is to produce a new, high-quality reference apple genome which will provide an important basis for not only the epigenetics studies described later in this manuscript but also numerous other apple related studies that rely on genetics. A few steps are needed to make this reference genome :

(1) Produce a highly contiguous genome assembly. We generated short (Illumina) and long (PacBio) reads, and BioNano optical maps. Using these three technologies altogether, jointly with an integrated linkage



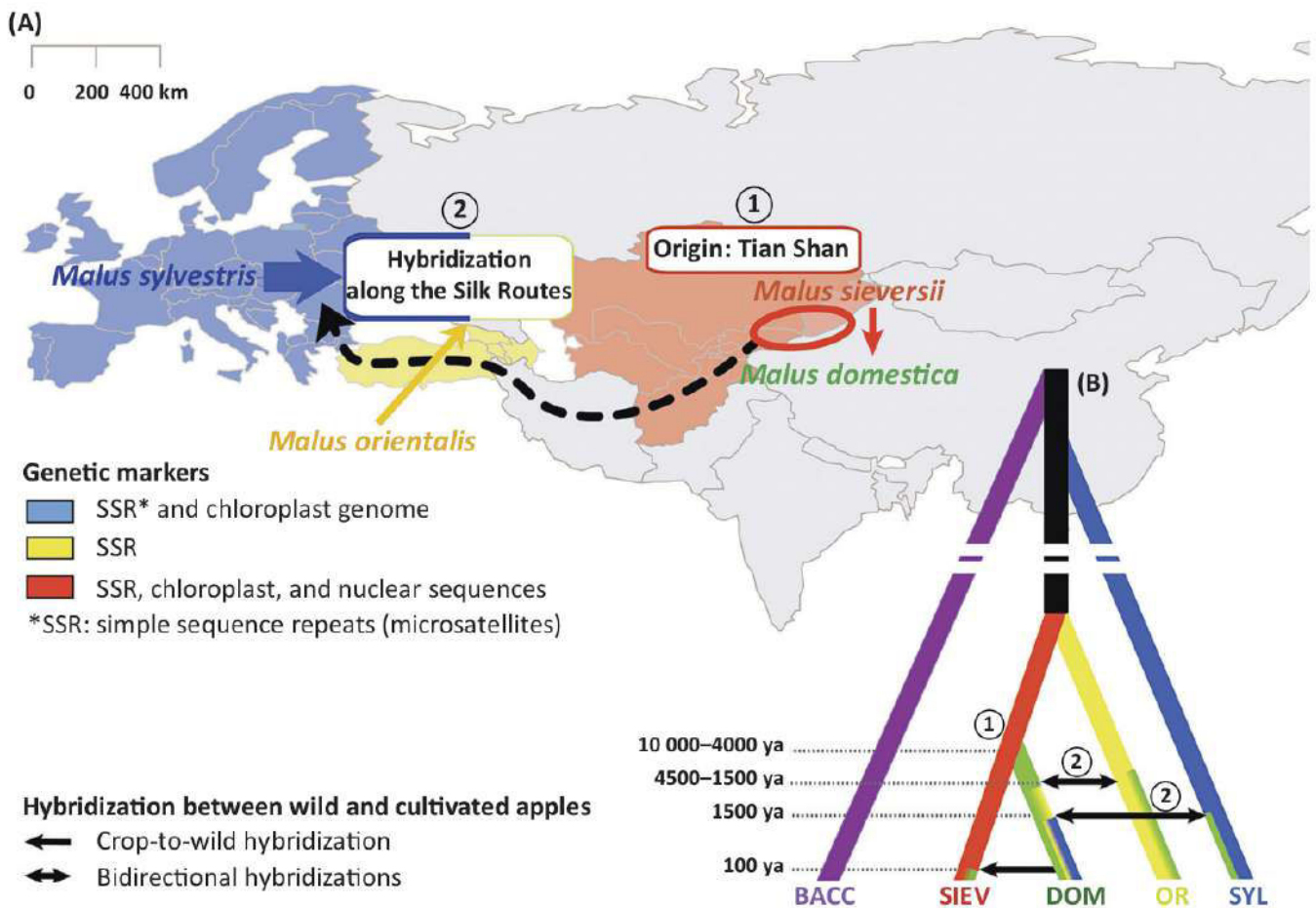


Figure 1.11 Evolutionary history of the cultivated apple. (A) This history was revealed by recent population studies using different types of molecular markers for evolutionary inferences. (1) Origin in the Tian Shan Mountains from *Malus sieversii*, followed by (2) dispersal from Asia to Europe along the Silk Route, facilitating hybridization and introgression from the Caucasian and European crabapples. Arrow thickness is proportional to the genetic contribution of various wild species to the genetic makeup of *Malus domestica*. (B) Genealogical relationships between wild and cultivated apples. Approximate dates of the domestication and hybridization events between wild and cultivated species are detailed in the legend. Abbreviations: BACC, *Malus baccata*; DOM, *M. domestica*; OR, *Malus orientalis*; SIEV, *M. sieversii*; SYL, *Malus sylvestris*; ya, years ago. Figure taken from [Cornille et al., 2014](#)

map, we produced a new apple genome sequence.

(2) Perform a gene annotation. First, we used transcript sequences coming from several libraries of RNA-seq of various organs, blast similarities against public protein sequences databases, *ab initio* prediction and various standalone tools to predict the exon structure of protein-coding and non-coding genes on the genome. Second, we assigned a function to each predicted gene.

(3) Integrate the above-mentioned results, and more metadata in a public genome browser, to allow the scientific community to use the genome to conduct their studies.

## 1.4 Epigenetics

Epigenetics generally describes all the heritable gene expression and phenotypic variations occurring without any changes in the nucleotide sequence of a genome [Allis and Jenuwein, 2016]. Several epigenetics marks are being studied : among them modifications of DNA bases like DNA methylation and post-translational modifications of histones [Dupont et al., 2009]. In this manuscript we will focus on DNA methylation changes, in particular on apple.

### 1.4.1 DNA Methylation

DNA methylation is one of the main epigenetic modifications. It consists in the binding of a methyl group on either a cytosine, an adenine or a guanine [Dupont et al., 2009]. However, cytosine methylation is the most frequent, in particular the 5C methylation (Fig. 1.12a) which consists in the binding of a methyl group on the fifth atom of a cytosine. Three different methylation contexts are being distinguished : CG, CHG and CHH (in which H = A, T or C) depending on the one or two nucleotides following a given cytosine. DNA methylation levels, patterns and contexts are variables depending on the kingdom, organism, family, tissues and cells. In animals, the predominant methylation context is CG [Bird, 2002] on an estimated 70% of the CG cytosines on the human genome [Ehrlich et al., 1982]. The unmethylated CG cytosines are often present upstream of the genes transcription starting sites (TSS) in CG-rich regions called CG islands. In plants, the three methylations contexts CG, CHG and CHH are frequently methylated [Vanyushin, 2006]. The global methylation levels between plants species is very variable but CHH cytosines are always less methylated than other sequences on average. In plants, DNA methylation primarily occurs in transposable elements and cytosines inside genes and exons in particular tend to be less methylated on average [Zhang et al., 2006] [Daccord et al., 2017].

Several studies have reported a reverse correlation between DNA methylation and gene expression [Bird, 1984] [Cedar, 1988] [Baylin, 2005] and more precisely between promoter methylation and gene expression [Di Croce et al., 2007] (example : Fig. 1.12b). In *A. thaliana*, a link between methylation and transcription has been found in *A. thaliana* [Zilberman et al., 2007] in which a loss of DNA methylation in the gene body results in increased transcription.

### 1.4.2 Methylation studies with NGS : bisulfite sequencing

To study DNA methylation at nucleotide resolution, the most commonly used technique is Whole Genome Bisulfite Sequencing (WGBS). This consists in a classic Illumina sequencing but a bisulfite treatment is applied on the DNA which converts all unmethylated cytosines into uracils and thymines subsequently. Ef-



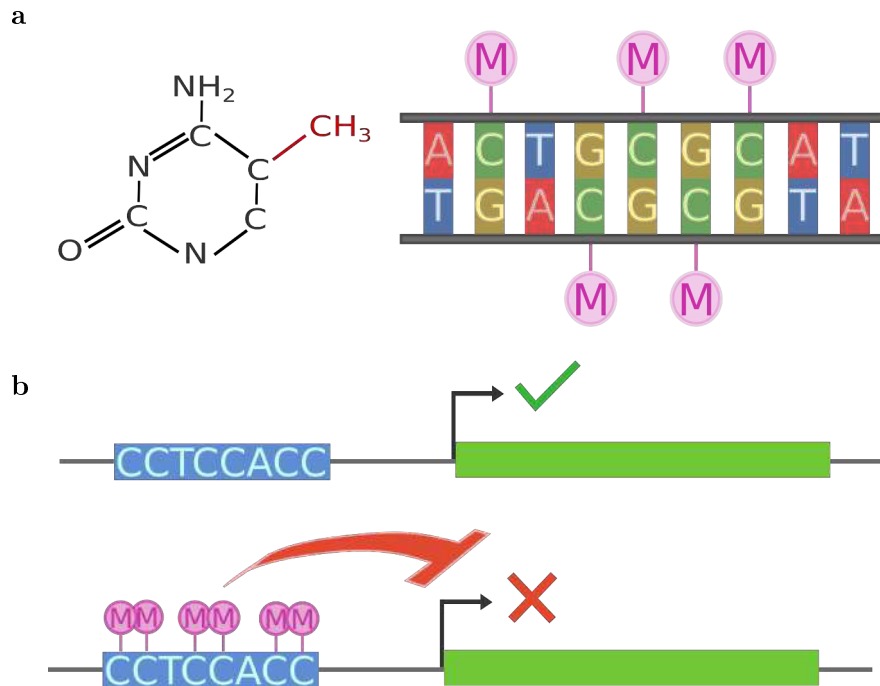


Figure 1.12 Base principle of DNA methylation and hypothesis about gene regulation by promoter methylation. (a) (Left) Representation of a methyl group (in red) bound to a nucleotide (in black). (Right) Representation of cytosine methylation on a simple DNA double strand. Pink circled "M" correspond to methyl groups. (b) Common hypothesis of gene regulation by DNA methylation of the promoter. (Top) Gene with unmethylated promoter is expressed. (Bottom) Some cytosines on the promoter sequence get methylated, which inhibits the expression of the corresponding gene.

fectively, thymines detected in bisulfite converted reads were originally either real thymines or unmethylated cytosines in the genome. The conversion rate is not perfect but generally above 95% [Holmes et al., 2014]. Specially designed reads mappers, like Bismark [Krueger and Andrews, 2011] or BSMAP [Xi and Li, 2009] are then used to map the bisulfite reads on the genome (Fig. 1.13).

Once the bisulfite reads mapping is performed, the extraction of methylation rate is performed for every covered cytosine on the genome (Fig. 1.14a). At a given position, the number of reads presenting a cytosine and the number of reads presenting a thymine are counted and the corresponding ratio is computed. A number between 0 and 1 corresponding to the number of "C reads" divided by the coverage ("C reads" + "T reads") is attributed to each position.

Theoretically, on a single haploid cell and with 100% bisulfite conversion rate, the methylation ratio should always be equal to either zero or one. The fact we have multiple diploid cells (two methylation alleles) which can have different methylation states, the biological variation between cells and tissues, the imperfect C to T conversion rate and the noise generated by the mapping because of repetitive regions in the genome can generate some values in-between. However the distribution of the ratios (Fig. 1.14b) shows that the values between zero and one excluded are a minority.

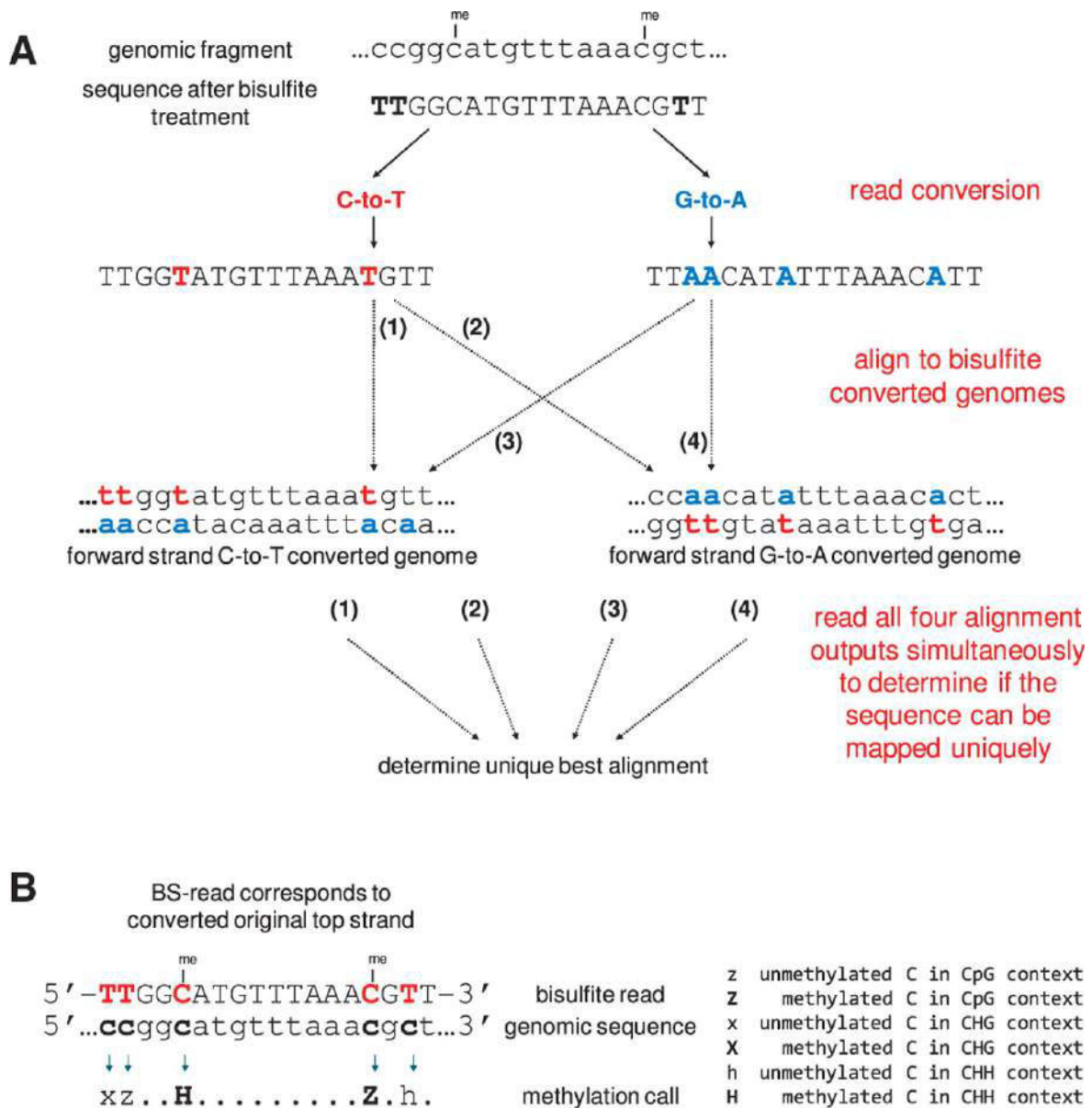


Figure 1.13 Bismark's approach to bisulfite mapping and methylation calling. (a) Reads from a BS-Seq experiment are converted into a C-to-T and a G-to-A version and are then aligned to equivalently converted versions of the reference genome. A unique best alignment is then determined from the four parallel alignment processes [in this example, the best alignment has no mismatches and comes from thread (1)]. (b) The methylation state of positions involving cytosines is determined by comparing the read sequence with the corresponding genomic sequence. Depending on the strand a read mapped against this can involve looking for C-to-T (as shown here) or G-to-A substitutions. Figure taken from [Krueger and Andrews, 2011](#)

### 1.4.3 Differential methylation analysis

The most common approach to detect methylation differences between two samples is to find Differentially Methylated Regions (DMRs) from Whole Genome Bisulfite Sequencing (WGBS) data. For each condition, the whole genome undergo bisulfite sequencing, the methylation levels are measured at single base resolution

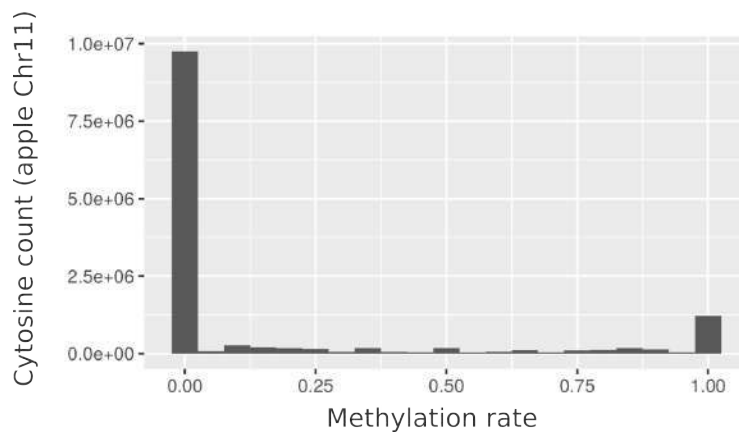
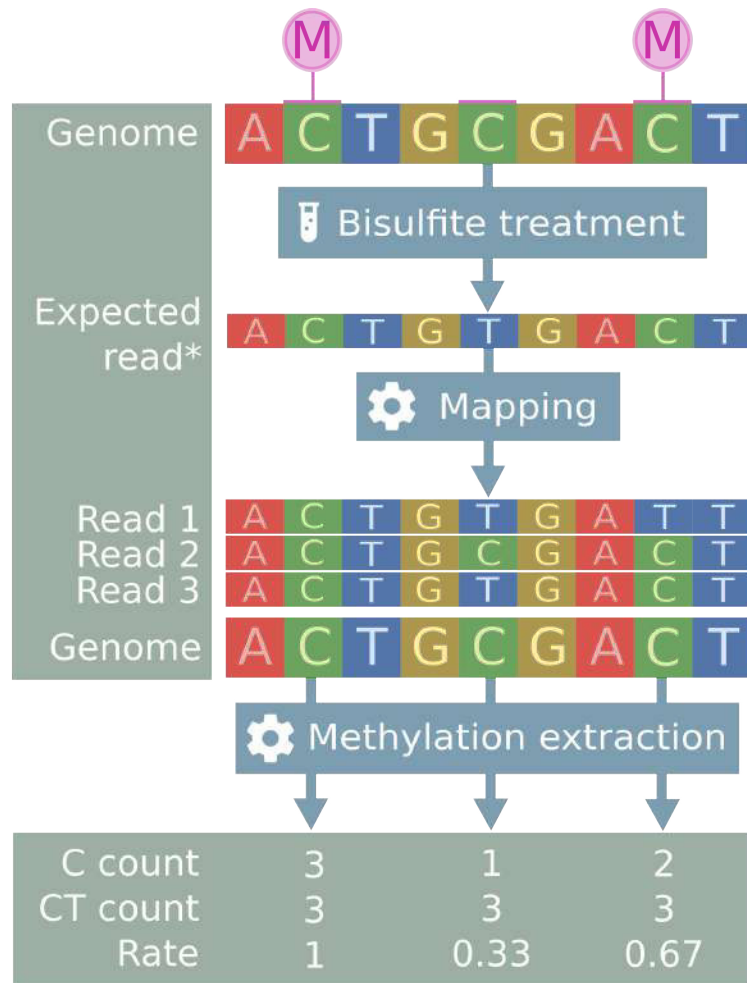


Figure 1.14 Methylation calling process illustration. (a) Bisulfite data processing summary. The genomic DNA is extracted and a bisulfite treatment is applied on the DNA fragment before PCR, which converts unmethylated cytosines into thymines. The reads are then mapped on the genome, and for each genome's cytosine, the ratio of reads which presents a cytosine (sign of methylation) is computed, resulting in a DNA methylation ratio for each individual cytosine on the genome. (b) Methylation ratios distribution in the chromosome 11 of apple. Bisulfite data coming from a leaf sample of GDDH13 was mapped on the genome and processed. Methylation ratios along the chromosome were counted to plot their distribution.

(see part [1.4.2](#)) as a number of methylated reads and a number of unmethylated reads, and compared between the two samples. Every locus on the genome where the methylation levels are significantly different between the two samples will be called a DMR.

Several DMRs finding tools are published ([Table 1.2](#)) using a wide range of statistical methods.

Each bisulfite sequencing experiment is performed on a population of cells. DNA methylation levels have a high biological variability between different cells and tissues [\[Robinson et al., 2014\]](#). Thus, several biological replicates are required in order to find robust and specific DMRs. There are two classes of DMR finding methods : methods which find Differentially Methylated Cytosines (DMC) first and merge DMC dense fragments into DMRs, and methods which directly compute methylation differences on regions.

### DMCs methods

For each covered cytosine on the genome, the methods of this class use a statistical model to determine if the cytosine is differentially methylated between each conditions. Thus, they need several biological replicates per condition to have enough values to account for the biological variability. However, if enough replicates are possible, they allow to find methylation differences at the cytosine level and generally more robust DMRs. Once all the DMCs are found, they find DMC rich regions and classify them as DMRs. Since the observations of methylation proportion are binomial distributed on a particular site [\[Robinson et al., 2014\]](#), the most common statistical model used by these method is a beta-binomial regression. This model is used in several published tools, such as DSS [\[Feng et al., 2014\]](#), BiSeq [\[Hebestreit et al., 2013\]](#), RAD-Meth [\[Dolzhenko and Smith, 2014\]](#), or methylSig [\[Park et al., 2014\]](#).

### Direct DMRs methods

Applying a statistical test directly to regions, which represent many cytosines at once, is an alternative method to find DMRs which requires less biological replicates because all cytosines of a region are considered altogether during testing which increases the test's power. Some of these methods achieve a decent specificity rate while working with as few as two biological replicates for each condition. These usually consist in two major steps : (1) determining the boundaries on regions on which the statistical model will be applied and (2) applying the statistical model on chosen regions. There exist different methods for windows defining step. The most straightforward approach is to use predefined regions, such as known gene boundaries, or on the entirety of a genome using sliding windows for example. This approach is performed by several tools such as methylSig [\[Park et al., 2014\]](#), methylKit [\[Akalin et al., 2012\]](#) or COHCAP [\[Warden et al., 2013\]](#). Once the windows are defined, a statistical test is applied in order to compare the different conditions. Several types of statistics are used by the different methods, among them the Student test, the Wilcoxon test or the Fisher

exact test.

### Specificity problems in DMRs finding

The major problem encountered in DMRs finding is the high rate of false positives, especially when the number of biological replicates is low, due to the natural biological variability of methylation. A recent study [Hesse et al., 2015] performed a comparative analysis of three DMRs tools : RADMeth [Dolzhenko and Smith, 2010], BSmooth [Hansen et al., 2012] and their own method, using parameters adapted to find DMRs containing roughly the same number of DMLs. The three methods each found a very different set of DMRs ( Fig. 1.15) which suggests that merging the results of different tools may be appropriate to obtain more specific DMRs at the cost of computing time.

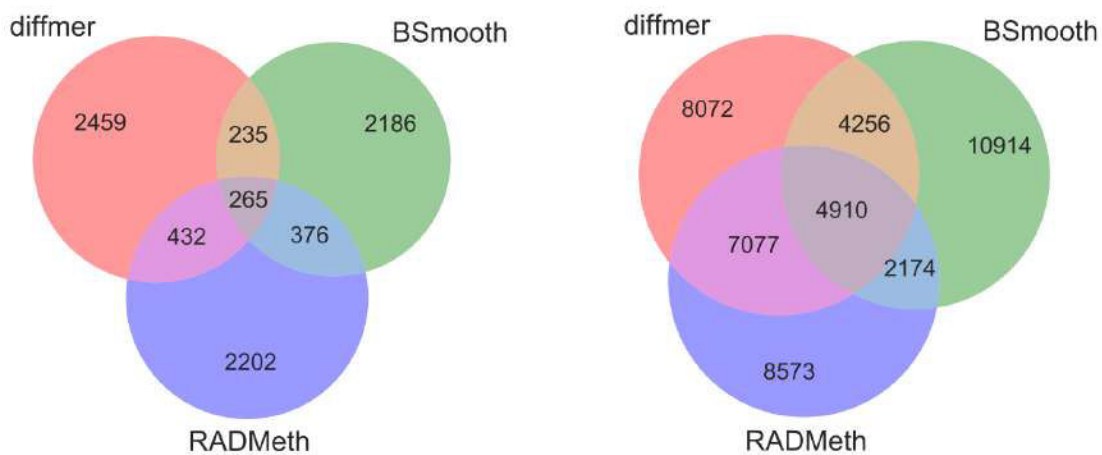


Figure 1.15 Venn diagrams comparing the number of DMLs in DMRs detected by diffmer, BSmooth and RADMeth. Left: macaques; right: uveal melanoma. Parameters were chosen to obtain an approximately equal number of DMCs in DMRs for each method. Figure taken from [Hesse et al., 2015]

#### 1.4.4 Thesis objectives

Using the apple genome generated in the first part of this work as a reference, the first objective of this part was to produce a genome-wide methylome map at nucleotide-level resolution for *Golden Delicious* apple in various conditions. Second, we compared the methylomes of apples having different sizes in order to find some genes playing a potential role in fruit development. Third, we generated an easy-to-use and complete pipeline to identify differentially methylated regions between two methylomes.

| Method     | Citation                                  | Designed for   | Test on regions | Covariates | Used statistics                       |
|------------|---|----------------|-----------------|------------|---------------------------------------|
| Minfi      | <a href="#">Aryee et al., 2014</a>        | 450k           | Determines      | Yes        | Bump hunting                          |
| IMA        | <a href="#">Wang et al., 2012</a>         | 450k           | Predefined      | No         | Wilcoxon                              |
| COHCAP     | <a href="#">Warden et al., 2013</a>       | 450k or BS-seq | Predefined      | Yes        | FET, t-test, ANOVA                    |
| BSmooth    | <a href="#">Hansen et al., 2012</a>       | BS-seq         | Determines      | No         | Bump hunting on smoothed t-like score |
| DSS        | <a href="#">Feng et al., 2014</a>         | BS-seq         | Determines      | No         | Wald                                  |
| MOABS      | <a href="#">Sun et al., 2014</a>          | BS-seq         | Determines      | No         | “Credible methylation difference”     |
| BiSeq      | <a href="#">Hebestreit et al., 2013</a>   | BS-seq         | Determines      | Yes        | Wald                                  |
| DMAP       | <a href="#">Stockwell et al., 2014</a>    | BS-seq         | Predefined      | Yes        | ANOVA, Chi-squared, FET               |
| methyKit   | <a href="#">Akalin et al., 2012</a>       | BS-seq         | Predefined      | Yes        | Logistic regression                   |
| RADMeth    | <a href="#">Dolzhenko and Smith, 2014</a> | BS-seq         | Determines      | Yes        | Likelihood-ratio                      |
| methySig   | <a href="#">Park et al., 2014</a>         | BS-seq         | Predefined      | No         | Likelihood-ratio                      |
| Bumphunter | <a href="#">Jaffe et al., 2012</a>        | General        | Determines      | Yes        | Permutation, smoothing                |
| ABCD-DNA   | <a href="#">Robinson et al., 2012</a>     | MeDIP-seq      | Predefined      | Yes        | Likelihood ratio                      |
| DiffBind   | <a href="#">Ross-Innes et al., 2012</a>   | MeDIP-seq      | Predefined      | Yes        | Likelihood ratio                      |

Table 1.2 List of recent methods to detect differentially methylated loci or regions. Adapted from [Robinson et al., 2014](#).

## Chapter 2

# Genome assembly and annotation

### 2.1 Introduction

Accurate sequence information, genome assemblies and annotations are the foundation for genetic and genome-wide studies. The major factors that limit *de novo* genome assembly are heterozygosity and repetitive sequences, such as TEs, which are often collapsed to single copies in draft genomes [Veeckman et al., 2016]. In recent years, however, evidence supporting the importance of TEs in genome evolution, genome structure, regulation of gene expression and epigenetics has been mounting [Consortium et al., 2012a] [Fedoroff, 2012] [Chénaïs et al., 2012]. The characterization of sequences and the distribution of TEs within a genome is, therefore, of great importance. Until now, the study of epigenetically controlled characteristics in perennial plants has been hampered by the draft status of their genome sequences. In the case of apple, a draft was produced [Velasco et al., 2010] but remained incomplete with inaccurate contig positions [Khan et al., 2012]; this hindered its utility for genetic and epigenetic studies. *de novo* sequencing and assembly of a new genome for apple, using technologies of the third generation, had thus become a necessity.

In the last few years, single-molecule sequencing and optical-mapping technologies have emerged [Ansorge, 2016], which are well suited for assembling genomic regions that contain long repetitive elements. Recently, several high-quality genome assemblies have been published using one or both technologies [Zhang et al., 2015] [VanBuren et al., 2015] [Zapata et al., 2016] [Redwan et al., 2016] [Mahesh et al., 2016] [Badouin et al., 2015]. The use of long-read sequencing technologies may also tackle potential assembly issues that are related to the presence of highly similar sequences resulting from whole-genome duplication events that frequently occurred in angiosperm genomes [Cui et al., 2006].

To produce a high-quality apple reference genome, we generated a *de novo* assembly of a ‘Golden Delicious’ doubled haploid tree (GDDH13), which results from a spontaneous chromosome doubling of a haploid tree, hence is entirely homozygote. The assembly is composed of 280 assembled scaffolds and arranged into 17

pseudomolecules, which represent the 17 chromosomes of apple. This assembly resulted from a combination of short (Illumina) and long sequencing reads (PacBio), along with scaffolding based on optical maps (BioNano) and a high-density integrated genetic linkage map [Di Pierro et al., 2016]. This chromosome-scale assembly was complemented by a detailed de novo annotation of genes based on RNA sequencing (RNA-seq) data, TE annotation and small RNA alignments. This work provides a solid foundation for future genetic and epigenomic studies in apple. Furthermore, our TE annotation provides novel insights into the evolutionary history of apple and may contribute to explaining its divergence from pear.

## 2.2 Methods

### 2.2.1 Sequencing (done by collaborators)

#### Plant material

*Previously done by collaborators in Angers*

Origin of the two doubled-haploid 'Golden Delicious' apple trees was described by [Lespinasse et al., 1996]. Hereto, among others, a 'Golden Delicious' progeny P21R1A50 deriving from a self-pollination of 'Golden Delicious' (1963) was self-pollinated (1986). At this time an ovule with an unfertilized egg rather than a zygote developed into a haploid plant of which leaves spontaneously produced two independent and simultaneous chromosome doubling events in vitro, named GDDH13 and GDDH18 (Fig. S1). These plants were then rooted and grown in the orchard (1989, first fruits in 1995).

#### DNA purification

*Done by Jean-Marc Celton in Angers*

For Illumina sequencing, genomic DNA was purified from young leaves using Macherey-Nagel NucleoSpin plant II DNA extraction kit (Germany), following the manufacturer's instructions. For BioNano and PacBio single Molecule Real Time Sequencing, genomic DNA was extracted using a modified nuclei preparation method 25 followed by an additional phenol-chloroform purification step.

#### Illumina Whole-genome shotgun sequencing

*Done by collaborators in San Michele All'adige and Angers*

One Paired-end library with an insert size of 350 bp was constructed with the Truseq DNA Library Prep Kit for Illumina according to the manufacturer's protocol. This library was sequenced on an Illumina HiSeq 2000 platform and yielded 45 Gb of paired 150 bp reads. A second Paired-end library with an insert size of 300bp was constructed with the Truseq DNA library preparation kit for Illumina according to the manufacturer's instructions. It was sequenced on an Illumina HiSeq 2000 platform and yielded 41.5



Gb of raw data as paired 100bp reads. Truseq adaptor sequences were removed from both libraries using scythe software (<https://github.com/vsbuffalo/scythe>). Reads were screened by alignment to the PhiX and E.Coli genomes and the published apple mitochondrial and chloroplast sequences [Velasco et al., 2010] using BWA mem with default settings. All Illumina reads were then subjected to kmer spectrum based error correction using SoapEC [Luo et al., 2012b] with a kmer size of 23 and all other parameters set at default. Three Illumina mate pair libraries, with target insert sizes of 2, 5 and 10 Kb, were prepared according to the Illumina Nextera Mate-pair protocol and sequenced on two lanes of an Illumina HiSeq 2000 platform yielding 82 Gb of raw sequence data as paired 100 bp reads. Mate pair data was processed using the NxTrim software [O'Connell et al., 2015] to remove short fragment read pairs in forward reverse orientation leaving only true mate pair fragments in forward reverse orientation. The FastUniq software [Xu et al., 2012] was subsequently used to remove duplicate read pairs. After cleaning and deduplication 8.5 Gb of data was available for scaffolding. Insert sizes of the Mate-pair libraries were estimated empirically by alignment to Illumina contigs over 10kb using the smalt aligner (<http://www.sanger.ac.uk/science/tools/smalt-0>) with independent mapping of forward and reverse reads.

### **PacBio single molecule real time sequencing**

*Done by collaborators in Wageningen and Angers*

In total twenty microgram of gDNA was sheared by a Megaruptor (Diagenode) device with 30 Kb settings. Sheared DNA was purified and concentrated with AmpureXP beads (Agencourt) and further used for Single-Molecule Real Time (SMRT) bell preparation according to manufacturer's protocol (Pacific Biosciences; 20-Kb template preparation using BluePippin size selection (Sagescience)). Size selected and isolated SMRT bell fractions were purified using AmpureXP beads and finally 20 nanogram of these purified SMRT bells were used for primer- and polymerase (P6) binding according to manufacturer's binding calculator (Pacific Biosciences). DNA-Polymerase complexes were used for Magbead binding and loaded at 0.1nM on-plate concentration spending 16 SMRT cells. Final sequencing was done on a PacBio RS-II platform, with 240 minutes movie time, one cell per well protocol and C4 sequencing chemistry. Raw sequence data was imported and further processed on a SMRT Analysis Server V2.3.0.

### **BioNano genomics genome mapping**

*Done by collaborators in Wageningen*

Agarose plug embedded nuclei were Proteinase K treated for two days followed by RNase treatment (Biorad CHEF Genomic DNA Plug Kit). DNA was recovered from agarose plugs according to IrysPrep™ Plug Lysis Long DNA Isolation guidelines (BioNano Genomics). Of the isolated DNA, 300 nano gram was used for subsequent DNA nicking using Nt.BspQ1 (NEB) incubating for 2 hours at 50°C. Labelling, repair and staining

reactions were done according to IrysPrep™ Assay NLRs (30024D) protocol. Finally, ultra-high molecular weight (U-HMW) NLRs DNA molecules were analyzed on two BioNano Genomics Irys instruments with optimized recipes using two Irys chips, three flowcells, twelve runs, for a total of 344 cycles. Data was collected and processed using IrisView software V 2.5 together with a XeonPhi (version v4704) accelerated cluster and special software (both BioNano Genomics, Inc.). A de novo map assembly was generated using molecules equal or bigger than 230 Kb, and containing a minimum of five labels per molecule. In total all molecules used for assembly encompassed 162 Gb equivalent space. For the assembly process, stringency settings for alignment and refineAlignment were set to 1e-8 and 1e-9 respectively. The assembly was performed by applying five iterations, where each iteration consisted of an extension and merging step. Hybrid scaffolding was done using "hybrid scaffolding\_config\_aggressive" of Irys View with minimal 80 Kb contig size.

## **mRNA-Seq**

*Done by collaborators in Angers*

To maximize the number and diversity of genes identified by RNA-Seq, mRNA was purified from various organs at multiple developmental stages derived from seven cultivars and hybrids. A total of 9 libraries were generated and included cDNA derived from roots ('Galaxy'), stem ('Granny Smith'), leaves (hybrid M49, pedigree described in [Segonne et al., 2014](#)), apex ('Granny Smith'), seedlings (derived from 'Golden Delicious' open pollinated), flowers ('Gala'), and parenchyme from mature fruits (two biological repetitions of hybrid M74, and one sample from hybrid M20, pedigrees of both hybrids presented in [Segonne et al., 2014](#)). With the exception of parenchyme fruit samples, RNA extraction was performed using the NucleoSpin RNA Plant extraction kit (Machery-Nagel, Germany). For fruit samples, total RNA was purified according to [Nobile et al., 2011](#). Nucleic acids were quantified (NanoDropTechnologies Inc., Wilmington, USA) and their quality was checked by electrophoresis on 1% agarose gel and stained in ethidium bromide. RNA were then treated with RQ1 DNase at 37°C for 10 min, and RQ1 DNase Stop Solution at 65°C for 10 min (Promega). The cDNA sequencing libraries were constructed following the manufacturer's instructions (Illumina, San Diego, CA, USA). Fragments of 200 to 350 bp were excised, enriched by 15 PCR cycles, and loaded onto flowcell channels at a concentration of 8 to 10 pM. Paired-end reads of varying length were generated (from 100 to 300 bp). The Illumina GA processing pipeline Cassava 1.7.0 was used for image analysis and base calling. Library preparation and final quality control of sequencing data of nine samples including leaves, roots, mature fruits, apex, stem, seedling and flower, were performed by the INRA-EPGV group while sequencing on GAIIx was implemented by the sequencing group of CEA-IG/CNG.

## 2.2.2 Genome assembly

**Hybrid assembly** The genome assembly was performed using a combination of sequencing technologies: PacBio RS II reads, Illumina paired-end reads (PE) and Illumina mate-pair reads (MP). First, the corrected Illumina PE reads were separately assembled using SoapDevo 2.223 [Luo et al., 2012b] in multi kmer mode with all kmer values from 51 to 127 and filtering out kmers with frequency lower than 3 prior to assembly. Since a doubled-haploid plant was sequenced we avoided the merging of similar sequences (bubble popping) during contig assembly (-M parameter SoapDeNovo). Next, the PacBio reads (24 Gb, approximate sequencing depth = 37X) and Illumina contigs were combined to perform a hybrid assembly using the DBG2OLC pipeline [Ye et al., 2016] with the following parameters: kmer size 17 as advised by the authors, removeChimera parameter 1. A broad range of the three critical parameters (AdaptiveTh, KmerCovTh and MinOverlap) were tested in different combinations in a way to optimize the N50 and to match the assembly as closely as possible to the expected genome size. The final used parameters were AdaptiveTh 0.005, KmerCovTh 3, MinOverlap 20. Reads were mapped to contigs with blasr [Chaisson and Tesler, 2012] before calling a consensus sequence with Sparc [Ye and Ma, 2016]. Parameter sweeps were performed for the critical DBG2OLC parameters in order to optimize the N50 and the Assembly size.

**Assembly polishing** A polishing of the assembly using the Illumina paired-end reads was performed. The 120X Illumina reads were mapped to the contigs using BWA-MEM v.0.7.12-r1044 [Li, 2013]. This alignment was then used with Pilon v1.17 [Walker et al., 2014] which computed a consensus base for each position. This process was performed twenty times iteratively.

**Mate-pair scaffolding** A total of 8.5 Gb of Illumina mate pair (MP) data (approximate sequencing depth = 15X), with an insert size varying between 2 kb and 10 kb was used to scaffold the assembly. The MP reads were mapped on the corrected contigs using BWA-MEM v.0.7.12-r1044. The alignments were used by BESST [Sahlin et al., 2014] using the default parameters.

### BioNano scaffolding

*done by collaborators in Wageningen*

A BioNano optical mapping was performed. Optical map reads were generated with the process previously described. Approximately 600 fold coverage of optical maps reads were generated and assembled in 397 BioNano maps (equivalent to BioNano contigs) with a N50 of 2.649 Mb and a total length of 649.7 Mb. The optical maps were used in a hybrid assembly with the scaffolds obtained from the mate-pair scaffolding to assemble the final scaffolds using the BioNano Irys software.

**Scaffold validation and anchoring to genetic map** An integrated multi-parental genetic linkage map of apple [Di Pierro et al., 2016] was used to organize and orientate the scaffolds and contigs into chromosome-sized sequences and to assess the quality of the assembly. The high-density linkage map, with a length of 1,267 cM, was produced in the framework of the EU-funded FruitBreedomics project, based on data from 21 full-sib families totaling 1,586 progenies and the 20K SNP Infinium<sup>®</sup> array [Bianco et al., 2014]. It is composed of 15,417 SNP markers which cluster into haploblocks from 10 Kb to 100 Kb that comprise up to 15 SNPs, and occur at 1 cM intervals along the genome. The probe sequence of the 15,417 markers were mapped on the genome using BWA-MEM v.0.7.12-r1044. The linkage group found for the majority of the mapped markers for a scaffold or contig was attributed to it. The position of each sequence relative to other sequences on the same linkage group was determined by the median position of the mapped markers on this sequence. The orientation of the scaffold and contigs was determined by the most common orientation indicated by all possible pairs of mapped markers when considering their order on the integrated genetic map, if at least two markers were mapped on the sequence.

**Illumina-based genome size estimation** Error corrected reads from the 150bp paired-end Illumina library were selected to perform genome size estimation. The library was submitted to 23 mer frequency distribution analysis using Jellyfish [Rizk et al., 2013]. The single peak obtained from the GDDH13 genome and corresponding to a kmer depth of 41 was used for genome size estimation. Based on the total number of kmers (26,715,896,120), the GDDH13 genome size was calculated using the following formula:

genome size = kmer\_Number/Peak\_Depth.

## Linkage disequilibrium

*Done by collaborators in Angers*

The "Old Dessert" INRA core collection, comprising 278 accessions [Lassois et al., 2016], was genotyped with the Axiom<sup>®</sup> Apple-480K SNP genotyping array [Bianco et al., 2016] as part of ongoing genome-wide association analyses. 264,861 markers out of the 275,076 markers (96%) polymorphic in the INRA core collection were localized at unique positions on the genome using BWA-MEM v.0.7.12-r1044. Linkage disequilibrium was estimated with the  $r^2$  statistics using the R package snpStats [Clayton and Leung, 2007] (R package version 1.16.0). Heatmaps of pairwise LD between markers were plotted using the R package LDheatmap [Shin et al., 2006]. For each chromosome, one marker every ten was used to illustrate LD at a whole genome scale.

### 2.2.3 Genome annotation

**Structural and functional gene annotation** RNA-seq data derived from nine different libraries, including six different organs (leaves, roots, fruits, apex, stems and flowers) was de novo assembled using Trinity [Grabherr et al., 2011] and SOAPdenovo-trans [Xie et al., 2014]. For each library, the assembly with the highest N50 was chosen to annotate the genes. 2,033 mRNAs and 326,941 EST extracted from the NCBI nucleotide and EST databases respectively were also used for gene prediction. Using the Eugene pipeline, repeat sequences were masked using LTRharvest [Ellinghaus et al., 2008], Red [Girgis, 2015] and BLASTx comparisons against Repbase [Bao et al., 2015]. The structural annotation of coding genes was performed using EuGene [Foissac et al., 2008] by combining Gmap transcript mapping [Wu and Watanabe, 2005], similarities detected with plant proteomes and Swiss-Prot, and ab initio predictions (Interpolated Marlow Model and Weight Array Matrix for donor and acceptor splicing sites). Moreover, the EuGene prediction has been completed by tRNAscan-SE [Lowe and Chan, 2016], RNAmmer [Lagesen et al., 2007] and RfamScan [Nawrocki et al., 2014] in order to annotate non-protein coding genes. Functional annotation of proteins was performed using InterProScan [Jones et al., 2014]. Additionnal functional data was generated using the best (if existing) Blastp hit against TAIR and Swissprot to attribute keywords succinctly describing each gene function if possible. The functional annotation was then completed by the prediction of targeted signals using the TargetP software [Emanuelsson et al., 2007].

### Comparison of annotation between the heterozygous 'Golden Delicious' and GDDH13 genomes

*Malus domestica* predicted genes (MDP) sequences obtained from the heterozygous genome annotation [Velasco et al., 2013] were mapped to the GDDH13 genome assembly using the best BLAT [Kent, 2002] hit including the following parameters: a minimum of 20% overlap between MDP sequence and new de novo predicted genes was required, with a minimum 96% base identity. Comparison of the two genome annotations was done using Bio++ [Guéguen et al., 2013].

**smallRNA alignment** Apple sRNA derived from mature fruit parenchyme [Celton et al., 2014] were aligned to the 'Golden Delicious' doubled-haploid pseudo-molecules using BWA-MEM v.0.7.12-r1044. Only perfectly mapped sequences were considered further (no SNP between sRNA sequence and target sequence), and reads with identical sequences were allowed to be mapped to two or more loci.

### 2.2.4 Genome synteny

SynMap (CoGe, [www.genomeevolution.org](http://www.genomeevolution.org)) was used to identify collinearity blocks using homologous CDS pairs using the following parameters: Maximum distance between two matches (-D): 20; Minimum number of

aligned pairs (-A): 10; Algorithm “Quota Align Merge” with Maximum distance between two blocks (-Dm): 500.

## 2.3 Results

A first version of the genome (GDDH13 V1.0) was made and corresponds to the published version in [Daccord et al., 2017](#). This version was improved in the latest release (GDDH13 V1.1). This section will first describe the published results (GDDH13 V1.0) then explain the motives and methods behind the latest version (GDDH13 V1.1).

### 2.3.1 GDDH13 V1.0

**Homozygosity of the doubled-haploid and genome size estimation** The doubled-haploid Golden Delicious line (GDDH13, also coded X9273) used in this study is the result of breeding efforts that were initiated at INRA in 1963. Homozygosity of this line was confirmed with microsatellite markers that are distributed along the apple genome (data not shown) and by observation of the k-mer spectrum of Illumina reads derived from GDDH13. In [Fig. 2.1a](#) k-mer spectra of GDDH13 and of the heterozygous ‘Golden Delicious’ [Li et al., 2016](#) are compared. Two peaks are clearly visible for the heterozygous cultivar (one containing heterozygous k-mers and the other with double coverage comprising k-mers shared by the two haplotypes) and only one peak is seen for the doubled-haploid.

We estimated the genome size of GDDH13 to be 651 Mb ([Table S1](#)), which suggested that the GDDH13 genome may be smaller than that of the heterozygous Golden Delicious line, which was recently estimated to be 701 Mb [Li et al., 2016](#).

**Genome assembly** To perform de novo assembly of the GDDH13 genome, we combined three different technologies: short-read sequencing, long-read sequencing and optical mapping ([Fig. 2.1b](#)). Using DNA from the leaves of GDDH13, we generated 120-fold coverage of Illumina paired-end reads (72 Gb), 80-fold coverage of Illumina Nextera mate-pair reads (58 Gb) at three different insert sizes (2, 5 and 10 kb) and 37-fold coverage of PacBio sequencing data (24 Gb; 2,837,045 subreads with a mean length of 8,474 bp). The Illumina paired-end reads were first assembled using SOAPdenovo [Luo et al., 2012b](#), and the resulting contigs ([Table 2.1](#)) were combined with the PacBio reads using the DBG2OLC assembler [Ye et al., 2016](#). A consensus step was performed using the raw hybrid contigs, the Illumina contigs, and the PacBio reads with Sparc [Ye and Ma, 2016](#).

**Assembly polishing** An correction procedure using Illumina paired-end reads (150bp, 120X) was performed with Pilon [Walker et al., 2014]. On the first Pilon run, 94,896 single-base assembly errors ; 1,054,709 insertions (1,466,015 bp) and 123,510 deletions (178,733 bp) were corrected.

**Assembly scaffolding** A first scaffolding was performed using Illumina mate pair reads (15X) with BESST [Sahlin et al., 2014]. The final scaffolding was performed using BioNano optical maps to obtain the final assembly. The genome assembly metrics for each step are reported **Table 2.1**. Finally, BioNano scaffolds were assembled into pseudo-molecules using an integrated linkage map of 15,417 markers [Di Pierro et al., 2016].

|                             | <b>Illumina<br/>assembly<br/>SOAPdenovo</b> | <b>Hybrid<br/>assembly<br/>DBG2OLC</b> | <b>Mate-pair<br/>scaffolding<br/>BESST</b> | <b>BioNano<br/>scaffolding</b> |
|-----------------------------|---|--|--|--------------------------------|
| <b>Number of sequences</b>  | 5,042,943                                   | 2,150                                  | 1,832                                      | 1,081                          |
| <b>Number of Bases (Mb)</b> | 1,316                                       | 625.2                                  | 625.5                                      | 649.7                          |
| <b>N50 (Kb)</b>             | 7.289                                       | 620                                    | 699  | 5,558                          |
| <b>L50</b>                  | 20,863                                      | 315                                    | 277  | 39                             |

Table 2.1 Metrics of the different steps performed for the GDDH13 genome assembly. The N50 number corresponds to the length of the median-sized contig if all contigs are sorted by size. The L50 number corresponds to the required number of largest contigs to obtain 50% of the assembly length.

**Assessment of genome quality** We assessed the quality of the assembly by using the SNP markers that were mapped on the previously mentioned integrated genetic linkage map. Of the 15,417 SNP probe sequences, we identified sequence homology in the GDDH13 genome for 14,732 of them. We then assessed their position on the scaffold assemblies by comparing their location on the integrated genetic linkage map (**Fig. 2.1c**). In total 14,117 of the mapped markers (95.8%) were found to be located at their expected positions (**Fig. S2**). In total, we identified 685 SNP probes without homology in the GDDH13 genome assembly (4.5% of the markers). These markers were found to be randomly distributed along the 17 linkage groups of the genetic linkage map. We also identified several markers showing discrepancy between their position within scaffolds and the genetic map. These markers were summed up to 47 groups that represented a total of 3.37 Mb (0.45% of the assembly; corresponding SNP markers have been flagged in the GDDH13 genome browser).

**Genome annotation** To obtain a global view of the apple transcriptome, we performed a high-throughput RNA-seq analysis on poly(A)-enriched RNAs from nine libraries that originated from different genotypes and tissues. RNA-seq reads were assembled, and the resulting contigs were mapped to the scaffolds and inte-

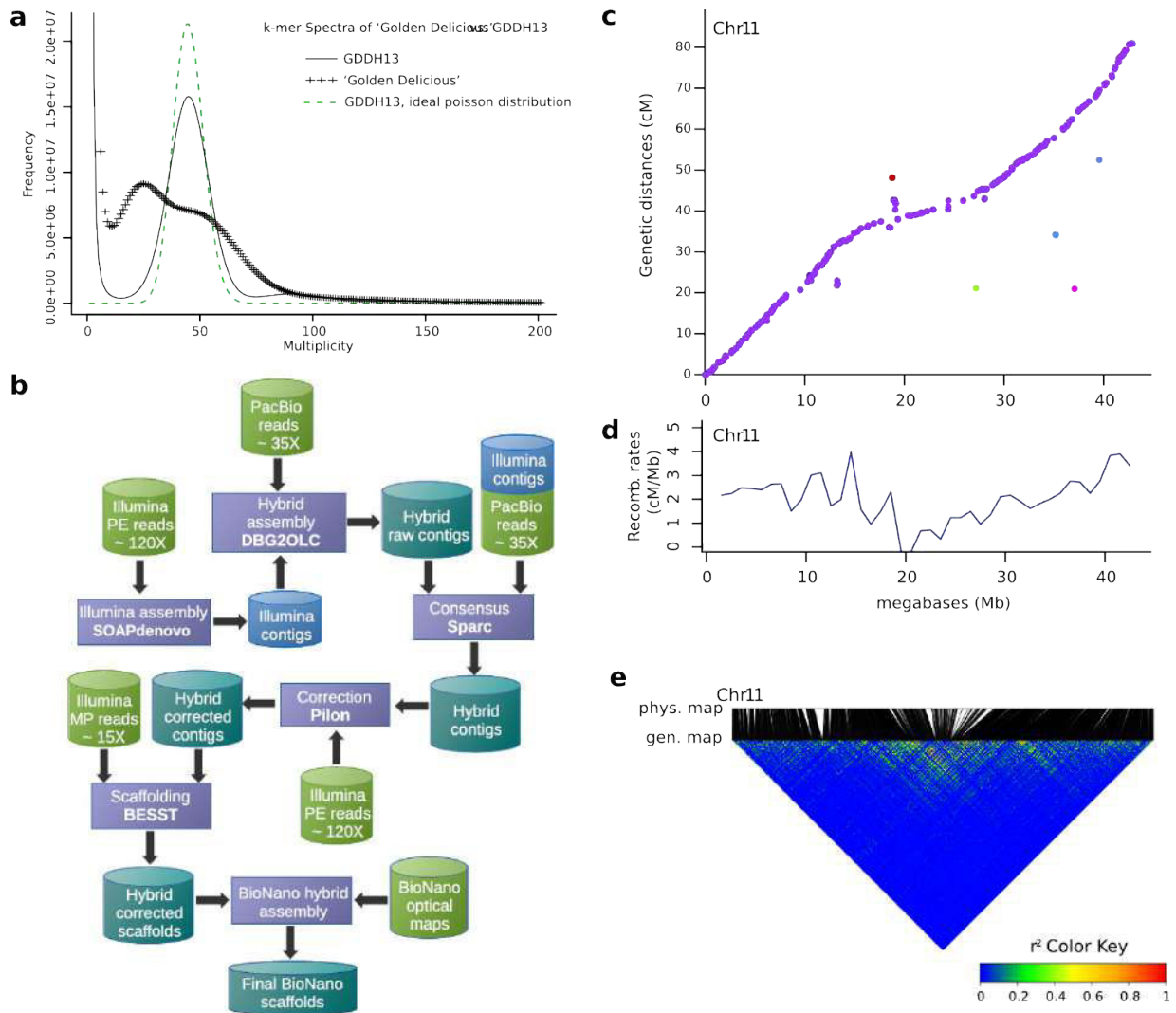


Figure 2.1 Assembly and validation of the GDDH13 doubled-haploid apple genome. **(a)** k-mer (23 bp) spectra of the doubled-haploid GDDH13 and the heterozygous Golden Delicious 33 genomes. The x axis represents k-mer multiplicity, and the y axis represents the number of k-mers with a given multiplicity in the sequencing data. The green dashed line represents the ideal Poisson distribution fitted on the data of GDDH13. **(b)** Overview of the processing pipeline used for the assembly of the GDDH13 genome (see Supplementary Note for details). **(c)** Graphical representation of the location of SNP markers on the physical map (x axis), as compared to their position on the integrated genetic map (y axis), for Chr11 of the GDDH13 genome. Each marker is depicted as a circle on the plot (1,069 data points). The colors depict the chromosomes as follows: red for Chr01, light green for Chr04, pink for Chr08, blue for Chr10 and violet for Chr11. **(d)** Graphical representation of the mean local recombination rates between successive SNP markers along Chr11 (3-Mb sliding window, 1-Mb shift, threshold 4). The x axis represents the physical positions of the means on Chr11, and the y axis indicates the recombination ratio (centiMorgan (cM)/Mb) in each 3-Mb sliding window. **(e)** Heat map of genotypic linkage disequilibrium (LD;  $r^2$ ) in Chr11 in the 'Old Dessert' INRA apple core collection. Shown are the graphical representation of the location of SNPs on the physical map (top) with correspondence to their order in a regular distribution (bottom) of Chr11 (1,461,195 data points). The color bar indicates the level of LD, from high LD (red) and low LD (blue).



grated in the EuGene combiner pipeline [Foissac et al., 2008]. In total, we identified 42,140 protein-coding genes (which represent 23.3% of the genome assembly) and 1,965 non-protein-coding genes. Evidence of transcription was found for 93% of the annotated genes. To further evaluate the quality of the annotation, a comparison with annotations of previous apple genome assemblies ([Li et al., 2016], [Velasco et al., 2010]) was performed using the BUSCO v2 method, which is based on a benchmark of 1,440 conserved plant genes [Simão et al., 2015]. The results indicate that our apple genome annotation is the most complete, despite having the lowest number of predicted genes (Table 2.2).

The de novo annotated genes were named using the following convention: MD (for *Malus domestica*) followed by the chromosome number and gene number on the chromosome (in steps of 100) going from top to bottom according to the linkage map, for example, MD13G0052100.

At least one Gene Ontology (GO) annotation was assigned to 63.4% of the newly predicted genes: 14,799 genes were tagged by ‘Biological Process’ GO term(s), 22,560 genes by ‘Molecular Function’ GO term(s) and 6,574 genes by ‘Cellular Component’ GO term(s). For gene family classification, 83.6% of genes matched to a domain signature according to at least one database of the Interpro consortium. Regarding only the PFAM resource [Finn et al., 2015], 32,109 genes (76%) were distributed among 3,853 gene families.

Previously published small RNA (sRNA) data [Celton et al., 2014] were also mapped to the genome. We found that most 21- and 22-nt-long sRNAs mapped to protein-coding genes, whereas most 24-nt-long sRNAs mapped to TEs. The distribution of 23-nt-long sRNAs was uniform in both types of genomic features (Fig. S4).

|   | This study | Velasco et al. (2010) | Li et al. (2016) |
|---|------------|-----------------------|------------------|
| <b>Total Number of Bases (Mb)</b>                       | 643.2      | 603.9                 | 632.4            |
| <b>N50 (Kb)</b>   | 5,558      | 16                    | 112              |
| <b>Annotated protein coding genes</b>                   | 42,140     | 63,541                | 53,922           |
| <b>Transposable elements proportion (%)</b>             | 57.3       | 42.4                  | NA               |
| <b>Pearson correlation coefficient with genetic map</b> | 0.90       | 0.67                  | NA               |
| <b>Complete BUSCOs</b>                                  | 94.9%      | 86.7%                 | 51.5%            |
| <b>Fragmented BUSCOs</b>                                | 2.6%       | 5.6%                  | 18.8%            |
| <b>Missing BUSCOs</b>                                   | 2.5%       | 7.7%                  | 29.7%            |

Table 2.2 Comparison of the GDDH13 genome with previously published assemblies of the apple genome.

## Transposable elements and annotation of repeat sequences

*Done by collaborators at URGI*

To produce a genome-wide annotation of repetitive sequences, TE consensus sequences (provided by the

TEdenovo detection pipeline (Flutre et al., 2011) were used to annotate their copies in the whole genome. To refine this annotation, we performed two iterations of the TEannot pipeline. In the GDDH13 genome, TEs represented 372.2 Mb (57.3% of the 649.7 Mb BioNano assembly; Fig. S1).

**Ancestral genome duplication** Intragenomic synteny of GDDH13 was assessed using SynMap (Lyons et al., 2008) (CoGe; <http://www.genomeevolution.org>) and visualized with Circos (Krzywinski et al., 2009). Results of this analysis (Fig. 2.2) showed an even clearer genome duplication pattern than has previously been reported (Velasco et al., 2010). Only very few regions showed no synteny to other parts of the genome (for example, the middle part of Chr04).

### 2.3.2 GDDH13 V1.1

**Problems encountered with GDDH13 V1.0** Following the release of GDDH13 V1.0, a set of 1233 genes involved in biotic stress responses was expertised by colleagues. 97 of these were found to have a structural inconsistency due to a small ponctual variation, compared to the transcripts and proteins homology data used to perform the gene prediction. In order to investigate if these inconsistencies were due to the pseudogene nature of the analyzed genes or to genomic sequence errors leading to an incorrect prediction, illumina DNA-seq reads were mapped on the assembly using BWA-MEM. For a significant part of these genes, a small inconsistency between the genomic sequence and the consensus given by the illumina reads was detected. The example of MD04G0026900 is shown Fig. 2.3a for which a small deletion in the genomic sequence resulted in an erroneus prediction of an additionnal intron at the beginning of the gene in order to bypass the frameshift.

**Genomic sequence improvement with additional polishing** In order to correct these errors, the genomic sequence was improved with more rounds of polishing using Pilon. Since these small sequencing errors could affect the genetic markers mapping used in the pseudo-molecules construction step, the BioNano scaffolds (Fig. 2.1b) were chosen to be subjected to several more Pilon polishing steps (Fig. 2.4). Each Pilon output was subsequently processed again until the number of sequence corrections reach a plateau. Over the 20 supplementary Pilon runs, 53,097 single-base assembly errors; 97,218 insertions (192,156 bp) and 44,063 deletions (120,304 bp) were corrected. Following this step, the corrected scaffolds resulting from this additionnal correction were oriented and assembled in pseudo-molecules by using the SNP markers of the integrated linkage map as described for the GDDH13 V1.0 version.

To assess the improvement made on the genomic sequence, a least stringent mapping of the SNP markers of this integrated linkage map was performed on GDDH13 V1.0 and GDDH13 V1.1. SNP markers were mapped on the genome using BLAT (Kent, 2002). Unique best hits were extracted and filtered with the

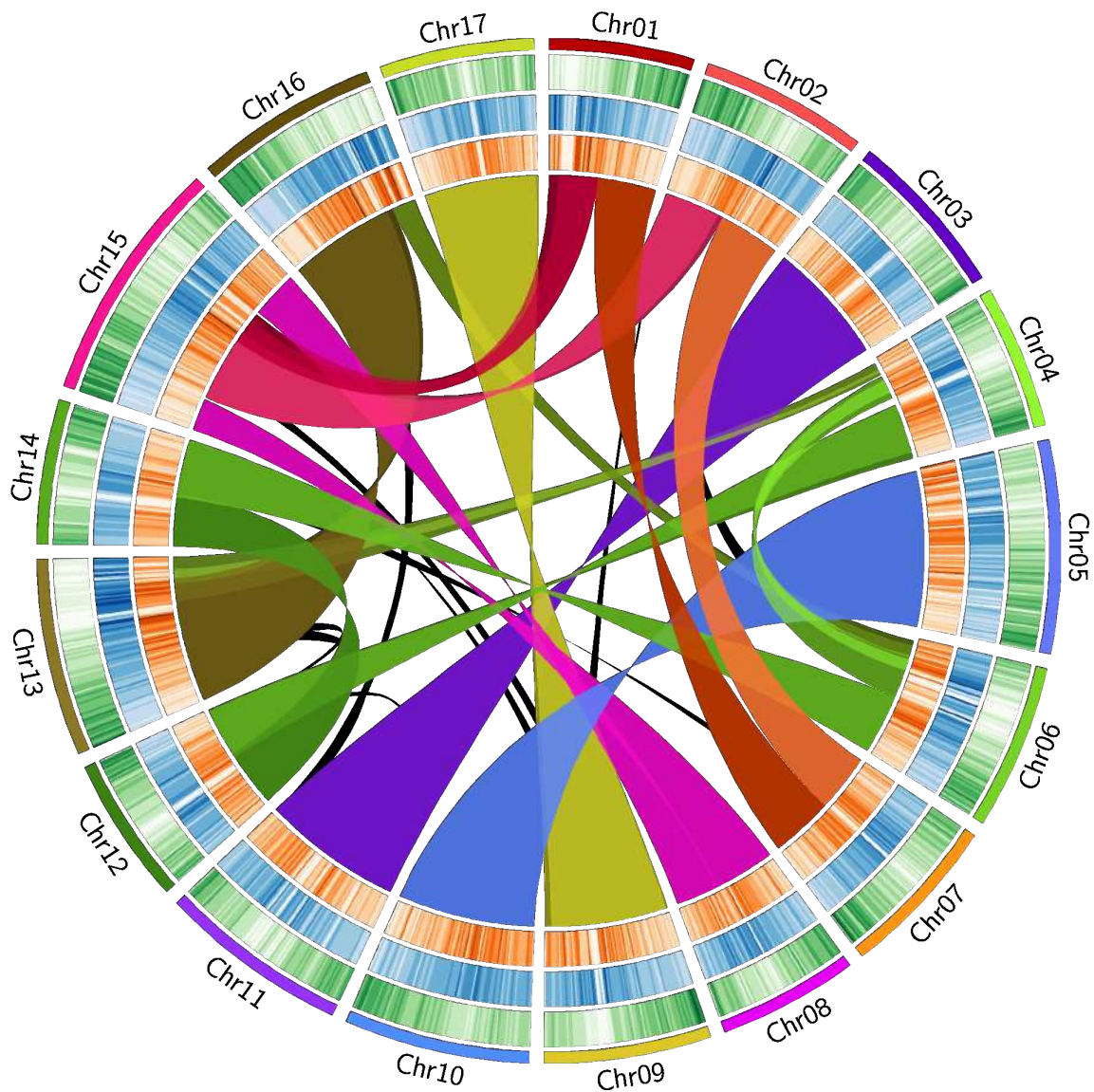


Figure 2.2 Synteny and distribution of genomic and epigenomic features of the apple genome. The rings indicate (from outside to inside, as indicated in the inset) chromosomes (Chr), heat maps representing gene density (green), TE density (blue) and DNA methylation levels (orange). A map connecting homologous regions of the apple genome is shown inside the figure. The colored lines link collinearity blocks that represent syntenic regions that were identified by SynMap.

following parameters : at least 95% match length, at most 4% mismatch length and no gaps. A total of 14,600 markers passed the final filter on GDDH13 V1.1 against 14,583 on GDDH13 V1.0 (Table 2.3).

|                    | Mapped markers | Uniquely mapped markers | Markers passing the filter |
|--------------------|----------------|-------------------------|----------------------------|
| <b>GDDH13 V1.0</b> | 15162          | 14874                   | 14583                      |
| <b>GDDH13 V1.1</b> | 15176          | 14885                   | 14600                      |

Table 2.3 Integrated linkage map SNP markers mapping statistics on GDDH13 V1.0 and GDDH13 V1.1.

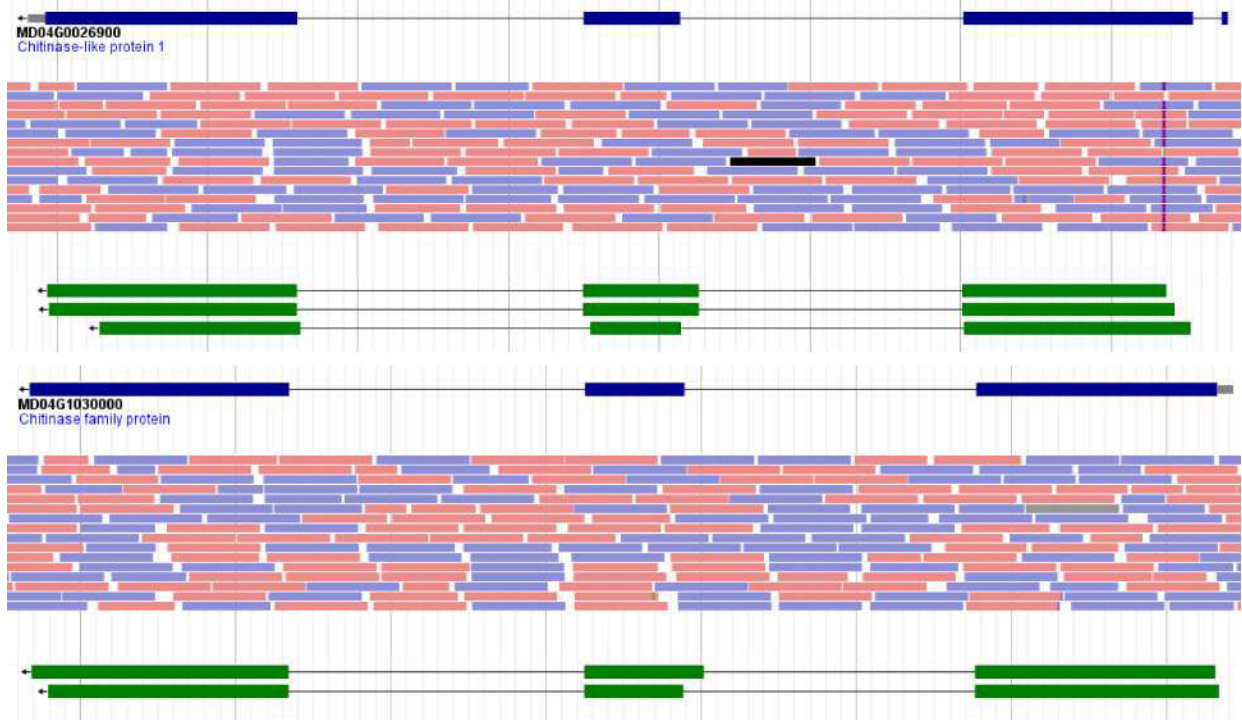


Figure 2.3 Screenshots of a gene affected by genomic sequence errors in GDDH13 V1.0. (a) GDDH13 V1.0 version of the gene (ID: MD04G0026900). The top track represents the gene structure : exons are represented by dark blue squares, introns by black lines and UTRs by grey squares. The middle track represents the mapped illumina reads : each read is represented by a red, blue or black square. A vertical bar in the middle of a read represents a small insertion (one or more nucleotides) in the read compared to the reference sequence. The bottom track represents the results of the BlastX data from Uniprot and Swissprot on the genome at this locus : matching sequences are represented by green squares and gaps by black lines. (b) Same as (a) but with the corresponding locus on the GDDH13 V1.1 version.

**Gene annotation optimization** In addition to the additional polishing, the gene structural and functional annotation was optimized and redone on the GDDH13 V1.1 genomic sequence. Parameter optimization concern various steps of the gene annotation process. Two parameters were used to assess the quality of a gene annotation : the BUSCO score [Simão et al., 2015], and the proportion of predicted genes having the exact same CDS structure compared to a manually expertized set of 865 complete genes ("PREDIRE" score). The GDDH13 V1.1 annotation has a BUSCO score of 96.8% (GDDH13 V1.0 = 94.9%) and a "PREDIRE" score of 78.75% (GDDH13 V1.0 = 73.64%). Complete results are reported table 2.4

The first optimized step was the RNA-seq data assembly with Trinity [Grabherr et al., 2011] in which two parameters were tested : *min\_kmer\_cov* and *jaccard\_clip*. The former, *min\_kmer\_cov*, represents the minimum count of k-mers to be assembled in the Trinity preliminary assembly step Inchworm. Thus, raising the value of this parameter makes the assembly more stringent at the cost of more fragmented contigs. We tried to fix this value at 1 (default) and 2. The gene annotation was substantially better, as showed by both the BUSCO and PREDIRE scores. The second parameter, *jaccard\_clip* can be set to activated or not and

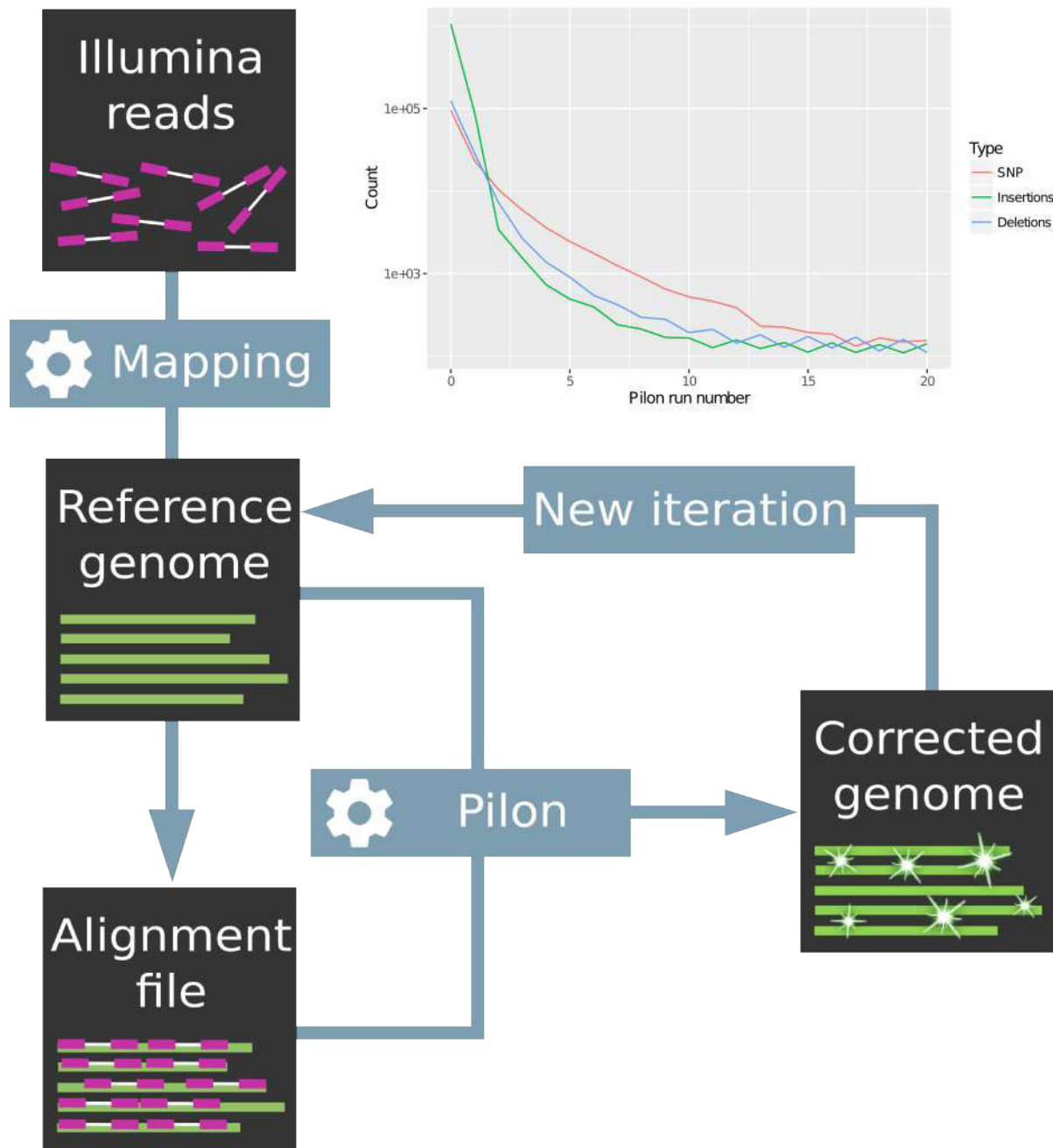


Figure 2.4 Results of the multiple rounds of Pilon polishing. (a) Pipeline used to perform the correction. (b) Number of corrections performed (y-axis, log<sub>10</sub> scale) at each of the 21 rounds of Pilon correction (x-axis). The number of corrected SNP are represented by the orange line, the number of corrected Insertions by the green line, and the number of corrected deletions by the blue line. The run n°0 corresponds to the first correction done on GDDH13 V1.0.

eliminates paired-end reads presenting an inconsistent mapping. Modifying this parameter did not modify the assembly results in an impactful way. Finally, we tried using SOAPdenovo-trans [Xie et al., 2014] to assemble some or all libraries, but this had little effect on the annotation.

The second optimized step is the RNA-seq library selection for the annotation. We tried to provide various

combinations of RNA-seq library to the Eugene combiner. Interestingly, providing more libraries slightly worsened the structural annotation. Thus, we tried to chose non-redundant libraries, resulting from the RNA-sequencing of samples as biologically different as possible, to perform the final annotation. Using a diverse pool of samples for RNA-seq-based helps to have transcripts coming from as many genes as possible, which is useful in order to obtain an exhaustive gene annotation.

Finally, various parameters can be optimized on the combiner step. The first one is the transcript mapping program : we tested GMAP [Wu and Watanabe, 2005] which maps previously assembled RNA-seq contigs on the genome and Cufflinks [Trapnell et al., 2012] which maps the RNA-seq reads directly on the genome and assemble them into contigs afterwards. The annotation quality significantly worsened using Cufflinks. Second, we tested different RNA-seq libraries on which the combiner training step will be performed. This parameter had no effect on the annotation. Third, we tried to change the *preserve\_uniprot* parameter. This parameter is applied during the interpretation of the results of the mapping of Uniprot protein sequences on the genome. Activating this parameter will unmask putative transposable elements (previously masked to avoid false predictions) if an Uniprot protein hit is found. Activating this parameter significantly worsened the specificity of the gene prediction : while the BUSCO metric was good, the number of predicted genes was a lot higher than in other annotation tests due to the numerous falsely predicted genes coming from the transposable element sequences contaminating Uniprot.

## 2.4 Discussion

**Results of the genome assembly and impact of doubled-haploid** As a prerequisite to epigenomic studies in apple, we decided to produce a high-quality reference genome for apple. An advantage for us was the availability of the homozygous GDDH13 doubled-haploid line. Assembling a genome that is both highly heterozygous and recently duplicated into a haploid consensus sequence presents a substantial challenge. This is exemplified by the comparison of our first assembly steps to a recently published report on a heterozygous Golden Delicious apple genome sequence [Li et al., 2016]. Following hybrid assembly of PacBio and Illumina reads, Li and colleagues [Li et al., 2016] reported a N50 of 112 kb, whereas we obtained a N50 of 620 kb at that same step, using the same assembly software. These results highlight the power of haploids or doubled haploids in genome sequencing projects [Zhang et al., 2014], particularly in those for apple, which is not only highly heterozygous but has also undergone a recent whole-genome duplication ([Velasco et al., 2010] and this study). The optical mapping then allowed us to produce scaffolds with a N50 of 5.5 Mb, which, in association with a high-density integrated linkage map, yielded highly contiguous pseudomolecules. In this new apple genome, we followed a newer convention [Di Pierro et al., 2016] in which the orientation of

Chr10 and Chr05 became aligned by the inversion of Chr05. We chose to invert Chr05 because it is the least frequently reported of the two in previous genetic studies on quantitative trait loci (QTL), gene discovery and characterization.

**Genome size** We estimated the genome size of GDDH13 to be 651 Mb, which suggested that the GDDH13 genome may be smaller than that of the heterozygous Golden Delicious line, which was recently estimated to be 701 Mb ([Li et al., 2016](#)). Although the GDDH13 tree looks similar to the heterozygous Golden Delicious counterpart (including tree architecture, flowering time and fruit appearance; **Fig. S1**), it is possible that through the consecutive steps of selfing, haploid development and chromosome doubling, some minor parts of the genome might have been lost or re-arranged. Thus, it is possible that some of the genome sequence might be missing in the GDDH13 assembly.

**Choice of the assembly pipeline** The main challenge in genome sequencing is assembling the repetitive sequences ([Treangen and Salzberg, 2012](#)). To allow this, the reads have to be longer than repetitive occurrences in order to span it completely and assemble it into only one fragment. To do this, we used the PacBio technology which produces long reads of about 10Kb mean length. However the downside of this technology is the high error rate of the reads ([Rhoads and Au, 2015](#)). To handle this in the assembly step, two main choices of pipeline are possible : either a "PacBio-only" assembly in which the sequencing depth will allow the reads to correct themselves, or a hybrid approach using PacBio reads combined with short Illumina reads which have a low error rate. We tested both approaches, using PBcR ([Koren et al., 2012](#)) for a "PacBio-only" assembly and DBG2OLC ([Ye et al., 2016](#)) for the hybrid assembly as described in Methods (**Fig. 2.1**). We observed a large contiguity improvement using the same data between the hybrid assembly (N50 = 620 Kb ; **Table 2.1**) and the "PacBio-only" assembly (best N50 obtained = 119 Kb) which suggests that the coverage of PacBio reads we used (37X) is too low to produce a contiguous assembly without combining it with another technology. However, DBG2OLC is a very sensitive pipeline and produced chimeric contigs which were subsequently corrected by the BioNano scaffolding. Therefore its use may not be suited if there are no means to break the chimeras downstream the assembly pipeline.

A common PacBio genome assembly practice is to use Quiver ([Chin et al., 2013](#)) which uses the PacBio reads after the assembly step to polish it. However we deemed our available coverage to be too low to safely correct sequencing mistakes in the assembly without reintroducing new errors due to the high error-rate of the PacBio reads.

**Predicted Number of genes** Our gene prediction reduced the estimated number of annotated genes in apple from 63,541 (www.rosaceae.org and [Velasco et al., 2010](#)) to 42,140 (V1.0), and 45,116 in the



following GDDH13 V1.1 version. It has been suggested that the number of genes was overestimated in the previous version of the genome because of the assembly and subsequent annotation of both haplotypes [Veeckman et al., 2016]. Another factor that might have contributed to this overestimation is the fragmentation of the original genome which led to the annotation of sections of genes located on different contigs [Denton et al., 2014]. Our new estimation of the number of genes in apple is also more in line with the number of genes reported for other Rosaceae crop species which do not have a duplicated genome, such as peach (27,852 genes, 265 Mb, [Verde et al., 2013]) and diploid strawberry (34,809 genes, 240 Mb, [Shulaev et al., 2011]). In the same way, the analysis of a few large gene families annotated in GDDH13 highlights a reduced family size, and is more in tune with other sequenced plant species (406 cytochromes P450, 49 terpene synthases, 90 pectinesterases, 43 cellulose synthases and 393 PPR proteins).

**Limits of Illumina RNA-seq for gene prediction** We used Illumina RNA-sequencing to perform the gene structural annotation. The inconvenience of this approach is the RNA-seq reads assembly into contigs which have to be done prior to the annotation. We obtained a mean of 140,687 RNA-seq contigs and predicted 45,116 genes in GDDH13 V1.1. This represents 3.12 RNA-seq contig per gene which suggests a high fragmentation of the transcriptome reconstruction. This can lead in structure prediction errors like gene splitting (**Fig. 2.5**), or gene fusion in some cases where reads of UTR overlapping genes are assembled into one chimeric contig. It can also lead to confusion between closely related paralogous genes because of the ambiguity brought by the high sequence similarity during the overlap step of the transcriptome assembly. One possible approach to avoid this is to use long reads as the source of transcripts. We sequenced one library of cDNA reads using the Nanopore technology [Clarke et al., 2009]. We mapped the reads on the genome using minimap2 [Li, 2017] with the default set of parameters for Nanopore reads [-k15 -w5 -Xp0 -m100 -g10000 max-chain-skip 25]. This mapping allows us to observe some cases of mis-annotated genes like MD06G1048600 and MD06G1048700 (**Fig. 2.5**) for which a gene splitting probably occurred. The mapped Nanopore reads at this locus suggest only one gene while the RNA-seq contigs suggest two genes. For this case the predictor, driven by the fragmentation of the RNA-seq assembly, wrongly predicted two distinct genes instead of one. This could be corrected by using the Nanopore reads as a supplementary source of transcript data.

## 2.5 Conclusion

We produced a new reference genome of a 'Golden Delicious' doubled haploid apple tree. We generated a de novo sequence assembly composed of 280 scaffolds with a N50 of 5,558 Kb spread into 17 pseudo-molecules corresponding to the 17 apple chromosomes. We predicted 45,116 protein-coding genes with a BUSCO score



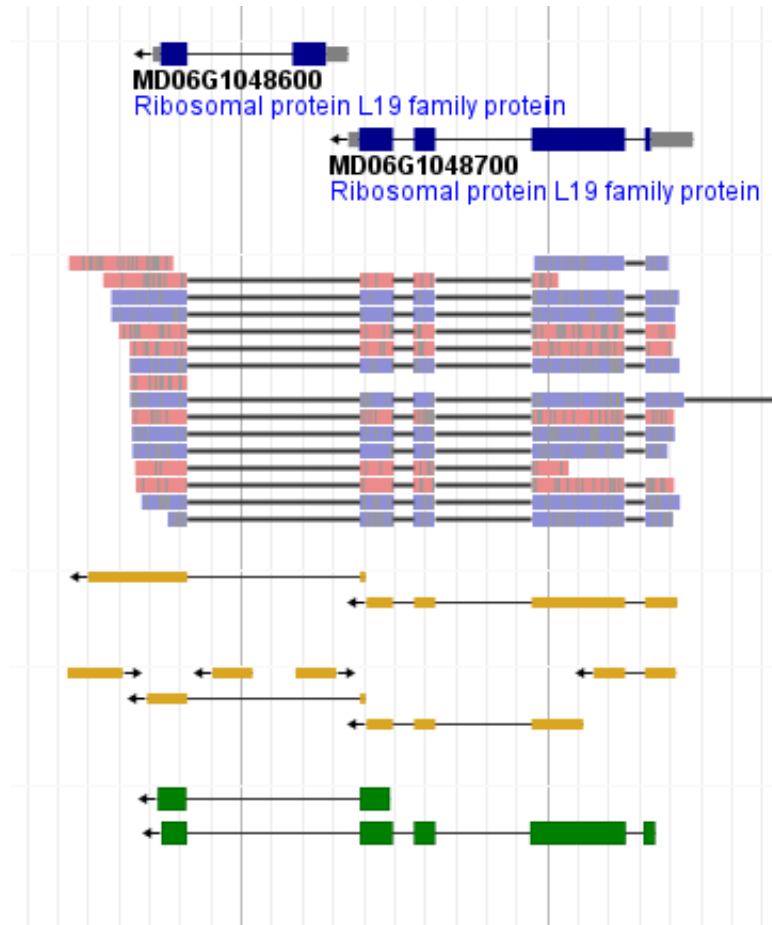


Figure 2.5 Screenshot of a gene prediction error (gene splitting) in GDDH13 V1.1. The top track represents the gene structure : exons are represented by dark blue squares, introns by black lines and UTRs by grey squares. The second track represents the mapped nanopore cDNA reads : red and blue squares represent the matching sequences. The second track represents the mapped nanopore cDNA reads : red and blue squares represent the matching sequences. The third track represents the mapped RNA-seq illumina contigs of the "Gala cv. Flowers" library : yellow squares represent the matching sequences. The third track represents the results of the BlastX data from Uniprot and Swissprot on the genome at this locus : green squares represent the matching sequences. For the three last tracks, the black lines represent gaps.

of 96.8% and annotated functional domains and functions on most of them. We integrated the genome, the gene annotation and various metadata (SNP markers, small RNAs, methylation data...) on a genome browser (<https://iris.angers.inra.fr/gddh13/jbrowse/?data=gddh13>) which we opened to allow the apple scientific community to use this reference genome for their genetic studies.

| <b>RNAseq lib.</b> | <b>Assembleur RNAseq</b> | <b>Trinity option</b> | <b>Mapper</b>  | <b>preserve uniprot</b> | <b>coding genes</b> | <b>BUSCO</b> | <b>PREDIRE</b> |
|--------------------|--------------------------|-----------------------|----------------|-------------------------|---------------------|--------------|----------------|
| base               | Trinity/SOAPdenovo       | kmer cov=1            | Gmap           | no                      | 42140               | 94.9         | 73.64          |
| base               | NA                       | NA                    | STAR+ Cufflink | no                      | 48443               | 92.9         | 73.53          |
| base               | Trinity                  | kmer cov=2            | Gmap           | yes                     | 55400               | 96.4         |                |
| base               | Trinity                  | Jaccard, kmer cov=2   | Gmap           | no                      | 46000               | 95.2         | 73.4           |
| base               | Trinity                  | Jaccard, kmer cov=1   | Gmap           | no                      | 46600               | 94.5         | 72.83          |
| base               | Trinity                  | Jaccard, kmer cov=2   | Gmap           | no                      | 45926               | 95.2         | 73.52          |
| base               | Trinity with genome      | Jaccard, kmer cov=2   | Gmap           | no                      | 46500               | 94.9         | 72.93          |
| base               | Soap                     | NA                    | Gmap           | no                      | 45900               | 95.7         | 73.75          |
| base               | Trinity/SOAPdenovo       | Jaccard, kmer cov=2   | Gmap           | no                      | 45900               | 95.7         | 73.75          |
| base+ GDDH13       | Trinity/SOAPdenovo       | Jaccard, kmer cov=2   | Gmap           | no                      | 45700               | 96.2         | 74.33          |
| base               | Trinity/SOAPdenovo       | Jaccard, kmer cov=2   | Gmap           | no                      | 44800               | 96.5         | 73.2           |
| base+GDDH13        | Soap                     |                       | Gmap           | no                      | 45116               | 96.8         | 78.75          |

Table 2.4 Description of tested parameters in the different steps of the gene structural prediction and corresponding results. Base = libraries used to perform the gene prediction on GDDH13 v1.0. GDDH13 = a RNA-seq library of fruit of GDDH13.

## Chapter 3

# DNA methylation in apple and DMRs

### 3.1 Introduction

In addition to DNA sequence modifications, it has been shown that epigenetic variations contribute to genome accessibility, functionality and structure [Roudier et al., 2009] [He et al., 2011]. Several studies have demonstrated that local DNA methylation variants, which are represented by differential cytosine methylation at particular loci, can have major effects on the transcription of nearby genes and can be inherited over generations [Cubas et al., 1999] [Becker et al., 2011] [Ong-Abdullah et al., 2015]. Apple, like most other fruit tree crops, is propagated by grafting onto rootstocks, which over time can allow the acquisition and propagation of epimutations, via variation in DNA methylation states that can influence various phenotypes, such as fruit color [El-Sharkawy et al., 2015] [Telias et al., 2011]. Thus, knowledge of the epigenetic landscape of apple cultivars could provide new tools to study somatic variants, leading to the development of epigenetic markers for marker-assisted selection. To understand the potential role of epigenetic marks on fruit development, we constructed genome-wide DNA methylation maps that compared different tissues and two isogenic apple lines that produce large or small fruits. This led to the identification of differential DNA methylation patterns that are associated with genes involved in fruit development. Moreover, we showed correlation between DNA methylation and gene expression for each methylation context.

### 3.2 Methods

**Plant material used in the GDDH13 (large fruit) versus GDDH18 (small fruit) diameter analysis**

*Done by Jean-Marc Celton*

The characterization of young fruits development of GDDH13 and GDDH18 was performed from three days

prior hand pollination of flowers to 28 days after pollination (DAP). Central fruit diameters were monitored using a random sample of 8 to 10 fruits representative of all fruits of each of the two biological replicates (clonally propagated trees of the same age, planted next to each-other in an orchard). At each measured date, fruitlets samples derived from both biological replicate were collected and stored appropriately for histological and DNA methylation studies.

### **DNA extraction from leaf and developing fruits and bisulfite sequencing**

*Done by Jean-Marc Celton*

Young leaves and developing fruits 9 days after pollination (DAP) were collected from two biological replicates of a GDDH13 tree and from two biological replicates of a GDDH18 tree. Following liquid nitrogen grinding, DNA was purified from young leaves using the Macherey-Nagel NucleoSpin plant II DNA extraction kit (Germany), following the manufacturer's instructions. Bisulfite treatment was applied to determine the cytosine methylation status using the Epiect bisulfite kit (Qiagen) and 100 ng of genomic DNA. Whole genome bisulfite sequencing was performed to an average of 16.3 fold coverage on the biological samples.

**Mapping of the bisulfite reads on the genome** Bisulfite sequencing reads were mapped on the genome using BSMAP [Xi and Li, 2009] with the following parameters : -q 20 -f 30. DNA methylation distribution plots and gene clustering by methylation patterns were performed with deepTools [Ramírez et al., 2014].

**Identification of DMRs between GDDH13 and GDDH18** DMRs were computed according to [Hagmann et al., 2015].

**Identification of SNPs between GDDH13 and GDDH18** 300 base pairs Illumina reads sequenced from GDDH18 were mapped on the GDDH13 reference genome using BWA-MEM [Li, 2013]. A SNP and small indels identification was performed using freebayes [Garrison and Marth, 2012] with the following parameters: -U 2 -p 1 -\$ 2 -e 1 standard-filters. Variants that had a sequencing depth smaller than 5, higher than 100, a quality score smaller than 50, an alternative allele frequency smaller than 0.8, non-specific to GDDH18, and that were outside annotated CDS were filtered out of the analysis. The results of the analysis are presented in **Table S2**.

**Gene expression quantification** All the gene expression levels are reported in this chapter come from a RNA-seq analysis using GDDH13 and GDDH18 fruit samples, at ten DAP with two replicates for each sample, with a coverage of 120X on genes for each replicate. The RNA-seq reads were mapped on the GDDH13 reference genome using BWA-MEM [Li, 2013] and the expression values were computed using

DESEQ2 [Love et al., 2014]. The expression level is expressed in logCPM, a normalized expression value computed by DESEQ2.

### 3.3 Results

#### 3.3.1 The apple methylome

To investigate the apple methylome, we produced genome-wide maps of DNA methylation content at single-base resolution for GDDH13 leaves and young fruits ([Lister et al., 2008], [Cokus et al., 2008]). Globally, in leaves we found DNA methylation levels of 49%, 39% and 12% in the CG, CHG and CHH sequence contexts (where H is adenine, thymine or cytosine), respectively (Fig. 3.1a). DNA methylation was not evenly spread throughout the chromosomes (Fig. 3.1b shows the profile for Chr11; see Fig. S5 for the profiles for all of the chromosomes), and peaks of methylation coincided with recombination cold spots. As expected ([Matzke and Mosher, 2014], [Law and Jacobsen, 2010]), there are reduced overall DNA methylation levels in gene sequences, whereas TEs are extensively methylated (Fig. 3.1c).

#### 3.3.2 DNA methylation and fruit development

To assess how DNA methylation contributes to fruit development, we first compared DNA methylation levels between leaves and fruits. We called differentially methylated regions (DMRs) using a hidden Markov model (HMM)-based approach [Hagmann et al., 2015]. In total, we identified 1,017 DMRs in all contexts between leaves and fruits, and we observed a very strong bias for DMRs containing methylation changes in the CHH context (875 DMRs; 86.0%) (Fig. 3.2a). We identified 294 genes that contained DMRs in their promoter region 14 DMRs were in the CHG context and showed increased amounts of DNA methylation in leaves, whereas the remaining 280 DMRs were found in the CHH context and showed increased amounts of DNA methylation in fruits. Thus, most methylation differences between leaves and fruits occurred at CHH sites, with a robust increase observed in the developing fruit. Among genes with DMRs that were 2 kb upstream of their transcription start site (TSS), we identified several apple orthologs of Arabidopsis genes with important roles in flower and fruit development and in epigenetic regulation (Fig. 3.2b).

Next we wanted to test whether DNA methylation could have a role in the regulation of fruit size. We took advantage of GDDH18, an isogenic line that was obtained from the same haploid that produced GDDH13. Whole-genome sequencing showed the presence of 27 homozygous SNPs within genes between the two trees, with nine of these SNPs resulting in amino acid changes (Table S2). Although the GDDH13 and GDDH18 trees were indistinguishable, the GDDH18 fruits were much smaller (Fig. 3.2c) because of a reduced number of cell layers in the parenchyma (Fig. 3.2d).

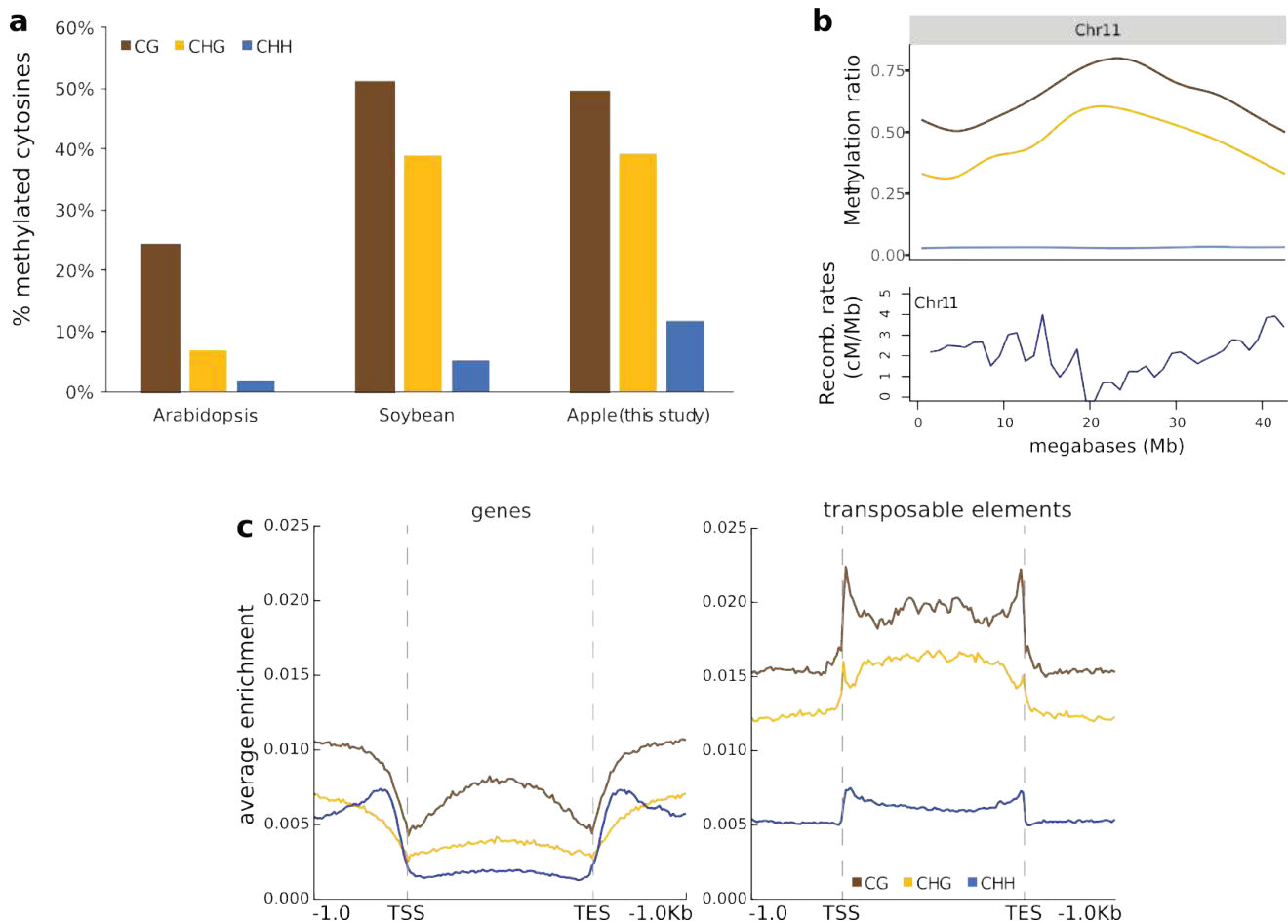


Figure 3.1 DNA methylation landscape of the GDDH13 genome. (a) Percentage of DNA methylation distributions of the three methylation contexts (CG, CHG or CHH) in Arabidopsis [Cokus et al., 2008], soybean [Schmitz et al., 2013] and apple. For apple, the percentages were estimated based on the number of cytosines that had a methylation ratio  $\geq 0.75$ . (b) Top, chromosomal distribution of the methylation ratios along Chr11. Bottom, the recombination rate plot from Figure 2.1d, for comparison purposes. (c) Global distribution of DNA methylation levels at protein-coding genes and TEs, including a 1-kb window upstream of the TSS and downstream of the transcription end site (TES). In all of the panels, the DNA methylation sequence contexts are color-coded as follows: brown for CG, yellow for CHG and blue for CHH.

To elucidate whether the difference in fruit size could have an epigenetic basis, whole-genome bisulfite sequencing was performed on samples that were collected at 3 days before pollination (or 3 days after pollination (DAP); when fruits have a similar size and number of cell layers) and at 9 DAP (a few days before observing significant phenotypic differences between the fruits). As expected from their common origin, only a limited number of high-confidence DMRs ( $n = 197$ ) could be found between young fruits of GDDH13 and GDDH18 at 3 DAP. Of these, 47 DMRs were located within 2 kb upstream of the TSS of genes. Similarly, we identified a total of 148 high-confidence DMRs between fruits of GDDH13 and GDDH18 at 9 DAP. From this analysis, we found that 53 genes contained DMRs in their promoter region (i.e., within 2 kb upstream of the TSS). At both time points a majority of genes with DMRs showed a decrease in methylation in their

promoter region for GDDH18 (Table S3). Notably, in both comparisons, DMRs in the CG CHG and CHG contexts were over-represented.

The overlap of DMRs between the two time points analyzed included 22 genes with DMRs in their promoter regions, with most of them ( $n = 17$ ) showing lower methylation in GDDH18 (Table S3). Several of the 22 genes have orthologs in other species with a role that could explain the observed size difference between the GDDH13 and GDDH18 fruits including SQUAMOSA PROMOTER-BINDING PROTEIN LIKE 13 (SPL13, MD16G0108400), 1-AMINO-CYCLOPROPANE-1-CARBOXYLATE SYNTHASE 8 (ACS8, MD15G0127800) and CYTOCHROME P450 FAMILY 71 SUBFAMILY A POLYPEPTIDE 25 (CYP71A25, MD14G0147300), which belong to the minority of genes with increased methylation in GDDH18.

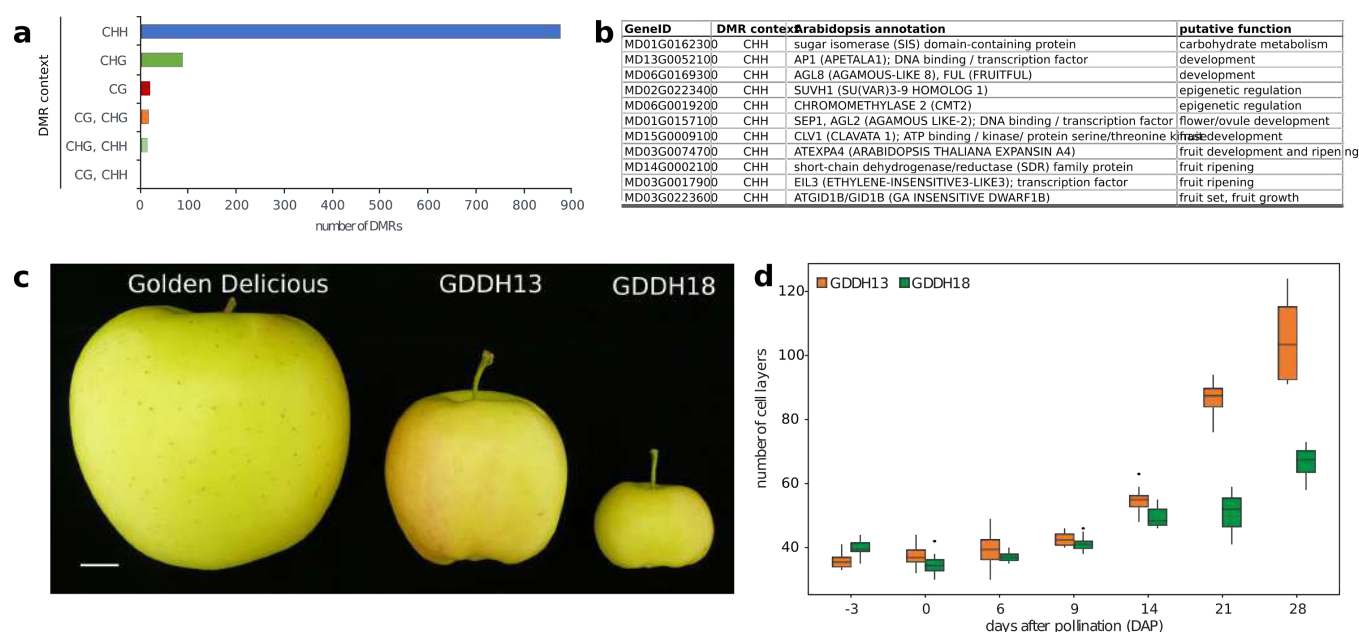


Figure 3.2 Differentially methylated regions between apple tree leaves and young fruits. (a) DMR content in samples of GDDH13 leaves and young fruits (CHH,  $n = 875$  DMRs; CHG,  $n = 88$  DMRs; CG,  $n = 21$  DMRs; CG and CHG,  $n = 17$  DMRs; CHG and CHH,  $n = 14$  DMRs; CG and CHH,  $n = 2$  DMRs). Most of the DMRs (86%) were identified in the CHH context. (b) Selection of GDDH13 genes that present a DMR within a region 2 kb upstream of the TSS. The apple gene ID, the methylation context of the DMR, the orthologous Arabidopsis gene annotation and the function of the encoded protein are listed. (c,d) Representative image comparing the fruit sizes of heterozygous Golden Delicious, GDDH13 and GDDH18 at harvest (c) and quantification of the number of cell layers in the parenchyma of GDDH13 (orange) and GDDH18 (green) fruits, as assessed by microscopy ( $n = 12$  data points per box plot) (d). The horizontal line in the box represents the median, the lower and upper hinges correspond to the first and third quartiles, the lower and upper whiskers extend from the hinge to the smallest and largest value (no further than 1.5-fold the inter-quartile range from the hinge), and outlying points are plotted individually. Scale bar, 1 cm.

### 3.3.3 Correlations between methylation and expression

#### Distribution of methylation level in gene putative promoters by methylation context

For each gene and each methylation context (CG, CHG, CHH), the mean methylation level was computed in the putative promoter region. Several regions upstream of the genes were considered (Fig. 3.3a). Each gene was classified in one of two classes based on the mean methylation ratios in an upstream region : mean comprised between 0 and 0.5 and mean comprised between 0.5 and 1. Several upstream regions ranges were tested, going from one hundred nucleotides before the gene to the TSS, to four thousands nucleotides before the gene to the TSS.

For the CG and CHG contexts, we observe that the bigger the upstream region is, the more genes are methylated in this region (average methylation ratio  $> 0.5$  in the tested region) ; for example, for the CG context in the [-100 :0] region, 3228 genes have more than 0.5 mean methylation ratio while 10404 genes have more than 0.5 mean methylation ratio in the [-4000 :0] region. The same observations can be done for the CHG context.

Concerning the CHH context, more genes have methylation in their upstream regions when the upstream region starts close to the TSS compared to when the upstream region starts far from the TSS : 448 genes for the [-100 :0] region and 5 and 2 genes for the [-400 :0] and [-4000 :0] regions respectively. This is the inverse observation than for the CG and CHG contexts and it suggests that the CHH methylation is very low in all the genome but higher near the genes TSS.

Moreover, it is noticeable that the shift in number of observations of high methylation between closer and more distant upstream regions is higher for the CHH context ( $448/5 = 89.6$  ratio between [-100 :0] and [-1000 :0]) than for the CG context (same ratio of 3.2) and the CHG context (same ratio of 3.1).

#### Methylation patterns

In order to explore the global effect of methylation on gene expression, we studied the effect of methylation patterns around the TSS. BPRmeth [Kapourani, 2016] was used to compute methylation patterns on a fruit sample of GDDH13, 9 DAP. Each methylation context (CG, CHG and CHH) was processed separately. Two regions around the TSS were tested : [-500 :500] and [-100 :100]. BPRmeth needs a predetermined number of clusters to classify each patterns in. Two approaches were tested : using three clusters or using five pre-determined clusters.

Concerning the CG methylation context (Fig. 3.4a), we can observe three distinct cluster : the orange cluster which shows high methylation on all the tested region, the green cluster which shows low methylation



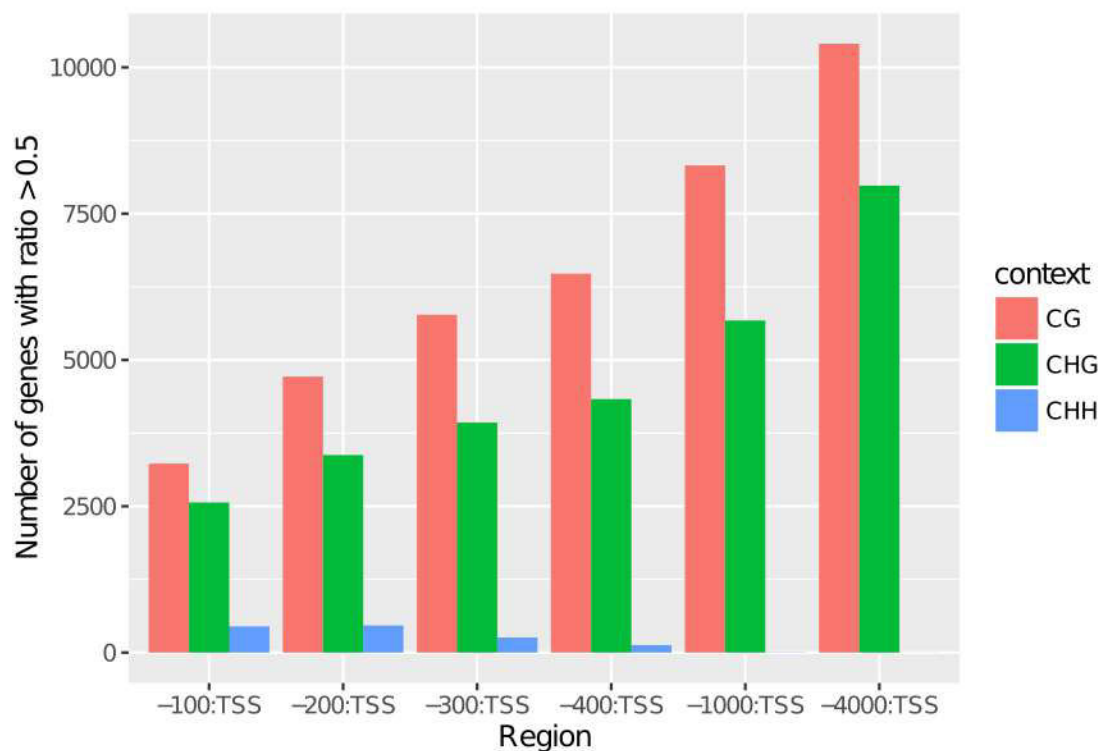
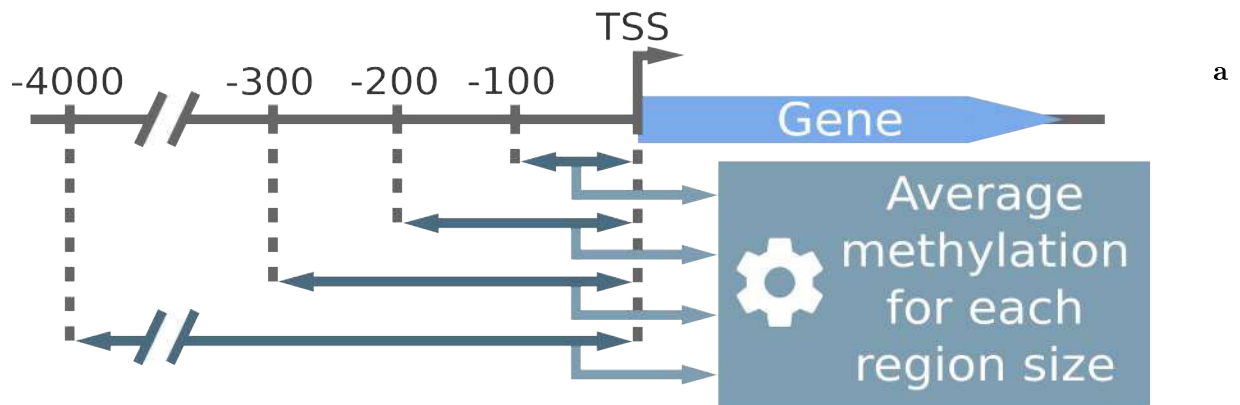


Figure 3.3 Results of average methylation analysis in gene promoters. (a) Protocol followed to compute average methylation in different regions upstream each genes. (b) Number of genes having an average methylation ratio  $> 0.5$  (y-axis) for each analyzed region (x-axis) and each methylation context. The DNA methylation sequence contexts are color-coded as follows: red for CG, green for CHG and blue for CHH.

on the tested region, and a blue cluster that show high methylation in the  $[-100 : 0]$  region and low methylation in the beginning of the gene body ( $[0 : 100]$  region). The genes in the blue cluster show the highest expression (11.5 logCPM median), the genes in the orange cluster show the lowest expression (6 logCPM median) and the genes in the green cluster show an intermediate expression (9 logCPM median). The same observations can be done for the CHG context (Fig. 3.4b).

Concerning the CHH methylation context, there are two similar clusters (in blue and green) which are

methylated upstream the gene and almost completely unmethylated in the gene body. The last cluster (in orange) shows an overall low methylation level, which is lower in the [-100 :0] region than the two other clusters, but higher in the [0 :100] region. This last cluster shows a lower median expression level (6 logCPM) than the two others (12 logCPM). This would suggest that a higher CHH methylation in the promoter would promote a higher expression of the gene. These results also show that the CHH methylation level is always low in the gene body.

## 3.4 Discussion

### 3.4.1 The apple methylome

The genome-wide distribution of DNA methylation peaked in putative centromeric regions of high Linkage Disequilibrium (LD) (Fig. 2.2). As has been observed in Arabidopsis [Lister et al., 2008], TEs were enriched and genes strongly depleted for DNA methylation. Compared to other plant methylomes, apple tends to be highly methylated (Fig. 3.1) but the relative methylation proportion between cytosines contexts remain similar.

### 3.4.2 DNA methylation and fruit development

The comparison of the apple leaf and fruit methylomes revealed a noteworthy pattern – the fruit globally had higher CHH DNA methylation levels, which suggested increased activity of the RNA-directed DNA methylation machinery in this organ [Matzke et al., 2015]. Consistent with this observation, it has been shown for Arabidopsis that cell-type-specific DNA methylation differences mainly occur at CHH sites [Kawakatsu et al., 2016]. Notably, DNA methylation differences in the CHH context between leaf and fruit tissues occurred next to 294 genes. Several of these were found to be orthologous to genes that are known to be important regulators of flower and fruit development in other species. This suggests that apple fruit development is regulated by epigenetic processes, which is consistent with data obtained in tomato, demonstrating that DNA methylation is important for fruit ripening ([Manning et al., 2006], [Liu et al., 2015], [Gallusci et al., 2016]).

In addition, among the major agronomical traits that contribute to both yield and quality, fruit size is one of the most important for many domesticated crops. Two of the key determinants that are known to alter plant organ size are cell number and cell size [Guo and Simmons, 2011]. Here we investigated fruit size difference between two isogenic doubled-haploid apple lines. We found that the number of cell layers in the parenchyma of GDDH13 fruits increased more rapidly than those in the parenchyma of the smaller GDDH18 fruits, with significant differences being observed as early as 21 DAP. To identify regulators that contributed to the

difference in fruit size between the two doubled-haploid apple lines, we found three genes that potentially contributed to the cell number difference, and these contained DMRs in their promoter regions.

The identification of potential molecular mechanisms that control cell-division-related processes by DNA methylation provides new insights into the understanding of this important process. However, by comparing the GDDH13 and GDDH18 genomes, we identified nine SNPs that affect protein sequences, and thus we cannot currently exclude a genetic effect.

### 3.4.3 Correlations between methylation and expression

#### DNA methylation levels upstream of genes

We found that the CG and CHG methylation is lower near the TSS of genes while the CHH methylation increases close to the TSS. This result implies that gene regulation by CG/CHH methylation occurs by demethylation of the promoter and regulation by CHH methylation occurs by active methylation of the promoter. We also found that the methylation levels correlation with distance to the TSS is greater for the CHH context. This could suggest that the methylation is more actively regulated in the CHH context than in the two other contexts near the genes.

#### Methylation patterns

The fact that the expression is higher when the CG methylation is lower upstream of the gene is already known, however it is surprising that a higher CG methylation in the upstream region of the TSS (blue cluster) correlates with even higher expression. One hypothesis is that the CG methylation peak in the [-100 :0] region correlates with a very low methylation in the gene body. It has already been demonstrated that gene body methylation lessens the gene expression [Aceituno et al., 2008]. This demonstrates that only considering the region upstream the TSS may be an inaccurate way to predict expression with CG methylation level.

We showed that a higher CHH methylation upstream of genes correlates with a higher gene expression.

## 3.5 Conclusion

Using the apple reference genome previously described, we generated whole-genome methylome maps for apple leaves and fruits, including the two phenotypes of doubled-haploid fruits GDDH13 and GDDH18 which present a different fruit size despite having the same genome. Then, we confirmed that a higher CG and CHG methylation upstream of the genes could inhibit gene expression, and a higher CHH methylation in this region could raise gene expression. This antagonist impact of the cytosine context hasn't been clearly identified before and shows that it is mandatory to treat each context separately when doing comparative

methylation studies. We also identified new correlations between the shape of methylation patterns upstream of genes and gene expression. This shows that the methylation pattern shape are important to predict gene expression, and that this information is complementary to the mean methylation level. Finally, we identified several DMRs along with candidate genes that could potentially affect fruit development by comparing the methylomes of GDDH13 and GDDH18. These genes's expression could be affected by hypo- or hyper-methylation in either of these two organisms resulting in the developpment of the two different phenotypes. In order to test this hypothesis, one possible approach is to perform knock-outs on each of these genes and observe any phenotype modification. If fruit size is affected after a knock-out experiment, targeted demethylation around the knocked-out gene could be performed to confirm that the expression changes of the gene is provoked by DNA methylation.

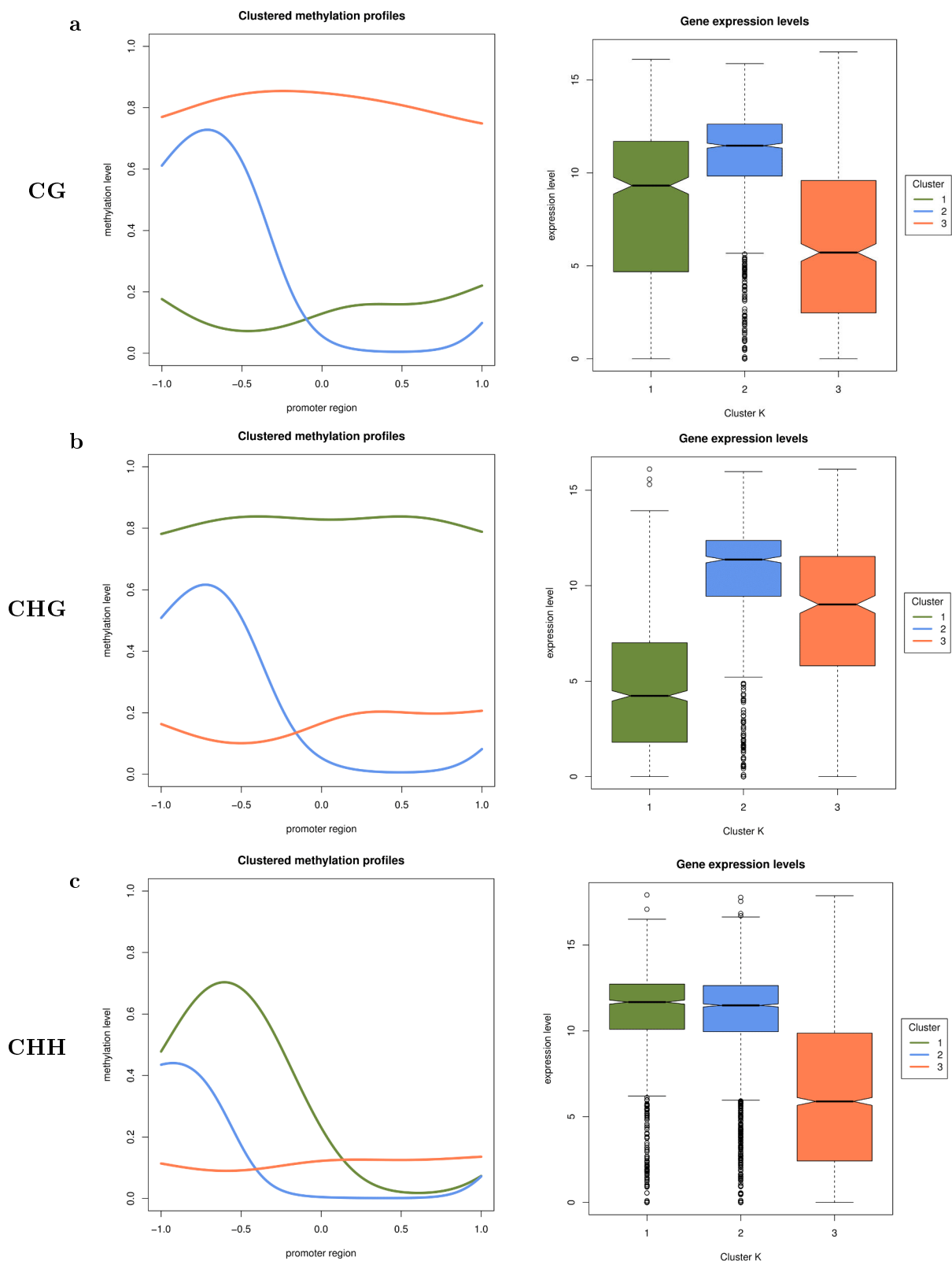


Figure 3.4 Results of the methylation patterns analysis, for the  $[-100 : 100]$  regions with three clusters of patterns. **(a (left))** Line plot of the methylation patterns for the CG methylation context. The y-axis represents the methylation ratio and the x-axis represent the position relatively to the TSS. On the x-axis, -1.0 represents the nucleotide which is one hundred bases before the TSS, 0.0 represents the TSS position and 1.0 represents the nucleotide which is one hundred bases after the TSS. Each line corresponds to a different cluster. **(a (right))** Boxplot representing the expression values of the genes of each clusters represented in (a (left)). The boxplot colors are corresponding to the line colors in (a (left)). **(b)** Line plot and boxplot expression values for the CHG methylation context. **(c)** Line plot and boxplot expression values for the CHH methylation context.

## Chapter 4

# Differentially Methylated Regions detection pipeline

### 4.1 Introduction

The most straightforward way to compare two methylomes is to find the Differentially Methylated Regions (DMRs). DMRs are genomic features on which the methylation level between two samples are significantly different. Many DMRs finding tools exist but a lot of specificity issues subsist, especially with a low number of biological replicates, which is a very common experimental design due to the cost of whole-genome bisulfite sequencing. Moreover, these tools are most of the time difficult to use which can hinder epigenetic studies if no bioinformatic resources are available. We propose an easy-to-use and complete pipeline which aims at automating the entire process, from read mapping to Differentially Methylated Regions (DMRs) detection taking into account biological replicates. The pipeline developed here integrates a new tool used to compute DMRs and provides clear metrics which the user can easily interpret to find high-quality DMRs. We tested this pipeline on various apple, maize and arabidopsis methylomes, and report the preliminary results.

### 4.2 Methods

#### 4.2.1 Programming language

Python 2.7.9 was used to write the pipeline.

## 4.2.2 Pipeline global description

We developed a pipeline which processes Whole Genome Wide Bisulfite Sequencing (WGBS) data from the raw reads to filtered Differentially Methylated Regions (DMRs) and global statistics and comparisons (Fig. 4.1). The reads are mapped onto the genome and the methylation values for each cytosine on the genome are extracted in methylation files using BSMAP [Xi and Li, 2009]. The methylation files are used for two different analysis. First, genome-wide statistics and comparisons are computed. Second, DMRs are identified on the genome then filtered to obtain the final DMRs.

## 4.2.3 Bisulfite reads mapping

The data files corresponding to different sequencing lanes are first concatenated in order to obtain two files per sample : one for the forward sequencing and one for the reverse sequencing. Bisulfite reads mapping and methylation extraction are performed using BSMAP. The following parameters are set by default : -q 20 -f 30, but the user is free to modify them. Each sample is processed separately. The output of this step is one methylation calling file (BSMAP format) for each sample.

## 4.2.4 Global individual statistics and comparisons

Genome-wide metrics are computed using the methylation calling files obtained at the end of the mapping step, for each sample separately. Mean CG, CHG, and CHH methylation are computed. Chromosome-scale plots of methylation density are generated using ggplot2 [Wickham, 2016]. Methylation patterns plot near genes are generated using deeptools [Ramírez et al., 2014] if a gene annotation is available.

## 4.2.5 DMRs computing

To identify Differentially Methylated Regions (DMRs) between two conditions, the genome is divided in small overlapping regions (Fig. 4.2). The choice of the regions size and overlap is left to the user but the default parameters are the following : window\_size = 200 nt, overlap = 100 nt. These default parameters were determined by empiric observations of DMRs sizes found with other tools on apple.

The biological replicates are then fused as described in Fig. 4.3 in order to obtain one methylation rate per cytosine per condition.

A pairwise Wilcoxon test is done between the two conditions in each defined genome window separately, using the following hypothesis (Eq. (4.1)), with  $D$  = all found DMRs sizes.

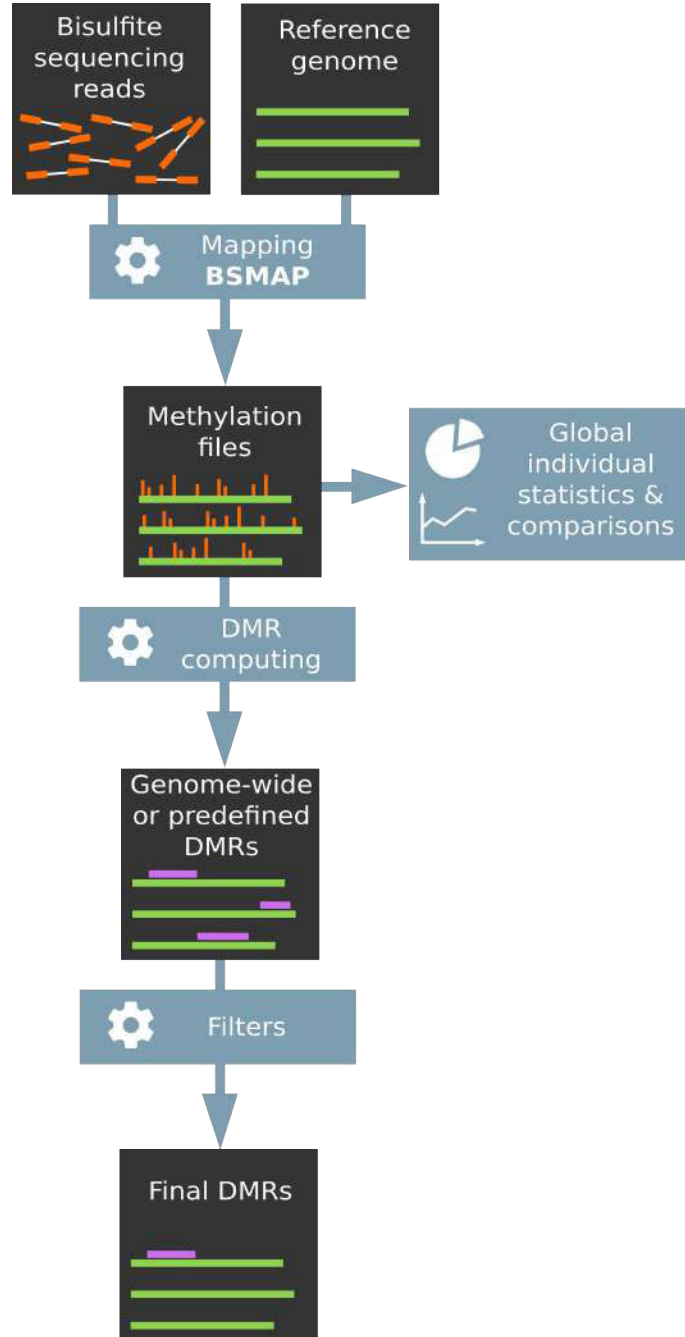


Figure 4.1 Global schema of the bisulfite sequencing analysis pipeline.

$$\forall i \in D : \begin{cases} H_{i0}: \text{There is no mean methylation difference between the two conditions} \\ H_{i1}: \text{There is a mean methylation difference between the two conditions} \end{cases} \quad (4.1)$$

For each tested region, aside from the p-value of the Wilcoxon test, the following values are computed : standard deviation of methylation averages between biological replicates, average of individual cytosines



standard deviations between biological replicates, average bisulfite reads coverage in the region, standard deviation of bisulfite reads coverage between biological replicates, and average methylation difference in the region between the two conditions. Finally, a multiple testing correction via False Discovery Rate (FDR) estimation is performed inside each set of hypothesis and all DMRs having a p-value  $\geq 0.05$  are filtered out of the analysis.

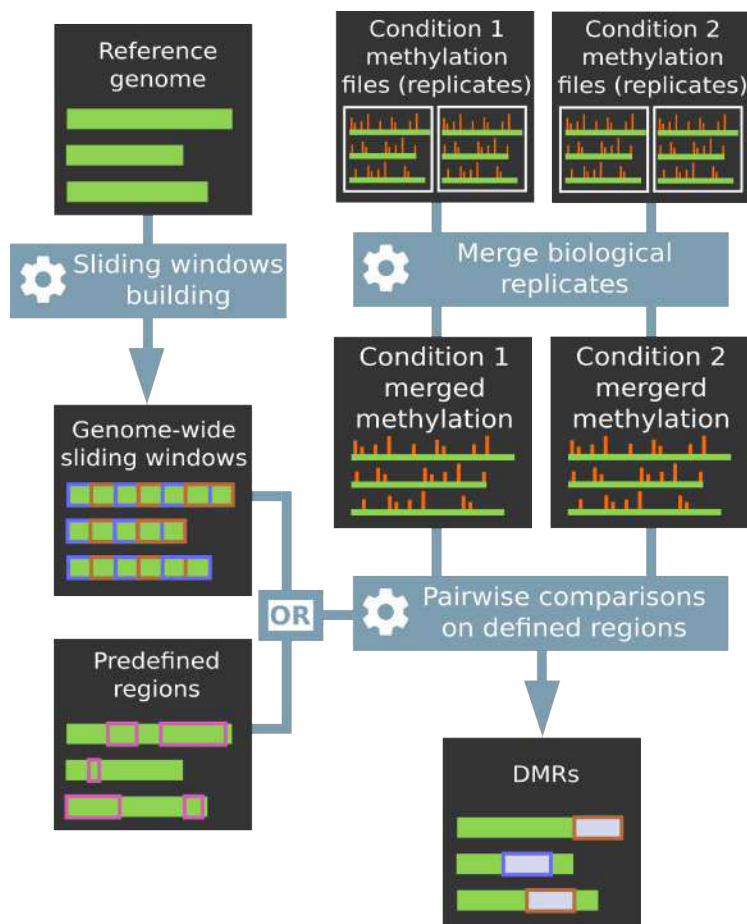


Figure 4.2 Schema of the DMRs identification process. First, windows on which DMRs will be tested are set to either user choice (predetermined windows) or computed on a overlapping windows model. Second, biological replicates are collapsed for each cytosine, in order to have only one value per cytosine per condition. Finally, DMRs are computed on the defined windows using the collapsed methylation values.

#### 4.2.6 DMRs filtering

Obtained DMRs are then filtered using the following criterias : a minimum threshold of average methylation difference (default 0.3 for the CG and CHG contexts and 0.1 for the CHH context, arbitrarily visually determined) and a maximum threshold of standard deviation of methylation averages between biological replicates.

|                      |  |                             |                          |
|----------------------|--|-----------------------------|--------------------------|
|                      | <b>—C<sub>1</sub>—C<sub>2</sub>—C<sub>3</sub>—</b> |                             |                          |
| <b>Replicate 1</b>   | <b>4/8</b>   | <b>1/1</b>                  | <b>7/9</b>               |
| <b>Replicate 2</b>   | <b>2/6</b>   | <b>2/10</b>                 | <b>NA</b>                |
| <b>Merged data</b>   | <b>6/14</b>  | <b>X</b>                    | <b>X</b>                 |
| <b>Justification</b> | <b>Ok</b>  | <b>Rep. 1<br/>cov &lt;4</b> | <b>Rep.2<br/>not cov</b> |

Figure 4.3 Description of the biological replicates merging process with three examples. Green numbers correspond to the number of "C" (methylated) bisulfite reads and red numbers correspond to the number of "T" (unmethylated) bisulfite reads coverage at this cytosine. **C<sub>1</sub>** : normal case. The number of "C" reads of each replicate are added to obtain the number of "C" reads of the merged data. Same calculation for the number of "T" reads. **C<sub>2</sub>** : replicate 1 has one "C" reads and one "T" reads which add to a total coverage of 2. This coverage is strictly lower than the threshold of 4 so this cytosine will not be taken into account in the merged data for this condition. **C<sub>3</sub>** : same as C<sub>2</sub> but because there is no coverage at all on the replicate 2.

### 4.3 Results

The pipeline was used to find DMRs during various side projects.

#### 4.3.1 Determining differential methylation on target regions

In this projects' ([Thieme et al., 2017](#)) context, the pipeline was used to evaluate the differential methylation between predefined regions, corresponding to transposable elements coordinates, between several *A. thaliana* methylomes that underwent various demethylation ( $\alpha$ -amanitine, zebularine) treatments and a control. Several transposable elements, like AT1TE12295 were found to be differentially methylated, especially when two drugs were combined (Fig. [4.4](#)).

#### 4.3.2 DMRs between tissues in apple

The aim of the experiment is to assess how DNA methylation patterns are inherited via sexual and asexual reproduction. We compared the methylomes between three different apple organs (grafts, seeds and trees) in order to find differentially methylated regions to associate with differentially expressed genes. We performed pairwise comparison of each condition (grafts against seeds, graft against trees and seeds against trees) using two biological replicates for grafts and trees samples and four replicates for seeds samples. We used a sliding window size of 200, overlapping by 100 nucleotides and performed the comparisons for each

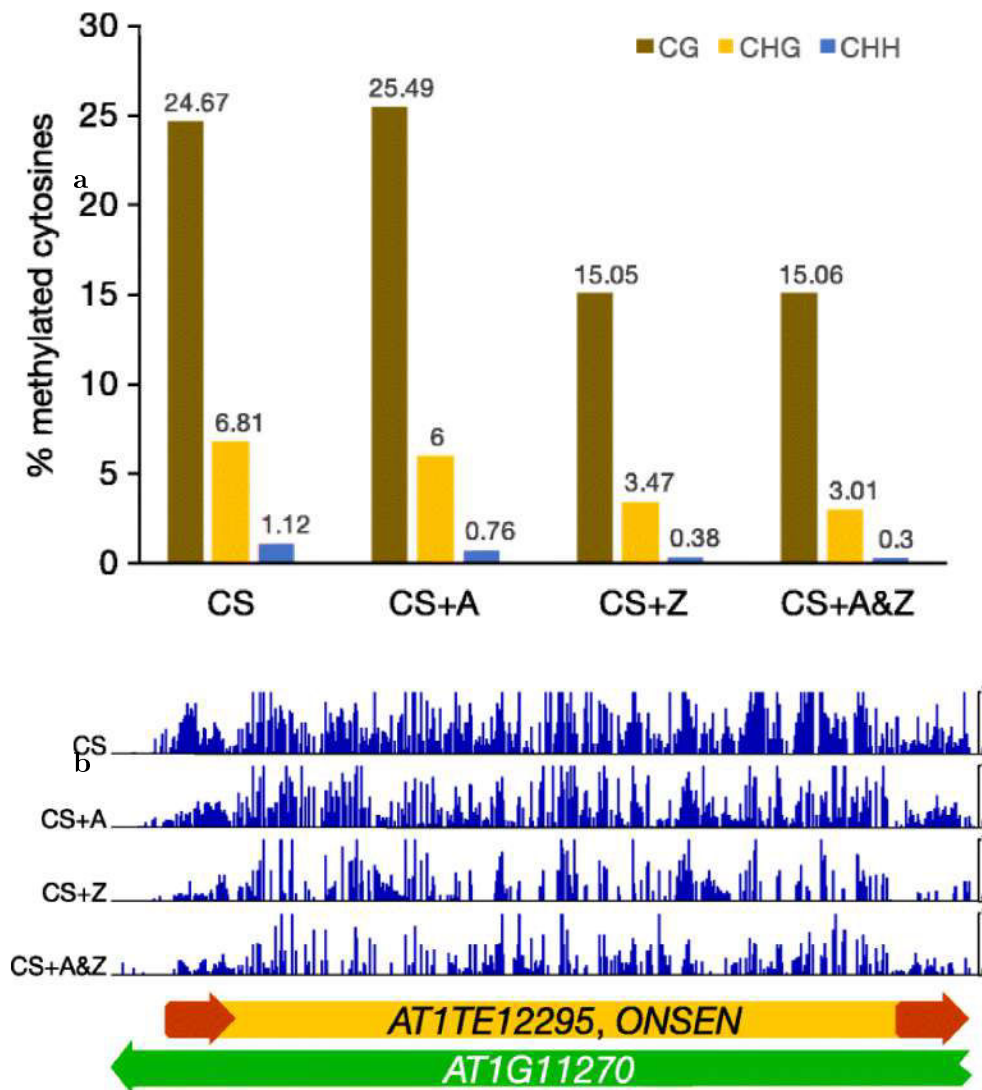


Figure 4.4 Methylation levels modifications observations in Arabidopsis with  $\alpha$ -amanitine and zebularine treatments. (a) Genome-wide DNA methylation levels in the WT after CS and CS plus A (5  $\mu\text{g}/\text{ml}$ ), Z (40  $\mu\text{M}$ ), or a combination of A and Z (A&Z) for three sequence contexts (brown for CG, yellow for CHG and blue for CHH). (b) Methylation data of treated and untreated plants at an ONSEN locus located on Chr 1 (ONSEN is indicated in yellow, its LTRs in red). Figure extracted from [Thieme et al., 2017].

methylation context (CG, CHG, CHH) separately. We then filtered DMRs based on the value of the standard deviation of average (see 4.2.6). The results are reported table 4.1.

| context | GvS    | TvG    | TvS    |
|---------|--------|--------|--------|
| CHH     | 210865 | 148895 | 227442 |
| CHG     | 624    | 2062   | 883    |
| CG      | 856    | 4538   | 1666   |

Table 4.1 Number of DMRs found for each comparison between different apple tissues. GvS = Grafts vs Seeds, TvG = Trees vs Grafts, TvS = Trees vs Seeds

## 4.4 Discussion

### 4.4.1 Biological validation of DMRs

Once potentially interesting DMRs have been identified, they should be biologically validated by independent experiments. One possible approach is to digest DNA with restriction enzymes like HpaII which digest unmethylated cytosines on its restriction site. This will result in no PCR product if the corresponding cytosines are unmethylated. However this validation method only works for DMRs that are completely hyper-methylated in one condition and completely hypo-methylated in the other.

Another validation method that bypass this limit is to perform localized, independent bisulfite sequencing. It allows to produce longer (500 bp) sequences which will ensure a more accurate alignment and assessment of methylation level at single-base level.

### 4.4.2 Working with few biological replicates

DNA methylation levels have a high variability between different individuals, tissues and cells [Alonso et al., 2016]. During DMRs identification, this can lead to specificity problems when the number of biological replicates for one condition is low [Ziller et al., 2015]. Some DMRs identification methods [Feng et al., 2014] compare each cytosine independently but require a lot of biological replicates to achieve a good specificity and power. In order to deal with lower number of replicates (three or less), we propose an alternative approach which consist in comparing entire regions at once instead of each cytosine independantely. This greatly increases the number of values used in the statistical tests, compensating for the low number of replicates specificity-wise, at the cost of losing the biological variability information. To balance this, we added post-treatment filters which ensure that biological replicates stay consistent at the region level within identified DMRs. This uses the same principle than other methods like BSmooth [Hansen et al., 2012], which works on entire regions at once but smooth the methylation profiles in order to be able to work with lower coverages. Not performing a smoothing allows to detect very small DMRs at the cost of losing accuracy on low coverage loci.

### 4.4.3 Selection of the regions to compare

We chose to let the user be free to submit custom predefined regions in which the methylation level will be compared or to analyze the whole genome using sliding windows. The former approach allows to precisely target regions of interest which raise the power and specificity on these regions, but make impossible to discover novel DMR somewhere else on the genome. The sliding windows approach allows this but the boundaries of a window have a high chance to not exactly represent the boundaries of a DMR. Thus, if the overlap between a window and the real DMR is only 70%, the analysis will lose sensitivity. Another

inconvenient of the sliding windows method is the need to fix a window size. Smaller and larger window sizes allow to gain sensitivity on the detection of smaller and larger DMRs respectively. It may be appropriate to run the analysis multiple times with different window sizes to optimize the sensitivity.

#### 4.4.4 DMRs filters

**Filtering on average methylation difference** At the end of the DMRs analysis, we perform a filter on the average methylation difference between the two conditions in order to select robust DMRs. While this allows to easily obtain highly varying DMRs, these filters are to be set with caution depending on the methylation context. Notably, we showed that the CHH context is globally less methylated than the CG and CHG contexts on the genome (Fig. 3.1). Thus we set a lower default threshold for CHH than for the two other contexts based on our experiences on apple. However, these threshold might need to be changed depending on the studied organism, depending on the relative global methylation levels in the three methylation contexts and the desired sensitivity and specificity.

**Separating methylation contexts in the analysis** We chose to treat the three methylation contexts separately for the analysis. We showed that the CHH context had different patterns than the the other contexts (Fig. 3.4), however the CG and CHG contexts have more common points, hence it would be interesting to be able to treat them alike in the pipeline.

## 4.5 Conclusion

We developed a complete and easy-to-use pipeline to find Differentially Methylated Regions. This entirely processes bisulfite sequencing data, from mapping on the reference genome to DMRs computing and filtering. The output consists in DMRs represented by genomic ranged, sorted using various easily interpretable metrics. It is easier to use than other available tools and handle all the steps of the analysis, thus is usable by scientists having little computational experience. The DMRs computing part of the pipeline could be improved in multiple ways. One possible improvement would be to make automatic determination of the sliding window size. The statistical model used to compare conditions could also be improved. Finally, it would be possible to implement already published tools in the pipeline in order to have multiples comparable sources of DMRs and raise the specificity of the results, even if running multiple algorithm is very time-consuming. Moreover, a part dedicated to a global analysis of found DMRs could be implemented at the end of the pipeline. Some possible perspectives concerning this part could be a GO enrichment analysis of genes having multiple DMRs, or establishing a correlation between the DMRs and gene expression if available. The pipeline will be accessible on github as soon as the development is completed.

# Chapter 5

## Side projects

This chapter will describe my participations to projects other than my main project, in independent parts.

### 5.1 *Tisochrysis lutea* genome assembly

*Work performed : genome size estimation and genome assembly*

*Publication : [Berthelier et al., 2018](#)*

#### 5.1.1 Introduction

The general aim of the project was to develop a pipeline to annotate autonomous transposable elements on non-model organisms and run it on the genome of *Tisochrysis lutea*. A draft genome was already available for this organism [Carrier et al., 2018](#) but was very fragmented (7659 contigs, N50 = 10.5 Kb, 54 Mb assembly size). In order to optimally annotate transposable elements, a new genome assembly was performed using PacBio reads. This resulted in a high-contiguity assembly (193 contigs, N50 = 853 Kb, 82 Mb assembly size) on which TEs were annotated.

#### 5.1.2 Methods

19-mer frequency of a library of Mate-Pair illumina data was computed using Jellyfish [Rizk et al., 2013](#). The single peak obtained, corresponding to a kmer depth of 13 was used to estimate the genome size using the following formula :  $\text{genome size} = \text{kmer\_Number} / \text{Peak\_Depth}$ . PacBio reads were assembled using Canu 1.3 [Koren et al., 2017](#). The resulting assembly was polished using Quiver [Chin et al., 2013](#). The polished assembly was subsequently corrected using Pilon v1.20 [Walker et al., 2014](#) and the previous Illumina hiseq mate-paire reads of *T. lutea* ([Carrier et al., 2018](#) SRA: SRR3156597).

### 5.1.3 Results

The genome size was estimated to be around 93 Mb from the K-mer spectrum (Fig. 5.1). The K-mer spectrum shows only one peak which suggests that the genome of *Tisochrysis lutea* has a low heterozygosity.

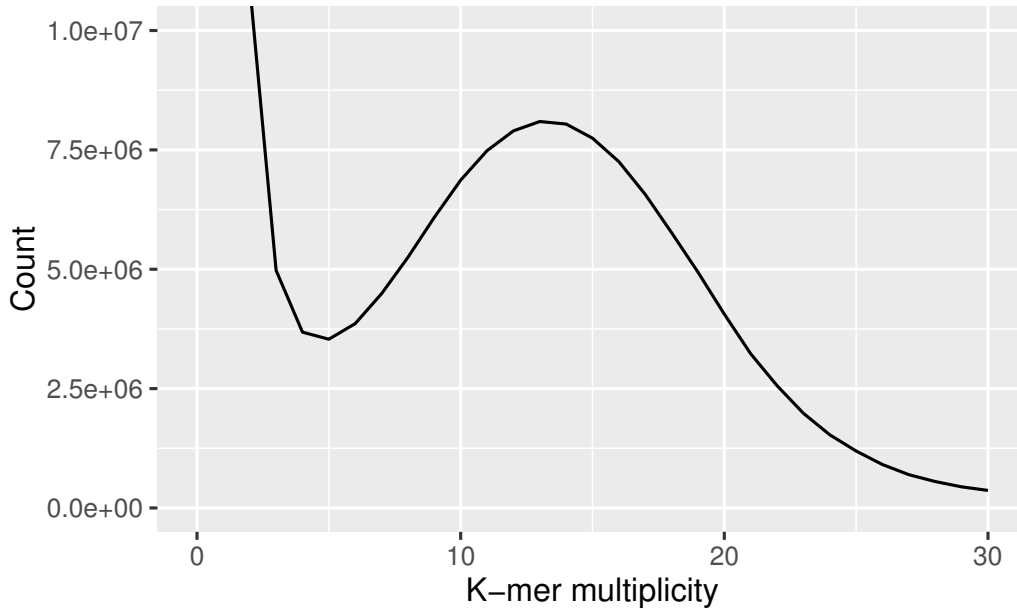


Figure 5.1 k-mer (19 bp) spectra of the *Tisochrysis lutea* genome. The x axis represents k-mer multiplicity, and the y axis represents the number of k-mers with a given multiplicity in the sequencing data.

The new *de novo* assembly of *Tisochrysis lutea* is fragmented in 193 contigs (previous assembly : 7659 contigs), has a N50 of 853 Kb (previous assembly N50 : 10.5Kb) for a total assembly size of 82 Mb (previous assembly size : 54 Mb).

### 5.1.4 Discussion and conclusion

The size of the *de novo* assembly has increased by 28 Mb between both genome versions while the size of the annotated coding region increased only by 3 Mb, from 25 Mb to 28 Mb between the old and new assembly respectively. Moreover the new assembly size is closer to the estimated size of 93 Mb than the old assembly. This suggests that most of the newly assembled sequences correspond to transposable elements, of which repetitive nature makes it difficult to assemble without long reads. Thus, the new assembly represents more accurately the repetitive content of the *Tisochrysis lutea* genome and is suited to an exhaustive detection and annotation of the transposable elements. It provided a strong testing support for the development of the pipeline PiRATE of which the aim is to annotate transposable elements in non-model organisms [Bertheliet al., 2018].

## 5.2 Participation on the rose genome sequencing project

*Work performed on this project : genome polishing, pseudo-molecules building*

*Publication : [\[Saint-Oyant et al., 2018\]](#)*

### 5.2.1 Introduction

Rose is a widely cultivated ornamental plant with great economic value. The aim of this project was to develop a high-quality reference genome to provide a tool to conduct genetic studies to the rose community. Using long and short reads, we generated a rose genome sequence at the pseudo-molecule scale (512 Mbp with N50 of 3.4 Mb and L75 of 97).

### 5.2.2 Methods

The genome polishing was performed using Pilon [\[Walker et al., 2014\]](#) for three iterative rounds like described in paragraph [2.3.2](#). Pseudo-molecules were built using high density female and male genetic maps for a total of 6746 SNP markers [\[Koning-Boucoiran et al., 2015\]](#) like described in paragraph [2.2.2](#). The anchoring of contigs on the genetic maps was done two times iteratively, breaking contigs in case of markers inconsistency as described in **Fig. 5.2** and using the synteny with *Fragaria vesca* in order to not cut the contig inside a gene.

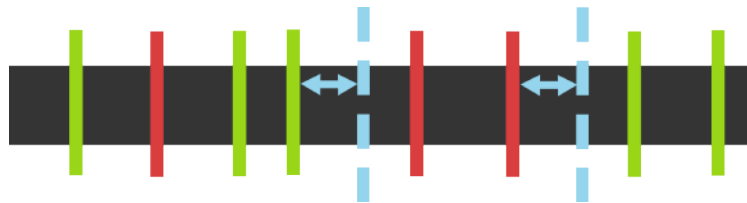


Figure 5.2 Illustration of the cutting of the contigs in case of genetic markers inconsistency. The black line represents the genomic sequence. Green and red lines represent mapped markers belonging to two different chromosomes. Blue dashed lines represent the cutting sites. A contig cutting is done when at least two markers in a row are inconsistent with previously encountered markers. The distance (represented by blue arrows) from the last "at least two in a row" markers encountered at which the contig cutting is performed is variable and decided individually for each cutting event depending on the synteny with *Fragaria vesca*.

### 5.2.3 Results

A total of 37.3k single-base assembly errors and 307.7k indels (341.1 Kbp) were corrected during the polishing. In total, 466 Mbp were anchored on the genetic maps and assembled into seven pseudo-chromosomes representing 90% of the assembly length [5.3](#). The remaining 368 contigs (52 Mbp) were assigned to Chr00.



#### 5.2.4 Discussion and conclusion

We performed the last steps of the rose genome assembly and produced the polished pseudo-molecules. Less round of polishing were needed to attain a very low number of corrections than on apple. While the read coverage was the same in both sequencing efforts, Canu [\[Koren et al., 2017\]](#) was used for the rose assembly and DBG2OLC [\[Ye et al., 2016\]](#) for the apple assembly. Canu performs a self-correction of the PacBio reads as part of its first assembly step which could explain that the non-polished assembly had a lower error rate than the apple non-polished assembly.

However, scaffolding using a genetic map was easier on apple for two main reasons. First, the genetic map was more dense on apple (15,417 markers for 650 Mb estimated genome size) than on rose (6,746 markers for 532 Mb estimated genome size). Second, the contigs used at the time of building the pseudo-molecules were longer on apple (N50 = 5.5 Mb) than on rose (N50 = 3.4 Mb) which raises the number of mapped markers per contig hence the accuracy of the ordering and orienting. This genome was published in [\[Saint-Oyant et al., 2018\]](#).

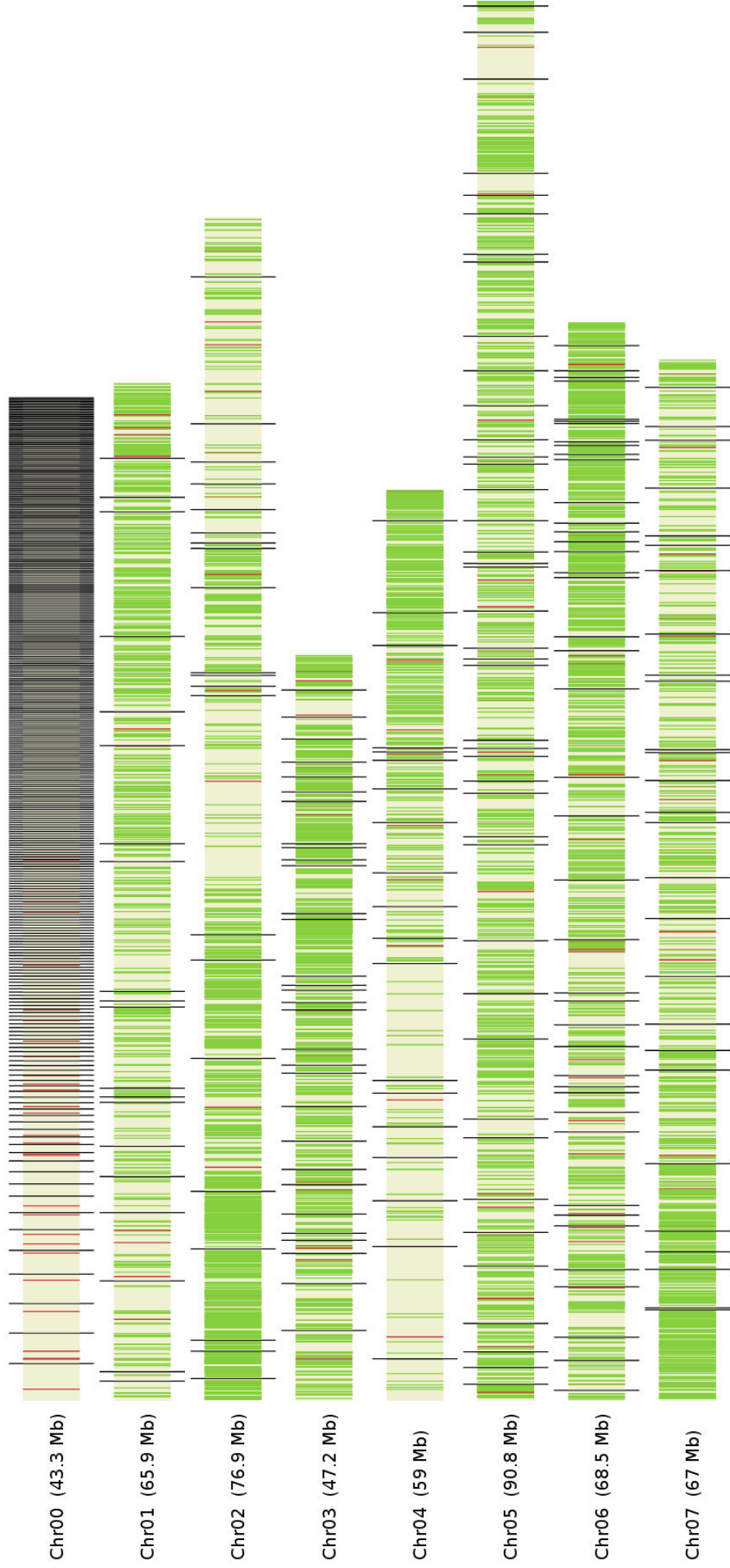


Figure 5.3 Graphical representation of the rose genome pseudo-molecules. Genomic sequence is represented in yellow. Green and red lines represent mapped markers consistent and inconsistent with the genetic map respectively. Black liens represent gaps (N x 10 Kbp) artificially put in order to separate the contigs during the pseudo-molecules building.

# General discussion and conclusion

This work was part of team EpiCenter's research theme which aim's to understand how epigenetics can affect gene expression and important traits in plants and apply this fundamental knowledge to crop breeding. Thus, the initial objective was to study DNA methylation using apple as a model organism. However it appeared that the apple reference genome at the time wasn't good enough to be a support for epigenetics studies, in this case methylation in particular. A good reference genome is characterized by an accurate and contiguous genomic sequence, and an as complete as possible gene annotation. Both of these criterias are frequently considered as bottlenecks when conducting studies on DNA methylation because this phenomenon occurs most of the time in repetitive sequences which typically are not or mis-assembled in reference genome resulting from older sequencing technologies. Moreover, a good gene annotation is mandatory to make interpretations about possible effects of observed epigenetics events.

We decided to produce a new reference apple genome using newer, long-reads sequencing technologies in order to be able to conduct epigenetic studies on apple but also to provide a solid working support for all researchers wanting to conduct genetic or genomic studies on this organism. This necessitated an integration of various types of DNA- and RNA-sequencing data, in particular recent technologies like long PacBio reads and BioNano optical maps which helped to overcome computational challenges such as the reconstitution of repetitive sequences. This work was performed in a collaborative effort with various research institutes and resulted in a new, high-quality reference genome which is described in the second chapter of this manuscript. However, the term "high-quality genome" is entirely subjective of the state of the art at the moment it is published. Indeed, sequencing technologies are evolving very quickly : the competition between the two main companies developing long reads technologies, Pacific Biosciences and Oxford Nanopore, results in a race for longer and more accurate reads, produced at higher throughputs. Moreover, some other companies like BioNano Genomics or Dovetail Genomics are emerging to produce additionnal means to obtain high-quality genomes at lower cost. As a consequence, the reference genome landscape is very different as I end my PhD thesis compared to when I started it in 2015 when long reads technologies were more recent. At this point

of time, aside from model organisms like *A. thaliana* or human, most of the sequenced genomes were in a draft state, which means that they were very fragmented and of a limited use for genetic and genomic studies, mainly due to the fact they weren't made with long reads. During the three years of my PhD thesis, long reads were popularized and more frequently used in sequencing projects. In 2018, a vast majority of the published genome assemblies are chromosome-scaled ; in other words molecules as long as the real chromosomes are produced. We finished to work on the apple genome and published it in 2017. Although the genome assembly is chromosome-scaled and the genome annotation is presumably of good quality (both being orders of magnitude better than in the 2010 reference genome), it still has flaws, like local misassemblies, unanchored scaffolds (Chr00), un-detected genes or collapsed highly paralogous sequences, especially because of the low PacBio coverage used. There is no doubt that it will become of sub-par quality relatively to other reference genomes in the near future, and that it should and will be resequenced at this moment, using more recent sequencing and annotation (long reads RNA-sequencing) technologies. In the meantime, the reference genome we produced can be used (and already began to) as a relatively solid working base for geneticists and genomicists interested in apple or other related plants.

Once we finished assembling and annotating the genome, the main objective of this project was to find if there existed any methylation differences between the two apple lines : GDDH13 and GDDH18 that may explain the fruit size difference. We performed whole genome bisulfite sequencing of these two trees, and compared their methylome using the previously assembled apple genome as a reference. The first approach we used was to compute Differentially Methylated Regions (DMRs) between GDDH13 and GDDH18, then find candidate genes spatially close to these DMRs thus possibly affected in their expression. However we encountered a few issues during this step. The DMRs computing itself is a complicated problem, and despite the fact many tools exist, many weren't adapted for plants because they were developed using mammals references, where only cytosines in the CG context are considered and where the methylation occur in "CG islands" which are methylation clusters upstream of genes. Moreover, because the biological variation of DNA methylation is high, the number of biological replicates we had was too low to compute DMRs specifically enough and some tools were inefficient on our data. Furthermore, especially with a low number of biological replicates, DMRs computing algorithms tend to produce a lot a false positives, which complicates the downstream analysis and necessitates a laborious step of human expertize. Finally, in addition to interpretation issues, DMRs computing also posed some technical problems, among them the numerous file format transformations necessary because of the lack of input format consistency, and the very high computing power and time needed for such calculations.

Having DMRs at our disposal, the course of action we chose was, for each DMR, to recense every neighbour-

ing genes, especially the ones having a DMR in their upstream region, based on the assumption a methylation difference in the putative promoter could provoke a gene expression difference, thus a phenotypic change. We found a list of genes which could be potentially be involved in fruit size and have one or more DMR upstream of their TSS. These results are described in the second part the third chapter of this manuscript. However, we couldn't find any expression differences between GDDH13 and GDDH18 for these genes. This made us reconsiderate the link between DNA methylation and gene expression. To ensure this correlation exist on the apple genome, we made a few genome wide analyses comparing the methylation patterns upstream and inside the genes and the corresponding expression. We confirmed a negative correlation between CG and CHG methylation upstream of the gene and expression, and a positive correlation concerning CHH methylation. These results are described in the first part the third chapter of this manuscript. However, while this global correlation exists, we couldn't validate it by looking each gene and DMRs individually since there wasn't any overlap between differentially methylated and differentially expressed genes between GDDH13 and GDDH18. A better way to potentially answer to the fruit size question would have been to start from the list of differentially expressed genes and seek any methylation differences around these genes specifically, because it diminishes drastically the number of loci that need to be checked for differential methylation and asks for less manual expertize. However, even if a localized methylation difference can be associated with a difference of gene expression, it is impossible to say if the expression change is a consequence or a simple correlation with the methylation change. It has to be biologically verified, which is expensive in time depending on the number of loci to check.

Considering the several technical and interpretations issues arising from DMRs computing, which makes it difficult for people unfamiliar with programming, and its importance, we decided to develop an easy-to-use pipeline handling all the bisulfite data analysis. While most of the work done is automatizing technical work, like file format conversion, and while there is no algorithm novelty, this pipeline can help anyone without programming experience to compute a set of relatively specific DMRs, with any number of biological replicates. While this pipeline works with high (>5) number of replicates, it was developed and optimized on experiments having only two replicates. The reason for this, aside from the fact our own sequencing data had four replicates at most, is that the financial cost of whole genome bisulfite sequencing usually doesn't allow to perform a lot of sequencing runs. However, this pipeline outputs a lot of DMRs and the user has the responsibility to choose the ones he deems significant. Indeed, there is a part of subjectivity in determining if a DMR is "real" or not. Theoretically, a DMR can be a region where the methylation difference between two samples is as high as 100% or as low as 5%. Thus, the user should arbitrarily choose a minimum methylation difference threshold where he thinks such a difference can have an impact on gene expression, although

the biologically relevant level of DNA methylation change is currently unknown. More generally, the main limit of DMRs computing with a low number of replicates is detecting small DMRs. It's possible that gene expression could be affected by the methylation of one or two cytosine, hypothetically at the transcription factors binding site. Differentially methylated cytosines are impossible to detect reliably with this type of method adapted for a low number of replicates. To do this, one should sequence at least five biological replicates and use appropriate tools.

The work performed during this PhD project allowed me to handle several types of -omics data and apply a wide range of common and less common methods on it to produce scientific results and tools. I applied the skills I learned working on my main subject in parallel of it, participating in a few side projects that are described in the fifth chapter of this manuscript. However, my inexperience at the beginning of the thesis was a little detrimental to the overall output quality and time to produce it. However being implied in this project was a chance and made me gain a lot of experience in genomics and programming which will allow me to be better on the future projects I will be associated with as a bioinformatician.

## Supplementary data

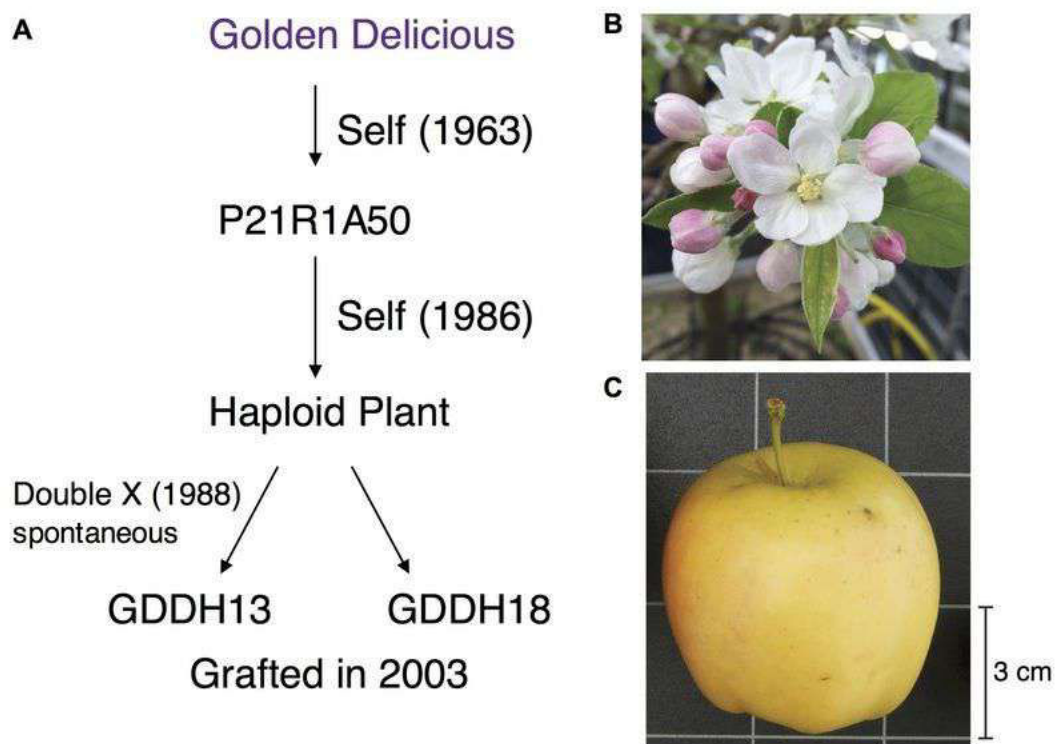


Figure S1 GDDH13 obtention process. (textbfa) In 1963 the 'Golden Delicious' variety was selfed once resulting in P21R1A50 (confirmed by isoenzymatic analysis). This line was then self-pollinated in 1986 resulting in a haploid plant from an unfertilized egg cell rather than a zygote in 1987. Samples of this haploid line were put in vitro, which resulted in spontaneous doubling events (1988, indicated by more vigorous growth). Root formation was induced in vitro and the plants were transferred to the orchard on their own roots resulting in GDDH13 (X9273). GDDH13 was then grafted in 2003. (textbfb) and (textbfc) show photographs of flowers and a fruit of GDDH13.

|   | Number  | Total size (Mb) | % of assembly | % of genome size |
|---|---------|-----------------|---------------|------------------|
| <b>Contigs</b>                                  | 912     | 57.1            | 8.8           | 8.0              |
| <b>Scaffolds</b>                                |         |                 |               |                  |
| With N  | 169     | 643.2           | 99.0          | 90.2             |
| Without N                                       | 169     | 568.2           | 87.5          | 79.7             |
| <b>Scaffolds and contigs anchored without N</b> | 280     | 580.37          | 89.3          | 81.4             |
| <b>Chr0 without N</b>                           | 801     | 45              | 6.9           | 6.3              |
| <b>Total without N</b>                          | 1081    | 625.37          | 96.3          | 87.7             |
| <b>Genes</b>                                    |         |                 |               |                  |
| Protein coding genes                            | 42,140  | 151.2           | 23.27         | 21.21            |
| Non protein coding genes                        | 1,965   | 4               | 0.62          | 0.56             |
| <b>TE</b>                                       |         |                 |               |                  |
| Class I   | 393,464 | 301.1           | 46.34         | 42.23            |
| Class II  | 299,637 | 87.3            | 13.44         | 12.24            |
| unclassified                                    | 56,487  | 12.4            | 1.91          | 1.74             |
| Total   | 749,588 | 400.8           | 61.69         | 56.21            |

Table S1 Summary of the genome assembly features and annotations of the apple GDDH13 genome.



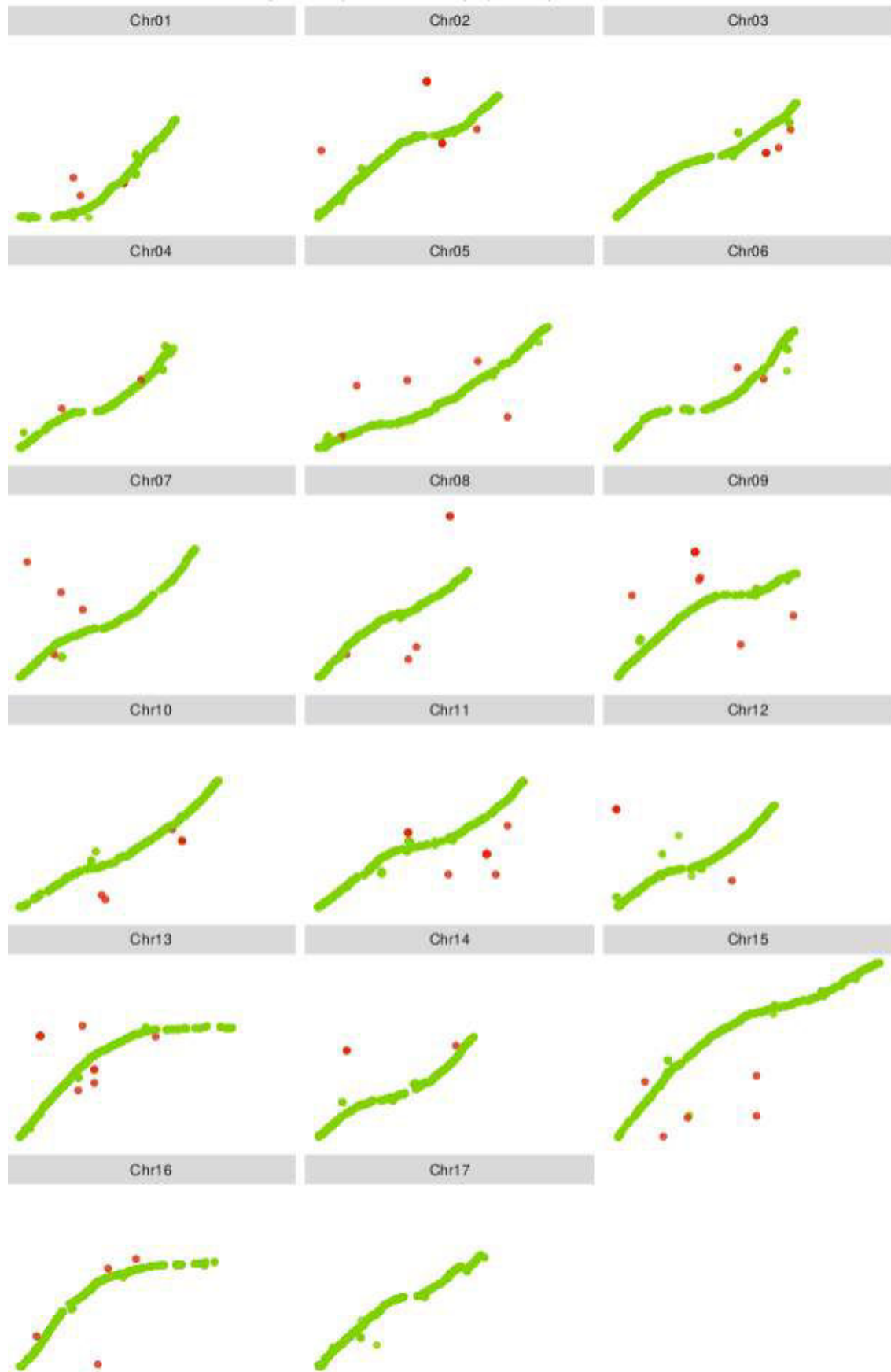


Figure S2 Comparison of the genetic and physical GDDH13 position of the SNP markers of the integrated genetic map. Graphical representation of the location of Single Nucleotide Polymorphism (SNP) markers on the physical map (x axis) compared to their position on the integrated genetic map (y axis) for all chromosomes. Each marker is plotted in the color of the chromosome to which it has been genetically mapped to. All markers mapping to the correct chromosome compared to the genetic map are plotted in green, the ones to the wrong chromosome in red

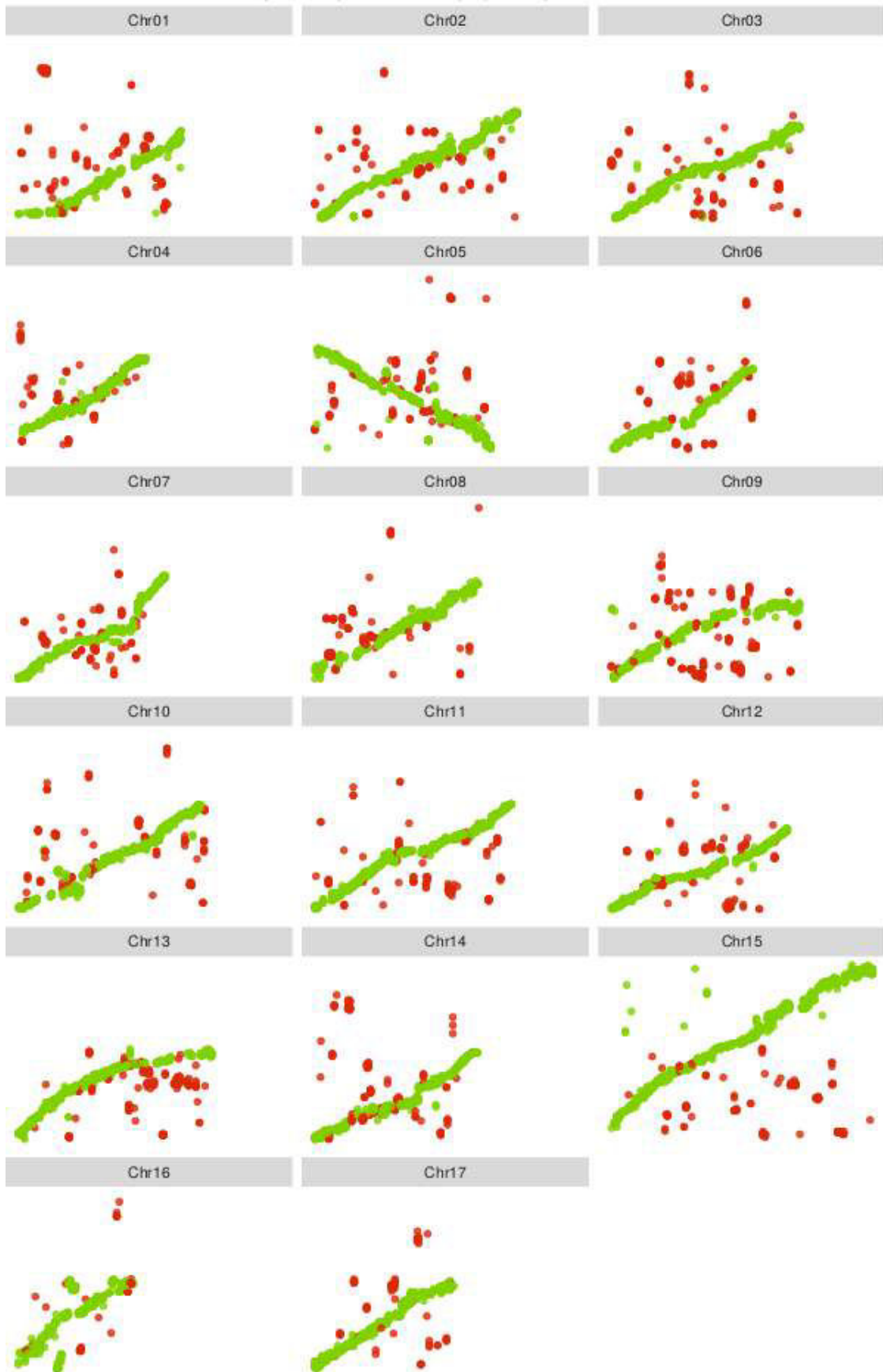


Figure S3 Like Fig. S2 but using the previous genome release [Velasco et al., 2010].

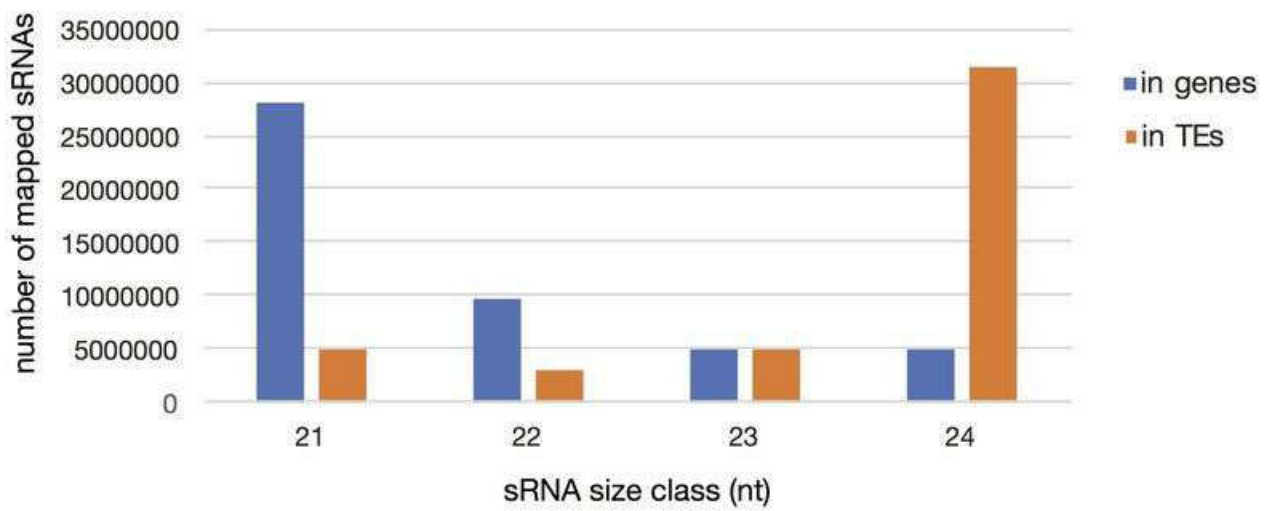


Figure S4 Histogram showing the number of mapped sRNAs of 21, 22, 23 and 24 nucleotides on genes and transposable elements of the GDDH13 genome.

| Exon ID            | Variant position | Variant type        | Gene element | Arabidopsis homologous gene | aa cl  |
|--------------------|------------------|---------------------|--------------|-----------------------------|--------|
| MD10G0010200.1     | Chr10:1822803    | SNP                 | UTR          | AT4G28390.1                 | NA     |
| MD10G0129700.1     | Chr10:24214044   | SNP heterozygous    | exon         | AT3G25280.1                 | silent |
| MD10G0135400.3     | Chr10:25126166   | Insertion in GDDH18 | UTR          | AT3G07220.1                 | NA     |
| MD10G0210300.utr0  | Chr10:34536717   | Insertion in GDDH18 | UTR          | AT1G62020.1                 | NA     |
| MD14G0029900.1     | Chr14:2986061    | SNP                 | UTR          | AT3G51680.1                 | NA     |
| MD15G0318900.utr0  | Chr15:45714269   | SNP                 | UTR          | AT1G78300.1                 | NA     |
| MD16G0085600.15    | Chr16:6683664    | SNP                 | UTR          | AT1G10670.4                 | NA     |
| MD16G0101300.9     | Chr16:8011011    | SNP                 | exon         | AT2G36670.2                 | Ala    |
| MD16G0108100.2     | Chr16:8730650    | SNP                 | UTR          | AT1G27290.2                 | NA     |
| MD16G0168400.4     | Chr16:16918723   | SNP                 | exon         | AT2G32150.1                 | Pro    |
| MD16G0177400.5     | Chr16:18324253   | SNP                 | UTR          | AT3G22890.1                 | NA     |
| MD16G0229600.1     | Chr16:35165536   | Insertion in GDDH18 | UTR          | AT1G32330.1                 | NA     |
| MD10G0039300.utr18 | Chr10:6542107    | Insertion in GDDH18 | UTR          | AT4G13070.1                 | NA     |
| MD00G0026800.2     | Chr00:6410135    | SNP                 | exon         | AT2G46660.1                 | Gly    |
| MD01G0115900.1     | Chr01:25604048   | SNP                 | exon         | AT2G37980.1                 | silent |
| MD02G0046700.1     | Chr02:4534877    | SNP                 | exon         | AT4G13360.1                 | Met    |
| MD02G0129000.1     | Chr02:12468297   | Insertion in GDDH18 | UTR          | AT4G36670.1                 | NA     |
| MD02G0169400.1     | Chr02:19140899   | Insertion in GDDH18 | UTR          | AT2G45650.1                 | NA     |
| MD05G0198000.2     | Chr05:36897333   | SNP                 | exon         | AT1G79480.1                 | Gln    |
| MD05G0246800.3     | Chr05:42512357   | SNP heterozygous    | exon         | AT1G17020.1                 | Silent |
| MD06G0028500.1     | Chr06:4267294    | SNP                 | exon         | AT1G32640.1                 | Silent |
| MD06G0030900.2     | Chr06:4750165    | SNP heterozygous    | exon         | AT5G16770.2                 | Arg    |
| MD07G0019400.1     | Chr07:2079538    | SNP                 | exon         | AT3G63460.1                 | Arg    |
| MD07G0085500.8     | Chr07:11860469   | SNP                 | UTR          | AT3G59670.1                 | NA     |
| MD07G0262300.8     | Chr07:36689374   | SNP                 | exon         | AT5G17520.1                 | Ile    |
| MD08G0170700.2     | Chr08:26943525   | SNP                 | exon         | AT1G58290.1                 | Arg    |
| MD09G0203000.2     | Chr09:29608286   | SNP                 | exon         | AT1G15210.1 (bad E-Value)   | NA     |

Table S2 List of CDS elements in which a genetic variant between GDDH13 and GDDH18 was found.

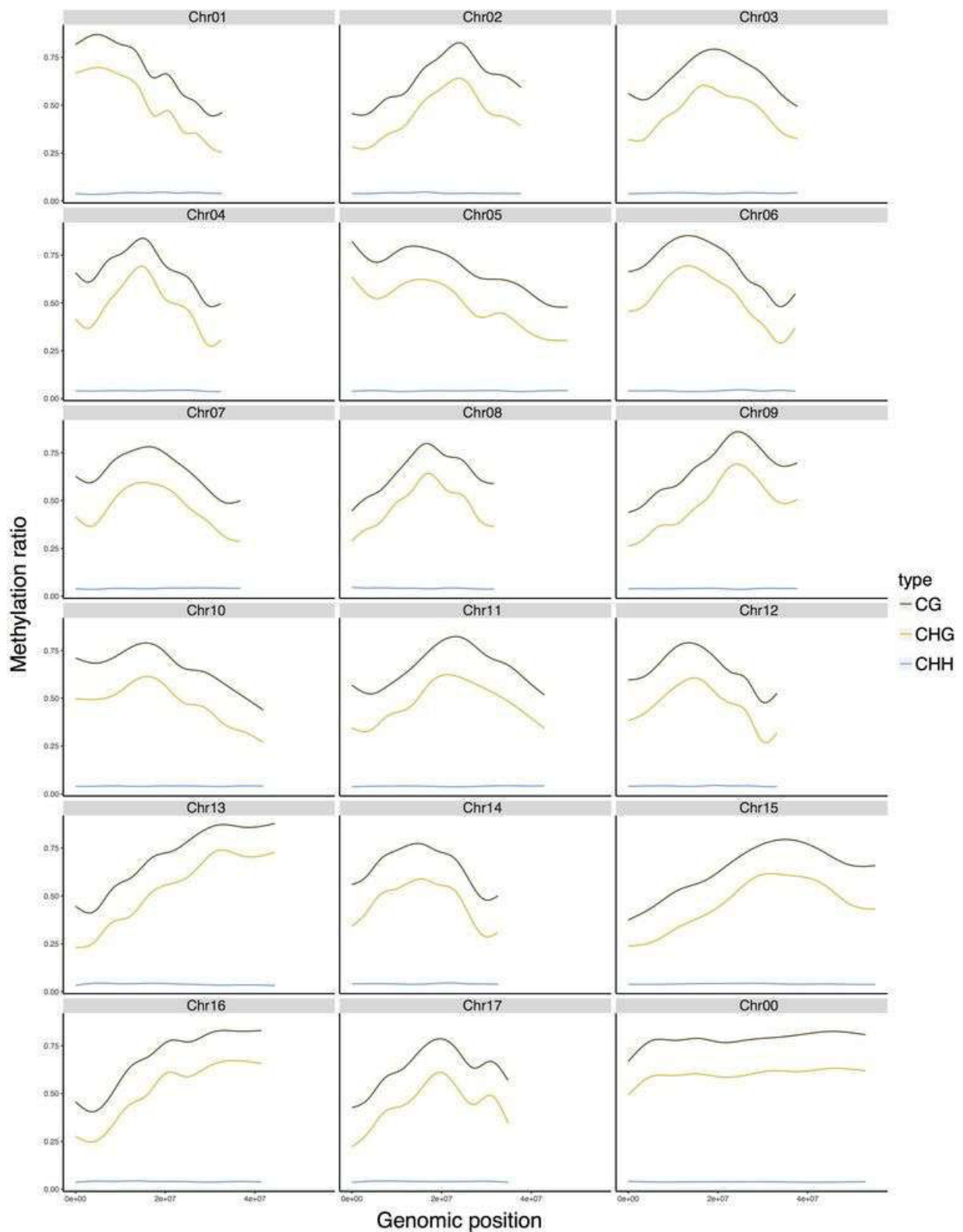


Figure S5 Genomic DNA methylation density in GDDH13. DNA methylation ratios were plotted for all chromosomes (including Chr00) and the three DNA sequence contexts: CG in brown, CHG in yellow and CHH in blue.

| MD gene nb   | DMR context  | Overmethylated in | Arabidopsis homolog | pvalue |
|--------------|--------------|-------------------|---------------------|--------|
| MD11G0093600 | CG           | 13                | AT3G46200.1         | 1e-138 |
| MD07G0163300 | CG           | 13                | AT3G08820.1         | 2e-32  |
| MD03G0045100 | CG           | 13                | AT2G13800.1         | 1.7    |
| MD02G0178100 | CG           | 13                | AT3G60680.1         | 3e-76  |
| MD17G0035800 | CG           | 13                | AT1G70180.2         | 9e-19  |
| MD15G0016600 | CG           | 13                | AT2G27190.1         | 0      |
| MD02G0158100 | CG           | 13                | AT1G64960.1         | 0      |
| MD11G0242100 | CG           | 13                | AT2G30780.1         | 1e-91  |
| MD10G0010900 | CG           | 13                | AT1G77120.1         | 0      |
| MD09G0211500 | CG           | 13                | AT2G45100.1         | 2.2    |
| MD01G0031500 | CG           | 13                | AT2G15430.1         | 0.31   |
| MD15G0127800 | CG, CHG      | 18                | AT4G37770.1         | 0      |
| MD15G0055300 | CG, CHG      | 18                | #N/A                | #N/A   |
| MD15G0055100 | CG, CHG      | 18                | AT1G06590.1         | 0.7    |
| MD01G0027800 | CG, CHG      | 18                | AT5G65540.1         | 0      |
| MD15G0054700 | CG, CHG      | 18                | AT5G65780.1         | 1e-139 |
| MD15G0055300 | CG, CHG      | 18                | #N/A                | #N/A   |
| MD10G0134000 | CG, CHG      | 13                | AT5G48630.1         | 2e-141 |
| MD07G0123200 | CG, CHG      | 13                | AT3G61760.1         | 0      |
| MD00G0050700 | CG, CHG      | 13                | AT3G11910.2         | 6e-42  |
| MD11G0212000 | CG, CHG      | 13                | AT1G78860.1         | 7e-118 |
| MD12G0013500 | CG, CHG      | 13                | AT5G27220.1         | 1e-28  |
| MD10G0281600 | CG, CHG      | 13                | AT4G22320.2         | 9e-47  |
| MD16G0163500 | CG, CHG      | 13                | AT5G19140.1         | 2e-74  |
| MD06G0088800 | CG, CHG      | 13                | AT5G46290.1         | 0      |
| MD06G0088700 | CG, CHG      | 13                | AT2G32460.1         | 0.36   |
| MD15G0020800 | CG, CHG      | 13                | AT1G47250.1         | 2e-159 |
| MD00G0032300 | CG, CHG      | 13                | AT4G09720.3         | 2.7    |
| MD00G0032200 | CG, CHG      | 13                | AT4G00150.1         | 1e-121 |
| MD03G0056400 | CG, CHG      | 13                | AT5G05580.1         | 0      |
| MD11G0002500 | CG, CHG      | 13                | AT5G56460.1         | 0      |
| MD11G0002400 | CG, CHG      | 13                | AT4G26330.1         | 7e-09  |
| MD09G0038900 | CG, CHG      | 13                | AT5G27670.1         | 2e-12  |
| MD14G0133700 | CG, CHG, CHH | 13                | AT5G50570.2         | 9e-58  |
| MD00G0035400 | CG, CHH      | 13                | AT1G49510.1         | 2e-07  |
| MD14G0147300 | CHG          | 18                | AT3G48280.1         | 2e-73  |
| MD16G0108400 | CHG          | 18                | AT5G50570.2         | 7e-43  |
| MD07G0069000 | CHG          | 13                | AT5G06600.3         | 9e-43  |
| MD14G0111000 | CHG          | 13                | AT4G23900.1         | 2e-136 |
| MD16G0111500 | CHG          | 13                | AT2G33490.1         | 0      |
| MD01G0131900 | CHG          | 13                | AT5G05330.1         | 9e-28  |
| MD15G0353900 | CHG          | 13                | AT5G44080.1         | 1e-45  |
| MD12G0150600 | CHG          | 13                | AT2G02030.1         | 9e-15  |
| MD17G0001900 | CHG          | 13                | #N/A                | #N/A   |
| MD10G0167800 | CHG          | 13                | AT5G16820.2         | 3e-142 |
| MD09G0086300 | CHG          | 13                | AT4G03230.1         | 2e-26  |
| MD10G0272900 | CHG          | 13                | AT3G16730.1         | 2.8    |

Table S3 Differentially methylated genes in the comparison between GDDH13 and GDDH18.

# Bibliography

- [Aceituno et al., 2008] Aceituno, F. F., Moseyko, N., Rhee, S. Y., and Gutiérrez, R. A. (2008). The rules of gene expression in plants: organ identity and gene body methylation are key factors for regulation of gene expression in *arabidopsis thaliana*. *BMC genomics*, 9(1):438.
- [Akalin et al., 2012] Akalin, A., Kormaksson, M., Li, S., Garrett-Bakelman, F. E., Figueroa, M. E., Melnick, A., and Mason, C. E. (2012). methylkit: a comprehensive r package for the analysis of genome-wide dna methylation profiles. *Genome biology*, 13(10):R87.
- [Allis and Jenuwein, 2016] Allis, C. D. and Jenuwein, T. (2016). The molecular hallmarks of epigenetic control. *Nature Reviews Genetics*, 17(8):487.
- [Alonso et al., 2016] Alonso, C., Pérez, R., Bazaga, P., Medrano, M., and Herrera, C. M. (2016). Msap markers and global cytosine methylation in plants: a literature survey and comparative analysis for a wild-growing species. *Molecular Ecology Resources*, 16(1):80–90.
- [Ansorge, 2016] Ansorge, W. (2016). Next-generation dna sequencing (ii): techniques, applications. *Next Generat. Sequenc. & Applic*, 1:1–10.
- [Aryee et al., 2014] Aryee, M. J., Jaffe, A. E., Corrada-Bravo, H., Ladd-Acosta, C., Feinberg, A. P., Hansen, K. D., and Irizarry, R. A. (2014). Minfi: a flexible and comprehensive bioconductor package for the analysis of infinium dna methylation microarrays. *Bioinformatics*, 30(10):1363–1369.
- [Badouin et al., 2015] Badouin, H., Hood, M. E., Gouzy, J., Aguilera, G., Siguenza, S., Perlin, M. H., Cuomo, C. A., Fairhead, C., Branca, A., and Giraud, T. (2015). Chaos of rearrangements in the mating-type chromosomes of the anther-smut fungus *microbotryum lychnidis-dioicae*. *Genetics*, 200(4):1275–1284.
- [Bao et al., 2015] Bao, W., Kojima, K. K., and Kohany, O. (2015). Repbase update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA*, 6(1):11.
- [Baylin, 2005] Baylin, S. B. (2005). Dna methylation and gene silencing in cancer. *Nature Reviews Clinical Oncology*, 2(S1):S4.
- [Becker et al., 2011] Becker, C., Hagmann, J., Müller, J., Koenig, D., Stegle, O., Borgwardt, K., and Weigel, D. (2011). Spontaneous epigenetic variation in the *arabidopsis thaliana* methylome. *Nature*, 480(7376):245.

- [Bennett, 2004] Bennett, S. (2004). Solexa ltd. *Pharmacogenomics*, 5(4):433–438.
- [Bertheliet al., 2018] Bertheliet, J., Casse, N., Daccord, N., Jamilloux, V., Saint-Jean, B., and Carrier, G. (2018). A transposable element annotation pipeline and expression analysis reveal potentially active elements in the microalga *tisochrysis lutea*. *BMC genomics*, 19(1):378.
- [Bianco et al., 2016] Bianco, L., Cestaro, A., Linsmith, G., Muranty, H., Denance, C., Théron, A., Poncet, C., Micheletti, D., Kerschbamer, E., Di Pierro, E. A., et al. (2016). Development and validation of the axiom® apple480k snp genotyping array. *The Plant Journal*, 86(1):62–74.
- [Bianco et al., 2014] Bianco, L., Cestaro, A., Sargent, D. J., Banchi, E., Derdak, S., Di Guardo, M., Salvi, S., Jansen, J., Viola, R., Gut, I., et al. (2014). Development and validation of a 20k single nucleotide polymorphism (snp) whole genome genotyping array for apple (*malus× domestica borkh*). *PLoS One*, 9(10):e110377.
- [Bird, 2002] Bird, A. (2002). Dna methylation patterns and epigenetic memory. *Genes & development*, 16(1):6–21.
- [Bird, 1984] Bird, A. P. (1984). Gene expression: Dna methylation – how important in gene control? *Nature*, 307(5951):503.
- [Carrier et al., 2018] Carrier, G., Baroukh, C., Rouxel, C., Duboscq-Bidot, L., Schreiber, N., and Bougaran, G. (2018). Draft genomes and phenotypic characterization of *tisochrysis lutea* strains. toward the production of domesticated strains with high added value. *Algal Research*, 29:1–11.
- [Cedar, 1988] Cedar, H. (1988). Dna methylation and gene activity. *Cell*, 53(1):3–4.
- [Celton et al., 2014] Celton, J.-M., Gaillard, S., Bruneau, M., Pelletier, S., Aubourg, S., Martin-Magniette, M.-L., Navarro, L., Laurens, F., and Renou, J.-P. (2014). Widespread anti-sense transcription in apple is correlated with sirna production and indicates a large potential for transcriptional and/or post-transcriptional control. *New Phytologist*, 203(1):287–299.
- [Chaisson and Tesler, 2012] Chaisson, M. J. and Tesler, G. (2012). Mapping single molecule sequencing reads using basic local alignment with successive refinement (blasr): application and theory. *BMC bioinformatics*, 13(1):238.
- [Chénais et al., 2012] Chénais, B., Caruso, A., Hiard, S., and Casse, N. (2012). The impact of transposable elements on eukaryotic genomes: from genome size increase to genetic adaptation to stressful environments. *Gene*, 509(1):7–15.
- [Chin et al., 2013] Chin, C.-S., Alexander, D. H., Marks, P., Klammer, A. A., Drake, J., Heiner, C., Clum, A., Copeland, A., Huddleston, J., Eichler, E. E., et al. (2013). Nonhybrid, finished microbial genome assemblies from long-read smrt sequencing data. *Nature methods*, 10(6):563.



- [Chin et al., 2016] Chin, C.-S., Peluso, P., Sedlazeck, F. J., Nattestad, M., Concepcion, G. T., Clum, A., Dunn, C., O'Malley, R., Figueroa-Balderas, R., Morales-Cruz, A., et al. (2016). Phased diploid genome assembly with single-molecule real-time sequencing. *Nature methods*, 13(12):1050.
- [Clarke et al., 2009] Clarke, J., Wu, H.-C., Jayasinghe, L., Patel, A., Reid, S., and Bayley, H. (2009). Continuous base identification for single-molecule nanopore dna sequencing. *Nature nanotechnology*, 4(4):265.
- [Clayton and Leung, 2007] Clayton, D. and Leung, H.-T. (2007). An r package for analysis of whole-genome association studies. *Human heredity*, 64(1):45–51.
- [Cokus et al., 2008] Cokus, S. J., Feng, S., Zhang, X., Chen, Z., Merriman, B., Haudenschild, C. D., Pradhan, S., Nelson, S. F., Pellegrini, M., and Jacobsen, S. E. (2008). Shotgun bisulphite sequencing of the arabidopsis genome reveals dna methylation patterning. *Nature*, 452(7184):215.
- [Consortium et al., 2012a] Consortium, E. P. et al. (2012a). An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57.
- [Consortium et al., 2014] Consortium, I. W. G. S. et al. (2014). A chromosome-based draft sequence of the hexaploid bread wheat (*triticum aestivum*) genome. *Science*, 345(6194):1251788.
- [Consortium et al., 2012b] Consortium, T. G. et al. (2012b). The tomato genome sequence provides insights into fleshy fruit evolution. *Nature*, 485(7400):635.
- [Cornille et al., 2014] Cornille, A., Giraud, T., Smulders, M. J., Roldán-Ruiz, I., and Gladieux, P. (2014). The domestication and evolutionary ecology of apples. *Trends in Genetics*, 30(2):57–65.
- [Cubas et al., 1999] Cubas, P., Vincent, C., and Coen, E. (1999). An epigenetic mutation responsible for natural variation in floral symmetry. *Nature*, 401(6749):157.
- [Cui et al., 2006] Cui, L., Wall, P. K., Leebens-Mack, J. H., Lindsay, B. G., Soltis, D. E., Doyle, J. J., Soltis, P. S., Carlson, J. E., Arumuganathan, K., Barakat, A., et al. (2006). Widespread genome duplications throughout the history of flowering plants. *Genome research*, 16(6):738–749.
- [Daccord et al., 2017] Daccord, N., Celton, J.-M., Linsmith, G., Becker, C., Choisne, N., Schijlen, E., van de Geest, H., Bianco, L., Micheletti, D., Velasco, R., et al. (2017). High-quality de novo assembly of the apple genome and methylome dynamics of early fruit development. *Nature genetics*, 49(7):1099.
- [Denton et al., 2014] Denton, J. F., Lugo-Martinez, J., Tucker, A. E., Schrider, D. R., Warren, W. C., and Hahn, M. W. (2014). Extensive error in the number of genes inferred from draft genome assemblies. *PLoS computational biology*, 10(12):e1003998.
- [Deshpande et al., 2013] Deshpande, V., Fung, E. D., Pham, S., and Bafna, V. (2013). Cerulean: A hybrid assembly using high throughput short and long reads. pages 349–363.

- [Di Croce et al., 2002] Di Croce, L., Raker, V. A., Corsaro, M., Fazi, F., Fanelli, M., Faretta, M., Fuks, F., Coco, F. L., Kouzarides, T., Nervi, C., et al. (2002). Methyltransferase recruitment and dna hypermethylation of target promoters by an oncogenic transcription factor. *Science*, 295(5557):1079–1082.
- [Di Pierro et al., 2016] Di Pierro, E. A., Gianfranceschi, L., Di Guardo, M., Koehorst-van Putten, H. J., Kruisselbrink, J. W., Longhi, S., Troglio, M., Bianco, L., Muranty, H., Pagliarani, G., et al. (2016). A high-density, multi-parental snp genetic map on apple validates a new mapping approach for outcrossing species. *Horticulture research*, 3:16057.
- [Dolzhenko and Smith, 2014] Dolzhenko, E. and Smith, A. D. (2014). Using beta-binomial regression for high-precision differential methylation analysis in multifactor whole-genome bisulfite sequencing experiments. *BMC bioinformatics*, 15(1):215.
- [Dudchenko et al., 2017] Dudchenko, O., Batra, S. S., Omer, A. D., Nyquist, S. K., Hoeger, M., Durand, N. C., Shamim, M. S., Machol, I., Lander, E. S., Aiden, A. P., et al. (2017). De novo assembly of the aedes aegypti genome using hi-c yields chromosome-length scaffolds. *Science*, 356(6333):92–95.
- [Dupont et al., 2009] Dupont, C., Armant, D. R., and Brenner, C. A. (2009). Epigenetics: definition, mechanisms and clinical perspective. In *Seminars in reproductive medicine*, volume 27, page 351. NIH Public Access.
- [Edger et al., 2017] Edger, P. P., VanBuren, R., Colle, M., Poorten, T. J., Wai, C. M., Niederhuth, C. E., Alger, E. I., Ou, S., Acharya, C. B., Wang, J., et al. (2017). Single-molecule sequencing and optical mapping yields an improved genome of woodland strawberry (*fragaria vesca*) with chromosome-scale contiguity. *GigaScience*.
- [Ehrlich et al., 1982] Ehrlich, M., Gama-Sosa, M. A., Huang, L.-H., Midgett, R. M., Kuo, K. C., McCune, R. A., and Gehrke, C. (1982). Amount and distribution of 5-methylcytosine in human dna from different types of tissues or cells. *Nucleic acids research*, 10(8):2709–2721.
- [Eid et al., 2009] Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., et al. (2009). Real-time dna sequencing from single polymerase molecules. *Science*, 323(5910):133–138.
- [El-Sharkawy et al., 2015] El-Sharkawy, I., Liang, D., and Xu, K. (2015). Transcriptome analysis of an apple (*malus × domestica*) yellow fruit somatic mutation identifies a gene network module highly associated with anthocyanin and epigenetic regulation. *Journal of experimental botany*, 66(22):7359–7376.
- [Ellinghaus et al., 2008] Ellinghaus, D., Kurtz, S., and Willhoeft, U. (2008). Ltrharvest, an efficient and flexible software for de novo detection of ltr retrotransposons. *BMC bioinformatics*, 9(1):18.

- [Emanuelsson et al., 2007] Emanuelsson, O., Brunak, S., Von Heijne, G., and Nielsen, H. (2007). Locating proteins in the cell using targetp, signalp and related tools. *Nature protocols*, 2(4):953.
- [Fedoroff, 2012] Fedoroff, N. V. (2012). Transposable elements, epigenetics, and genome evolution. *Science*, 338(6108):758–767.
- [Feng et al., 2014] Feng, H., Conneely, K. N., and Wu, H. (2014). A bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data. *Nucleic acids research*, 42(8):e69–e69.
- [Feschotte et al., 2002] Feschotte, C., Jiang, N., and Wessler, S. R. (2002). Plant transposable elements: where genetics meets genomics. *Nature Reviews Genetics*, 3(5):329.
- [Finn et al., 2015] Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., Potter, S. C., Punta, M., Qureshi, M., Sangrador-Vegas, A., et al. (2015). The pfam protein families database: towards a more sustainable future. *Nucleic acids research*, 44(D1):D279–D285.
- [Florea et al., 2011] Florea, L., Souvorov, A., Kalbfleisch, T. S., and Salzberg, S. L. (2011). Genome assembly has a major impact on gene content: a comparison of annotation in two bos taurus assemblies. *PLoS One*, 6(6):e21400.
- [Flutre et al., 2011] Flutre, T., Duprat, E., Feuillet, C., and Quesneville, H. (2011). Considering transposable element diversification in de novo annotation approaches. *PloS one*, 6(1):e16526.
- [Foissac et al., 2008] Foissac, S., Gouzy, J., Rombauts, S., Mathé, C., Amselem, J., Sterck, L., de Peer, Y. V., Rouzé, P., and Schiex, T. (2008). Genome annotation in plants and fungi: Eugene as a model platform. *Current Bioinformatics*, 3(2):87–97.
- [Gallusci et al., 2016] Gallusci, P., Hodgman, C., Teyssier, E., and Seymour, G. B. (2016). Dna methylation and chromatin regulation during fleshy fruit development and ripening. *Frontiers in plant science*, 7:807.
- [Garrison and Marth, 2012] Garrison, E. and Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. *arXiv preprint arXiv:1207.3907*.
- [Girgis, 2015] Girgis, H. Z. (2015). Red: an intelligent, rapid, accurate tool for detecting repeats de-novo on the genomic scale. *BMC bioinformatics*, 16(1):227.
- [Gore et al., 2009] Gore, M. A., Chia, J.-M., Elshire, R. J., Sun, Q., Ersoz, E. S., Hurwitz, B. L., Peiffer, J. A., McMullen, M. D., Grills, G. S., Ross-Ibarra, J., et al. (2009). A first-generation haplotype map of maize. *Science*, 326(5956):1115–1117.
- [Grabherr et al., 2011] Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., et al. (2011). Full-length transcriptome assembly from rna-seq data without a reference genome. *Nature biotechnology*, 29(7):644.

- [Guéguen et al., 2013] Guéguen, L., Gaillard, S., Boussau, B., Gouy, M., Groussin, M., Rochette, N. C., Bigot, T., Fournier, D., Pouyet, F., Cahais, V., et al. (2013). Bio++: efficient extensible libraries and tools for computational molecular evolution. *Molecular biology and evolution*, 30(8):1745–1750.
- [Guo and Simmons, 2011] Guo, M. and Simmons, C. R. (2011). Cell number counts the fw2. 2 and cnr genes and implications for controlling plant fruit and organ size. *Plant Science*, 181(1):1–7.
- [Guo et al., 2013] Guo, S., Zhang, J., Sun, H., Salse, J., Lucas, W. J., Zhang, H., Zheng, Y., Mao, L., Ren, Y., Wang, Z., et al. (2013). The draft genome of watermelon (*Citrullus lanatus*) and resequencing of 20 diverse accessions. *Nature genetics*, 45(1):51.
- [Hackl et al., 2014] Hackl, T., Hedrich, R., Schultz, J., and Förster, F. (2014). proovread: large-scale high-accuracy pacbio correction through iterative short read consensus. *Bioinformatics*, 30(21):3004–3011.
- [Hagmann et al., 2015] Hagmann, J., Becker, C., Müller, J., Stegle, O., Meyer, R. C., Wang, G., Schneeberger, K., Fitz, J., Altmann, T., Bergelson, J., et al. (2015). Century-scale methylome stability in a recently diverged *Arabidopsis thaliana* lineage. *PLoS genetics*, 11(1):e1004920.
- [Hansen et al., 2012] Hansen, K. D., Langmead, B., and Irizarry, R. A. (2012). Bsmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome biology*, 13(10):R83.
- [Hatakeyama et al., 2017] Hatakeyama, M., Aluri, S., Balachadran, M. T., Sivarajan, S. R., Patrignani, A., Grüter, S., Poveda, L., Shimizu-Inatsugi, R., Baeten, J., Francoijs, K.-J., et al. (2017). Multiple hybrid de novo genome assembly of finger millet, an orphan allotetraploid crop. *DNA Research*.
- [He et al., 2011] He, G., Elling, A. A., and Deng, X. W. (2011). The epigenome and plant development. *Annual review of plant biology*, 62:411–435.
- [Heather and Chain, 2016] Heather, J. M. and Chain, B. (2016). The sequence of sequencers: the history of sequencing dna. *Genomics*, 107(1):1–8.
- [Hebestreit et al., 2013] Hebestreit, K., Dugas, M., and Klein, H.-U. (2013). Detection of significantly differentially methylated regions in targeted bisulfite sequencing data. *Bioinformatics*, 29(13):1647–1653.
- [Hesse et al., 2015] Hesse, N., Schröder, C., and Rahmann, S. (2015). An optimization approach to detect differentially methylated regions from whole genome bisulfite sequencing data. *PeerJ PrePrints*.
- [Holmes et al., 2014] Holmes, E. E., Jung, M., Meller, S., Leisse, A., Sailer, V., Zech, J., Mengdehl, M., Garbe, L.-A., Uhl, B., Kristiansen, G., et al. (2014). Performance evaluation of kits for bisulfite-conversion of dna from tissues, cell lines, ffpe tissues, aspirates, lavages, effusions, plasma, serum, and urine. *PloS one*, 9(4):e93933.
- [Initiative et al., 2000] Initiative, A. G. et al. (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *nature*, 408(6814):796.

- [Istace et al., 2017] Istace, B., Friedrich, A., d’Agata, L., Faye, S., Payen, E., Beluche, O., Caradec, C., Davidas, S., Cruaud, C., Liti, G., et al. (2017). de novo assembly and population genomic survey of natural yeast isolates with the oxford nanopore minion sequencer. *Gigascience*, 6(2):1–13.
- [Jaffe et al., 2012] Jaffe, A. E., Murakami, P., Lee, H., Leek, J. T., Fallin, M. D., Feinberg, A. P., and Irizarry, R. A. (2012). Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. *International journal of epidemiology*, 41(1):200–209.
- [Jaillon et al., 2007] Jaillon, O., Aury, J.-M., Noel, B., Policriti, A., Clepet, C., Casagrande, A., Choisne, N., Aubourg, S., Vitulo, N., Jubin, C., et al. (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *nature*, 449(7161):463.
- [Jaramillo-Correa et al., 2010] Jaramillo-Correa, J. P., Verdú, M., and González-Martínez, S. C. (2010). The contribution of recombination to heterozygosity differs among plant evolutionary lineages and life-forms. *BMC evolutionary biology*, 10(1):22.
- [Jiao et al., 2017] Jiao, Y., Peluso, P., Shi, J., Liang, T., Stitzer, M. C., Wang, B., Campbell, M. S., Stein, J. C., Wei, X., Chin, C.-S., et al. (2017). Improved maize reference genome with single-molecule technologies. *Nature*, 546(7659):524.
- [Jones et al., 2014] Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., et al. (2014). Interproscan 5: genome-scale protein function classification. *Bioinformatics*, 30(9):1236–1240.
- [Judge et al., 2015] Judge, K., Harris, S. R., Reuter, S., Parkhill, J., and Peacock, S. J. (2015). Early insights into the potential of the oxford nanopore minion for the detection of antimicrobial resistance genes. *Journal of Antimicrobial Chemotherapy*, 70(10):2775–2778.
- [Kaplan and Dekker, 2013] Kaplan, N. and Dekker, J. (2013). High-throughput genome scaffolding from in vivo dna interaction frequency. *Nature biotechnology*, 31(12):1143.
- [Kapourani, 2016] Kapourani, C.-A. (2016). Bprmeth: Higher order methylation features for clustering and prediction in epigenomic studies.
- [Kawakatsu et al., 2016] Kawakatsu, T., Stuart, T., Valdes, M., Breakfield, N., Schmitz, R. J., Nery, J. R., Urich, M. A., Han, X., Lister, R., Benfey, P. N., et al. (2016). Unique cell-type-specific patterns of dna methylation in the root meristem. *Nature plants*, 2(5):16058.
- [Kent, 2002] Kent, W. J. (2002). Blat – the blast-like alignment tool. *Genome research*, 12(4):656–664.
- [Khan et al., 2012] Khan, M. A., Han, Y., Zhao, Y. F., Troggio, M., and Korban, S. S. (2012). A multi-population consensus genetic map reveals inconsistent marker order among maps likely attributed to structural variations in the apple genome. *PLoS One*, 7(11):e47864.

- [Koning-Boucoiran et al., 2015] Koning-Boucoiran, C. F., Esselink, G. D., Vukosavljev, M., van't Westende, W. P., Gitonga, V. W., Krens, F. A., Voorrips, R. E., van de Weg, W. E., Schulz, D., Debener, T., et al. (2015). Using rna-seq to assemble a rose transcriptome with more than 13,000 full-length expressed genes and to develop the waghsnp 68k axiom snp array for rose (*rosa l.*). *Frontiers in plant science*, 6:249.
- [Koren et al., 2013] Koren, S., Harhay, G. P., Smith, T. P., Bono, J. L., Harhay, D. M., Mcvey, S. D., Radune, D., Bergman, N. H., and Phillippy, A. M. (2013). Reducing assembly complexity of microbial genomes with single-molecule sequencing. *Genome biology*, 14(9):R101.
- [Koren et al., 2012] Koren, S., Schatz, M. C., Walenz, B. P., Martin, J., Howard, J. T., Ganapathy, G., Wang, Z., Rasko, D. A., McCombie, W. R., Jarvis, E. D., et al. (2012). Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nature biotechnology*, 30(7):693.
- [Koren et al., 2017] Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., and Phillippy, A. M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome research*, 27(5):722–736.
- [Krueger and Andrews, 2011] Krueger, F. and Andrews, S. R. (2011). Bismark: a flexible aligner and methylation caller for bisulfite-seq applications. *bioinformatics*, 27(11):1571–1572.
- [Krzywinski et al., 2009] Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S. J., and Marra, M. A. (2009). Circos: an information aesthetic for comparative genomics. *Genome research*, 19(9):1639–1645.
- [Lagesen et al., 2007] Lagesen, K., Hallin, P., Rødland, E. A., Stærfeldt, H.-H., Rognes, T., and Ussery, D. W. (2007). Rnammer: consistent and rapid annotation of ribosomal rna genes. *Nucleic acids research*, 35(9):3100–3108.
- [Lassois et al., 2016] Lassois, L., Denancé, C., Ravon, E., Guyader, A., Guisnel, R., Hibrand-Saint-Oyant, L., Poncet, C., Lasserre-Zuber, P., Feugey, L., and Durel, C.-E. (2016). Genetic diversity, population structure, parentage analysis, and construction of core collections in the french apple germplasm based on ssr markers. *Plant molecular biology reporter*, 34(4):827–844.
- [Law and Jacobsen, 2010] Law, J. A. and Jacobsen, S. E. (2010). Establishing, maintaining and modifying dna methylation patterns in plants and animals. *Nature Reviews Genetics*, 11(3):204.
- [Lee et al., 2014] Lee, H., Gurtowski, J., Yoo, S., Marcus, S., McCombie, W. R., and Schatz, M. (2014). Error correction and assembly complexity of single molecule sequencing reads. *BioRxiv*, page 006395.
- [Lespinasse et al., 1996] Lespinasse, Y., Bouvier, L., Djulbic, M., and Chevreau, E. (1996). Haploidy in apple and pear. In *Eucarpia Symposium on Fruit Breeding and Genetics 484*, pages 49–54.

- [Li, 2013] Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with bwa-mem. *arXiv preprint arXiv:1303.3997*.
- [Li, 2017] Li, H. (2017). Minimap2: fast pairwise alignment for long dna sequences. *arXiv preprint arXiv:1708.01492*.
- [Li et al., 2016] Li, X., Kui, L., Zhang, J., Xie, Y., Wang, L., Yan, Y., Wang, N., Xu, J., Li, C., Wang, W., et al. (2016). Improved hybrid de novo genome assembly of domesticated apple (*malus x domestica*). *GigaScience*, 5(1):35.
- [Li et al., 2012] Li, Z., Chen, Y., Mu, D., Yuan, J., Shi, Y., Zhang, H., Gan, J., Li, N., Hu, X., Liu, B., et al. (2012). Comparison of the two major classes of assembly algorithms: overlap layout consensus and de-bruijn-graph. *Briefings in functional genomics*, 11(1):25–37.
- [Lister et al., 2008] Lister, R., O’Malley, R. C., Tonti-Filippini, J., Gregory, B. D., Berry, C. C., Millar, A. H., and Ecker, J. R. (2008). Highly integrated single-base resolution maps of the epigenome in arabidopsis. *Cell*, 133(3):523–536.
- [Liu et al., 2015] Liu, R., How-Kit, A., Stammitti, L., Teyssier, E., Rolin, D., Mortain-Bertrand, A., Halle, S., Liu, M., Kong, J., Wu, C., et al. (2015). A demeter-like dna demethylase governs tomato fruit ripening. *Proceedings of the National Academy of Sciences*, 112(34):10804–10809.
- [Love et al., 2014] Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15(12):550.
- [Lowe and Chan, 2016] Lowe, T. M. and Chan, P. P. (2016). trnascan-se on-line: integrating search and context for analysis of transfer rna genes. *Nucleic acids research*, 44(W1):W54–W57.
- [Luo et al., 2012a] Luo, C., Tsementzi, D., Kyrpides, N., Read, T., and Konstantinidis, K. T. (2012a). Direct comparisons of illumina vs. roche 454 sequencing technologies on the same microbial community dna sample. *PloS one*, 7(2):e30087.
- [Luo et al., 2012b] Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q., Liu, Y., et al. (2012b). Soapdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience*, 1(1):18.
- [Lyons et al., 2008] Lyons, E., Pedersen, B., Kane, J., and Freeling, M. (2008). The value of nonmodel genomes and an example using synmap within coge to dissect the hexaploidy that predates the rosids. *Tropical Plant Biology*, 1(3-4):181–190.
- [Mahesh et al., 2016] Mahesh, H., Shirke, M. D., Singh, S., Rajamani, A., Hittalmani, S., Wang, G.-L., and Gowda, M. (2016). Indica rice genome assembly, annotation and mining of blast disease resistance genes. *BMC genomics*, 17(1):242.

- [Manning et al., 2006] Manning, K., Tör, M., Poole, M., Hong, Y., Thompson, A. J., King, G. J., Giovannoni, J. J., and Seymour, G. B. (2006). A naturally occurring epigenetic mutation in a gene encoding an sbp-box transcription factor inhibits tomato fruit ripening. *Nature genetics*, 38(8):948.
- [Martin et al., 2016] Martin, G., Baurens, F.-C., Droc, G., Rouard, M., Cenci, A., Kilian, A., Hastie, A., Doležel, J., Aury, J.-M., Alberti, A., et al. (2016). Improvement of the banana “*musa acuminata*” reference sequence using ngs data and semi-automated bioinformatics methods. *BMC genomics*, 17(1):243.
- [Mashayekhi and Ronaghi, 2007] Mashayekhi, F. and Ronaghi, M. (2007). Analysis of read length limiting factors in pyrosequencing chemistry. *Analytical biochemistry*, 363(2):275–287.
- [Matzke et al., 2015] Matzke, M. A., Kanno, T., and Matzke, A. J. (2015). Rna-directed dna methylation: the evolution of a complex epigenetic pathway in flowering plants. *Annual review of plant biology*, 66:243–267.
- [Matzke and Mosher, 2014] Matzke, M. A. and Mosher, R. A. (2014). Rna-directed dna methylation: an epigenetic pathway of increasing complexity. *Nature Reviews Genetics*, 15(6):394–408.
- [Meyers and Levin, 2006] Meyers, L. A. and Levin, D. A. (2006). On the abundance of polyploids in flowering plants. *Evolution*, 60(6):1198–1206.
- [Miller et al., 2010] Miller, J. R., Koren, S., and Sutton, G. (2010). Assembly algorithms for next-generation sequencing data. *Genomics*, 95(6):315–327.
- [Myers et al., 2000] Myers, E. W., Sutton, G. G., Delcher, A. L., Dew, I. M., Fasulo, D. P., Flanigan, M. J., Kravitz, S. A., Mobarry, C. M., Reinert, K. H., Remington, K. A., et al. (2000). A whole-genome assembly of drosophila. *Science*, 287(5461):2196–2204.
- [Namiki et al., 2012] Namiki, T., Hachiya, T., Tanaka, H., and Sakakibara, Y. (2012). Metavelvet: an extension of velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic acids research*, 40(20):e155–e155.
- [Nawrocki et al., 2014] Nawrocki, E. P., Burge, S. W., Bateman, A., Daub, J., Eberhardt, R. Y., Eddy, S. R., Floden, E. W., Gardner, P. P., Jones, T. A., Tate, J., et al. (2014). Rfam 12.0: updates to the rna families database. *Nucleic acids research*, 43(D1):D130–D137.
- [Nobile et al., 2011] Nobile, P. M., Wattebled, F., Quecini, V., Girardi, C. L., Lormeau, M., and Laurens, F. (2011). Identification of a novel  $\alpha$ -l-arabinofuranosidase gene associated with mealiness in apple. *Journal of experimental botany*, 62(12):4309–4321.
- [O’Connell et al., 2015] O’Connell, J., Schulz-Trieglaff, O., Carlson, E., Hims, M. M., Gormley, N. A., and Cox, A. J. (2015). Nxtrim: optimized trimming of illumina mate pair reads. *Bioinformatics*, 31(12):2035–2037.



- [Ong-Abdullah et al., 2015] Ong-Abdullah, M., Ordway, J. M., Jiang, N., Ooi, S.-E., Kok, S.-Y., Sarpan, N., Azimi, N., Hashim, A. T., Ishak, Z., Rosli, S. K., et al. (2015). Loss of karma transposon methylation underlies the mantled somaclonal variant of oil palm. *Nature*, 525(7570):533.
- [Park et al., 2014] Park, Y., Figueroa, M. E., Rozek, L. S., and Sartor, M. A. (2014). MethySig: a whole genome dna methylation analysis pipeline. *Bioinformatics*, 30(17):2414–2422.
- [Persson and Tegenfeldt, 2010] Persson, F. and Tegenfeldt, J. O. (2010). Dna in nanochannels directly visualizing genomic information. *Chemical Society Reviews*, 39(3):985–999.
- [Pevzner et al., 2001] Pevzner, P. A., Tang, H., and Waterman, M. S. (2001). An eulerian path approach to dna fragment assembly. *Proceedings of the National Academy of Sciences*, 98(17):9748–9753.
- [Pootakham et al., 2017] Pootakham, W., Sonthirod, C., Naktang, C., Ruang-Areerate, P., Yoocha, T., Sangsrakru, D., Theerawattanasuk, K., Rattanawong, R., Lekawipat, N., and Tangphatsornruang, S. (2017). De novo hybrid assembly of the rubber tree genome reveals evidence of paleotetraploidy in hevea species. *Scientific reports*, 7:41457.
- [Putnam et al., 2016] Putnam, N. H., O’Connell, B. L., Stites, J. C., Rice, B. J., Blanchette, M., Calef, R., Troll, C. J., Fields, A., Hartley, P. D., Sugnet, C. W., et al. (2016). Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome research*, 26(3):342–350.
- [Ramírez et al., 2014] Ramírez, F., Dündar, F., Diehl, S., Grüning, B. A., and Manke, T. (2014). deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic acids research*, 42(W1):W187–W191.
- [Redwan et al., 2016] Redwan, R. M., Saidin, A., and Kumar, S. V. (2016). The draft genome of md-2 pineapple using hybrid error correction of long reads. *DNA Research*, 23(5):427–439.
- [Rhoads and Au, 2015] Rhoads, A. and Au, K. F. (2015). Pacbio sequencing and its applications. *Genomics, proteomics & bioinformatics*, 13(5):278–289.
- [Rizk et al., 2013] Rizk, G., Lavenier, D., and Chikhi, R. (2013). Dsk: k-mer counting with very low memory usage. *Bioinformatics*, 29(5):652–653.
- [Robinson et al., 2014] Robinson, M. D., Kahraman, A., Law, C. W., Lindsay, H., Nowicka, M., Weber, L. M., and Zhou, X. (2014). Statistical methods for detecting differentially methylated loci and regions. *Frontiers in genetics*, 5:324.
- [Robinson et al., 2012] Robinson, M. D., Strbenac, D., Stirzaker, C., Statham, A. L., Song, J. Z., Speed, T. P., and Clark, S. J. (2012). Copy-number-aware differential analysis of quantitative dna sequencing data. *Genome research*, pages gr 139055.
- [Ronaghi, 2001] Ronaghi, M. (2001). Pyrosequencing sheds light on dna sequencing. *Genome research*, 11(1):3–11.

- [Ross-Innes et al., 2012] Ross-Innes, C. S., Stark, R., Teschendorff, A. E., Holmes, K. A., Ali, H. R., Dunning, M. J., Brown, G. D., Gojis, O., Ellis, I. O., Green, A. R., et al. (2012). Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature*, 481(7381):389.
- [Roudier et al., 2009] Roudier, F., Teixeira, F. K., and Colot, V. (2009). Chromatin indexing in arabidopsis: an epigenomic tale of tails and more. *Trends in genetics*, 25(11):511–517.
- [Sahlin et al., 2014] Sahlin, K., Vezzi, F., Nystedt, B., Lundeberg, J., and Arvestad, L. (2014). Best-efficient scaffolding of large fragmented assemblies. *BMC bioinformatics*, 15(1):281.
- [Saiki et al., 1988] Saiki, R. K., Chang, C.-A., Levenson, C. H., Warren, T. C., Boehm, C. D., Kazazian Jr, H. H., and Erlich, H. A. (1988). Diagnosis of sickle cell anemia and  $\beta$ -thalassemia with enzymatically amplified dna and nonradioactive allele-specific oligonucleotide probes. *New England Journal of Medicine*, 319(9):537–541.
- [Saint-Oyant et al., 2018] Saint-Oyant, L. H., Ruttink, T., Hamama, L., Kirov, I., Lakhwani, D., Zhou, N., Bourke, P., Daccord, N., Leus, L., Schulz, D., et al. (2018). A high-quality genome sequence of *rosa chinensis* to elucidate ornamental traits. *Nature plants*, page 1.
- [Salmela and Rivals, 2014] Salmela, L. and Rivals, E. (2014). Lordec: accurate and efficient long read error correction. *Bioinformatics*, 30(24):3506–3514.
- [Sanger et al., 1977] Sanger, F., Nicklen, S., and Coulson, A. R. (1977). Dna sequencing with chain-terminating inhibitors. *Proceedings of the national academy of sciences*, 74(12):5463–5467.
- [Schatz et al., 2012] Schatz, M. C., Witkowski, J., and McCombie, W. R. (2012). Current challenges in de novo plant genome sequencing and assembly. *Genome biology*, 13(4):243.
- [Schmid and Deininger, 1975] Schmid, C. W. and Deininger, P. L. (1975). Sequence organization of the human genome. *Cell*, 6(3):345–358.
- [Schmidt et al., 2017] Schmidt, M. H.-W., Vogel, A., Denton, A. K., Istace, B., Wormit, A., van de Geest, H., Bolger, M. E., Alseekh, S., Maß, J., Pfaff, C., et al. (2017). De novo assembly of a new *solanum pennellii* accession using nanopore sequencing. *The Plant Cell*, 29(10):2336–2348.
- [Schmitz et al., 2013] Schmitz, R. J., He, Y., Valdés-López, O., Khan, S. M., Joshi, T., Urich, M. A., Nery, J. R., Diers, B., Xu, D., Stacey, G., et al. (2013). Epigenome-wide inheritance of cytosine methylation variants in a recombinant inbred population. *Genome research*, 23(10):1663–1674.
- [Schmutz et al., 2010] Schmutz, J., Cannon, S. B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., Hyten, D. L., Song, Q., Thelen, J. J., Cheng, J., et al. (2010). Genome sequence of the palaeopolyploid soybean. *nature*, 463(7278):178.





- [Schnable et al., 2009] Schnable, P. S., Ware, D., Fulton, R. S., Stein, J. C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T. A., et al. (2009). The b73 maize genome: complexity, diversity, and dynamics. *science*, 326(5956):1112–1115.
- [Segonne et al., 2014] Segonne, S. M., Bruneau, M., Celton, J.-M., Le Gall, S., Francin-Allami, M., Juchaux, M., Laurens, F., Orsel, M., and Renou, J.-P. (2014). Multiscale investigation of mealiness in apple: an atypical role for a pectin methyltransferase during fruit maturation. *BMC plant biology*, 14(1):375.
- [Shin et al., 2006] Shin, J.-H., Blay, S., McNeney, B., Graham, J., et al. (2006). Ldheatmap: an r function for graphical display of pairwise linkage disequilibria between single nucleotide polymorphisms. *Journal of Statistical Software*, 16(3):1–10.
- [Shulaev et al., 2011] Shulaev, V., Sargent, D. J., Crowhurst, R. N., Mockler, T. C., Folkerts, O., Delcher, A. L., Jaiswal, P., Mockaitis, K., Liston, A., Mane, S. P., et al. (2011). The genome of woodland strawberry (*fragaria vesca*). *Nature genetics*, 43(2):109.
- [Simão et al., 2015] Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M. (2015). Busco: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, 31(19):3210–3212.
- [Staden, 1979] Staden, R. (1979). A strategy of dna sequencing employing computer programs. *Nucleic acids research*, 6(7):2601–2610.
- [Stockwell et al., 2014] Stockwell, P. A., Chatterjee, A., Rodger, E. J., and Morison, I. M. (2014). Dmap: differential methylation analysis package for rrbs and wgbs data. *Bioinformatics*, 30(13):1814–1822.
- [Sun et al., 2014] Sun, D., Xi, Y., Rodriguez, B., Park, H. J., Tong, P., Meong, M., Goodell, M. A., and Li, W. (2014). Moabs: model based analysis of bisulfite sequencing data. *Genome biology*, 15(2):R38.
- [Takata et al., 2007] Takata, M., Kiyohara, A., Takasu, A., Kishima, Y., Ohtsubo, H., and Sano, Y. (2007). Rice transposable elements are characterized by various methylation environments in the genome. *BMC genomics*, 8(1):469.
- [Telias et al., 2011] Telias, A., Lin-Wang, K., Stevenson, D. E., Cooney, J. M., Hellens, R. P., Allan, A. C., Hoover, E. E., and Bradeen, J. M. (2011). Apple skin patterning is associated with differential expression of myb10. *BMC Plant Biology*, 11(1):93.
- [Thieme et al., 2017] Thieme, M., Lanciano, S., Balzergue, S., Daccord, N., Mirouze, M., and Bucher, E. (2017). Inhibition of rna polymerase ii allows controlled mobilisation of retrotransposons for plant breeding. *Genome biology*, 18(1):134.

- [Trapnell et al., 2012] Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., Pimentel, H., Salzberg, S. L., Rinn, J. L., and Pachter, L. (2012). Differential gene and transcript expression analysis of rna-seq experiments with tophat and cufflinks. *Nature protocols*, 7(3):562.
- [Treangen and Salzberg, 2012] Treangen, T. J. and Salzberg, S. L. (2012). Repetitive dna and next-generation sequencing: computational challenges and solutions. *Nature Reviews Genetics*, 13(1):36.
- [VanBuren et al., 2015] VanBuren, R., Bryant, D., Edger, P. P., Tang, H., Burgess, D., Challabathula, D., Spittle, K., Hall, R., Gu, J., Lyons, E., et al. (2015). Single-molecule sequencing of the desiccation-tolerant grass *oropetium thomaeum*. *Nature*, 527(7579):508.
- [Vanyushin, 2006] Vanyushin, B. (2006). Dna methylation in plants. In *DNA methylation: basic mechanisms*, pages 67–122. Springer.
- [Veeckman et al., 2016] Veeckman, E., Ruttink, T., and Vandepoele, K. (2016). Are we there yet? reliably estimating the completeness of plant genome sequences. *The Plant Cell*, 28(8):1759–1768.
- [Velasco et al., 2010] Velasco, R., Zharkikh, A., Affourtit, J., Dhingra, A., Cestaro, A., Kalyanaraman, A., Fontana, P., Bhatnagar, S. K., Troggio, M., Pruss, D., et al. (2010). The genome of the domesticated apple (*malus × domestica* borkh.). *Nature genetics*, 42(10):833.
- [Verde et al., 2013] Verde, I., Abbott, A. G., Scalabrin, S., Jung, S., Shu, S., Marroni, F., Zhebentyayeva, T., Dettori, M. T., Grimwood, J., Cattonaro, F., et al. (2013). The high-quality draft genome of peach (*prunus persica*) identifies unique patterns of genetic diversity, domestication and genome evolution. *Nature genetics*, 45(5):487.
- [Walker et al., 2014] Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C. A., Zeng, Q., Wortman, J., Young, S. K., et al. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PloS one*, 9(11):e112963.
- [Wang et al., 2012] Wang, D., Yan, L., Hu, Q., Sucheston, L. E., Higgins, M. J., Ambrosone, C. B., Johnson, C. S., Smiraglia, D. J., and Liu, S. (2012). Ima: an r package for high-throughput analysis of illumina’s 450k infinium methylation data. *Bioinformatics*, 28(5):729–730.
- [Warden et al., 2013] Warden, C. D., Lee, H., Tompkins, J. D., Li, X., Wang, C., Riggs, A. D., Yu, H., Jove, R., and Yuan, Y.-C. (2013). Cohcap: an integrative genomic pipeline for single-nucleotide resolution dna methylation analysis. *Nucleic acids research*, 41(11):e117–e117.
- [Waterland and Jirtle, 2003] Waterland, R. A. and Jirtle, R. L. (2003). Transposable elements: targets for early nutritional effects on epigenetic gene regulation. *Molecular and cellular biology*, 23(15):5293–5300.
- [Wickham, 2016] Wickham, H. (2016). *ggplot2: elegant graphics for data analysis*. Springer.

- [Wu et al., 2013] Wu, J., Wang, Z., Shi, Z., Zhang, S., Ming, R., Zhu, S., Khan, M. A., Tao, S., Korban, S. S., Wang, H., et al. (2013). The genome of the pear (*pyrus bretschneideri* rehder). *Genome research*, 23(2):396–408.
- [Wu and Watanabe, 2005] Wu, T. D. and Watanabe, C. K. (2005). Gmap: a genomic mapping and alignment program for mrna and est sequences. *Bioinformatics*, 21(9):1859–1875.
- [Xi and Li, 2009] Xi, Y. and Li, W. (2009). Bsmapp: whole genome bisulfite sequence mapping program. *BMC bioinformatics*, 10(1):232.
- [Xiang et al., 2016] Xiang, Y., Huang, C.-H., Hu, Y., Wen, J., Li, S., Yi, T., Chen, H., Xiang, J., and Ma, H. (2016). Evolution of rosaceae fruit types based on nuclear phylogeny in the context of geological times and genome duplication. *Molecular biology and evolution*, 34(2):262–281.
- [Xie et al., 2014] Xie, Y., Wu, G., Tang, J., Luo, R., Patterson, J., Liu, S., Huang, W., He, G., Gu, S., Li, S., et al. (2014). Soapdenovo-trans: de novo transcriptome assembly with short rna-seq reads. *Bioinformatics*, 30(12):1660–1666.
- [Xu et al., 2012] Xu, H., Luo, X., Qian, J., Pang, X., Song, J., Qian, G., Chen, J., and Chen, S. (2012). Fastuniq: a fast de novo duplicates removal tool for paired short reads. *PLoS one*, 7(12):e52249.
- [Ye et al., 2016] Ye, C., Hill, C. M., Wu, S., Ruan, J., and Ma, Z. S. (2016). Dbg2olc: efficient assembly of large genomes using long erroneous reads of the third generation sequencing technologies. *Scientific reports*, 6:31900.
- [Ye and Ma, 2016] Ye, C. and Ma, Z. S. (2016). Sparc: a sparsity-based consensus algorithm for long erroneous sequencing reads. *PeerJ*, 4:e2016.
- [Yu et al., 2002] Yu, J., Hu, S., Wang, J., Wong, G. K.-S., Li, S., Liu, B., Deng, Y., Dai, L., Zhou, Y., Zhang, X., et al. (2002). A draft sequence of the rice genome (*oryza sativa* l. ssp. *indica*). *science*, 296(5565):79–92.
- [Zapata et al., 2016] Zapata, L., Ding, J., Willing, E.-M., Hartwig, B., Bezdán, D., Jiao, W.-B., Patel, V., James, G. V., Koornneef, M., Ossowski, S., et al. (2016). Chromosome-level assembly of *Arabidopsis thaliana* Ler reveals the extent of translocation and inversion polymorphisms. *Proceedings of the National Academy of Sciences*, 113(28):E4052–E4060.
- [Zerbino and Birney, 2008] Zerbino, D. R. and Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de bruijn graphs. *Genome research*, 18(5):821–829.
- [Zhang et al., 2015] Zhang, G., Tian, Y., Zhang, J., Shu, L., Yang, S., Wang, W., Sheng, J., Dong, Y., and Chen, W. (2015). Hybrid de novo genome assembly of the chinese herbal plant danshen (*Salvia miltiorrhiza* Bunge). *GigaScience*, 4(1):62.

- [Zhang et al., 2014] Zhang, H., Tan, E., Suzuki, Y., Hirose, Y., Kinoshita, S., Okano, H., Kudoh, J., Shimizu, A., Saito, K., Watabe, S., et al. (2014). Dramatic improvement in genome assembly achieved using doubled-haploid genomes. *Scientific reports*, 4:6780.
- [Zhang et al., 2012] Zhang, Q., Chen, W., Sun, L., Zhao, F., Huang, B., Yang, W., Tao, Y., Wang, J., Yuan, Z., Fan, G., et al. (2012). The genome of prunus mume. *Nature communications*, 3:1318.
- [Zhang et al., 2006] Zhang, X., Yazaki, J., Sundaresan, A., Cokus, S., Chan, S. W.-L., Chen, H., Henderson, I. R., Shinn, P., Pellegrini, M., Jacobsen, S. E., et al. (2006). Genome-wide high-resolution mapping and functional analysis of dna methylation in arabidopsis. *Cell*, 126(6):1189–1201.
- [Zilberman et al., 2007] Zilberman, D., Gehring, M., Tran, R. K., Ballinger, T., and Henikoff, S. (2007). Genome-wide analysis of arabidopsis thaliana dna methylation uncovers an interdependence between methylation and transcription. *Nature genetics*, 39(1):61.
- [Ziller et al., 2015] Ziller, M. J., Hansen, K. D., Meissner, A., and Aryee, M. J. (2015). Coverage recommendations for methylation analysis by whole-genome bisulfite sequencing. *Nature methods*, 12(3):230.
- [Zolan and Pukkila, 1986] Zolan, M. and Pukkila, P. (1986). Inheritance of dna methylation in coprinus cinereus. *Molecular and Cellular Biology*, 6(1):195–200.
- [Zou et al., 2017] Zou, C., Chen, A., Xiao, L., Muller, H. M., Ache, P., Haberer, G., Zhang, M., Jia, W., Deng, P., Huang, R., et al. (2017). A high-quality genome assembly of quinoa provides insights into the molecular basis of salt bladder-based salinity tolerance and the exceptional nutritional value. *Cell research*, 27(11):1327.

# High-quality *de novo* assembly of the apple genome and methylome dynamics of early fruit development

Nicolas Daccord<sup>1,11</sup>, Jean-Marc Celton<sup>1,11</sup>, Gareth Linsmith<sup>2</sup>, Claude Becker<sup>3,10</sup> , Nathalie Choisne<sup>4</sup>, Elio Schijlen<sup>5</sup>, Henri van de Geest<sup>5</sup>, Luca Bianco<sup>2</sup>, Diego Micheletti<sup>2</sup>, Riccardo Velasco<sup>2</sup>, Erica Adele Di Pierro<sup>6</sup>, Jérôme Gouzy<sup>7</sup>, D Jasper G Rees<sup>8</sup>, Philippe Guérif<sup>1</sup>, Hélène Muranty<sup>1</sup>, Charles-Eric Durel<sup>1</sup>, François Laurens<sup>1</sup>, Yves Lespinasse<sup>1</sup>, Sylvain Gaillard<sup>1</sup>, Sébastien Aubourg<sup>1</sup>, Hadi Quesneville<sup>4</sup> , Detlef Weigel<sup>3</sup> , Eric van de Weg<sup>9</sup>, Michela Troglio<sup>2</sup> & Etienne Bucher<sup>1</sup> 

Using the latest sequencing and optical mapping technologies, we have produced a high-quality *de novo* assembly of the apple (*Malus domestica* Borkh.) genome. Repeat sequences, which represented over half of the assembly, provided an unprecedented opportunity to investigate the uncharacterized regions of a tree genome; we identified a new hyper-repetitive retrotransposon sequence that was over-represented in heterochromatic regions and estimated that a major burst of different transposable elements (TEs) occurred 21 million years ago. Notably, the timing of this TE burst coincided with the uplift of the Tian Shan mountains, which is thought to be the center of the location where the apple originated, suggesting that TEs and associated processes may have contributed to the diversification of the apple ancestor and possibly to its divergence from pear. Finally, genome-wide DNA methylation data suggest that epigenetic marks may contribute to agronomically relevant aspects, such as apple fruit development.

Accurate sequence information, genome assemblies and annotations are the foundation for genetic and genome-wide studies. The major factors that limit *de novo* genome assembly are heterozygosity and repetitive sequences, such as TEs, which are often collapsed to single copies in draft genomes<sup>1</sup>. In recent years, however, evidence supporting the importance of TEs in genome evolution, genome structure, regulation of gene expression and epigenetics has been mounting<sup>2–5</sup>. The characterization of sequences and the distribution of TEs within a genome is, therefore, of great importance.

Until now, the study of epigenetically controlled characteristics in perennial plants has been hampered by the draft status of their genome sequences. In the case of apple, a draft was produced<sup>6</sup> but remained incomplete with inaccurate contig positions<sup>7</sup>; this hindered its utility for genetic and epigenetic studies. *De novo* sequencing and assembly of a new genome for apple, using technologies of the third generation, had thus become a necessity.

In the last few years, single-molecule sequencing and optical-mapping technologies have emerged<sup>8</sup>, which are well suited for assembling genomic regions that contain long repetitive elements. Recently, several high-quality genome assemblies have been published using one or both technologies<sup>9–14</sup>. The use of long-read sequencing technologies

may also tackle potential assembly issues that are related to the presence of highly similar sequences resulting from whole-genome duplication events that frequently occurred in angiosperm genomes<sup>15</sup>.

In addition to DNA sequence modifications, it has been shown that epigenetic variations contribute to genome accessibility, functionality and structure<sup>16,17</sup>. Several studies have demonstrated that local DNA methylation variants, which are represented by differential cytosine methylation at particular loci, can have major effects on the transcription of nearby genes and can be inherited over generations<sup>18–20</sup>.

Apple, like most other fruit tree crops, is propagated by grafting onto rootstocks, which over time can allow the acquisition and propagation of epimutations, via variation in DNA methylation states that can influence various phenotypes, such as fruit color<sup>21,22</sup>. Thus, knowledge of the epigenetic landscape of apple cultivars could provide new tools to study somatic variants, leading to the development of epigenetic markers for marker-assisted selection.

To produce a high-quality apple reference genome and methylome, we generated a *de novo* assembly of a ‘Golden Delicious’ doubled-haploid tree (GDDH13) composed of 280 assembled scaffolds and arranged into 17 pseudomolecules, which represent the 17 chromosomes of apple. This assembly resulted from a combination of short

<sup>1</sup>Institut de Recherche en Horticulture et Semences (IRHS), Université d'Angers, INRA, AGROCAMPUS-Ouest, SFR4207 QUASAV, Université Bretagne Loire, Angers, France. <sup>2</sup>Research and Innovation Center, Department of Genomics and Biology of Fruit Crops, Fondazione E Mach di San Michele all'Adige, Italy. <sup>3</sup>Department of Molecular Biology, Max Planck Institute for Developmental Biology, Tübingen, Germany. <sup>4</sup>UR1164 URGI, Research Unit in Genomics-Info, INRA, Université Paris-Saclay, Versailles, France. <sup>5</sup>Wageningen UR–Bioscience, Wageningen, the Netherlands. <sup>6</sup>Department of Biosciences, University of Milan, Milan, Italy. <sup>7</sup>LIPM, Université de Toulouse, INRA, CNRS, Castanet-Tolosan, France. <sup>8</sup>Agricultural Research Council, Biotechnology Platform, Onderstepoort, Pretoria, South Africa. <sup>9</sup>Wageningen UR–Plant Breeding, Wageningen, the Netherlands. <sup>10</sup>Present address: Gregor Mendel Institute of Molecular Plant Biology, Austrian Academy of Sciences, Vienna Biocenter (VBC), Vienna, Austria. <sup>11</sup>These authors contributed equally to this work. Correspondence should be addressed to E.B. ([etienne.bucher@inra.fr](mailto:etienne.bucher@inra.fr)).

Received 17 October 2016; accepted 3 May 2017; published online 5 June 2017; doi:10.1038/ng.3886

(Illumina) and long sequencing reads (PacBio), along with scaffolding based on optical maps (BioNano) and a high-density integrated genetic linkage map<sup>23</sup>. This chromosome-scale assembly was complemented by a detailed *de novo* annotation of genes based on RNA sequencing (RNA-seq) data, TE annotation and small RNA alignments.

To understand the potential role of epigenetic marks on fruit development, we constructed genome-wide DNA methylation maps that compared different tissues and two isogenic apple lines that produce large or small fruits. This led to the identification of differential DNA methylation patterns that are associated with genes involved in fruit development.

This work provides a solid foundation for future genetic and epigenomic studies in apple. Furthermore, our TE annotation provides novel insights into the evolutionary history of apple and may contribute to explaining its divergence from pear.

## RESULTS

### Genome sequencing, assembly and scaffolding

The doubled-haploid Golden Delicious line (GDDH13, also coded X9273) used in this study is the result of breeding efforts that were initiated at INRA in 1963 (ref. 24) (Supplementary Fig. 1 and Online Methods). Homozygosity of this line was confirmed with microsatellite markers that are distributed along the apple genome (data not shown) and by observation of the k-mer spectrum of Illumina reads derived from GDDH13 (Fig. 1a and Supplementary Note).

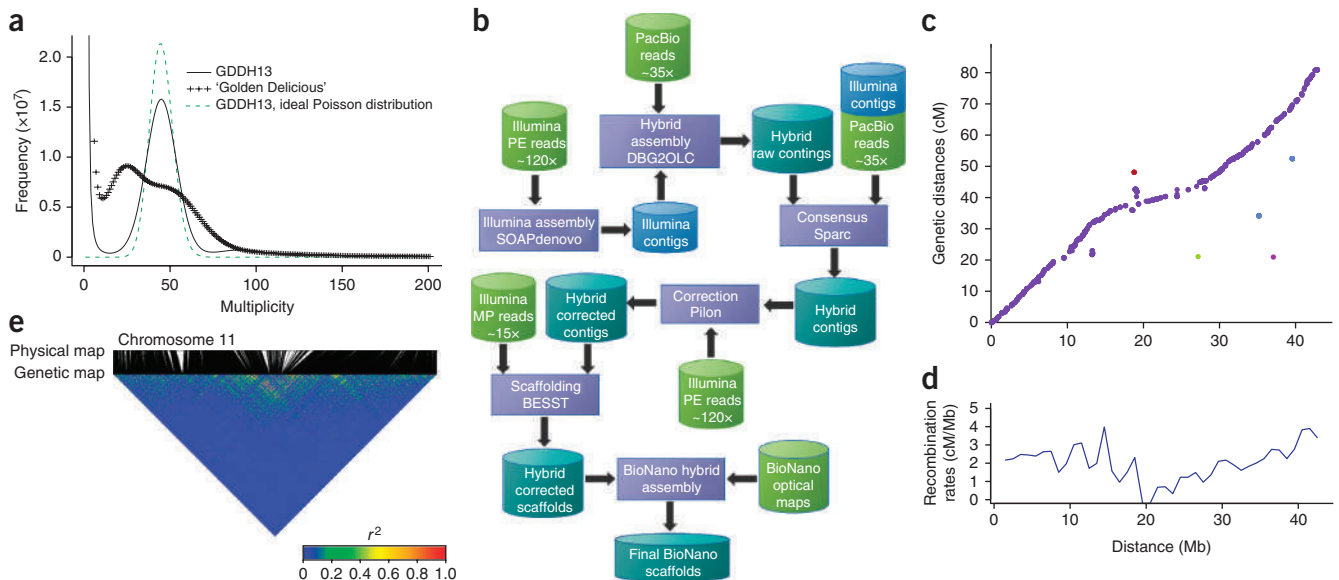
To perform *de novo* assembly of the GDDH13 genome, we combined three different technologies: short-read sequencing, long-read sequencing and optical mapping (Fig. 1b). Using DNA from the leaves of GDDH13, we generated ~120-fold coverage of Illumina paired-end

reads (72 Gb), 80-fold coverage of Illumina Nextera mate-pair reads (58 Gb) at three different insert sizes (2, 5 and 10 kb) and ~35-fold coverage of PacBio sequencing data (24 Gb; 2,837,045 subreads with a mean length of 8,474 bp). The Illumina paired-end reads were first assembled using SOAPdenovo<sup>25</sup>, and the resulting contigs were combined with the PacBio reads using the DBG2OLC assembler<sup>26</sup>. This resulted in an assembly that consisted of 2,150 contigs with an N50 of 620 kb (i.e., 50% of the assembly was contained in contigs  $\geq 620$  kb in size) (Supplementary Table 1) and a total length of 625.2 Mb, which were subsequently corrected by using the Illumina paired-end reads (94,896 single-base assembly errors corrected; 1,054,709 insertions (1,466,015 bp) and 123,510 deletions (178,733 bp)) and scaffolded by using Illumina mate-pair reads with BESST (assembly N50 increased from 620 kb to 699 kb).

Next, using a ~600-fold-coverage BioNano optical map, we generated a consensus map that resulted in an assembly of 649.7 Mb. This consensus map was then used for the hybrid assembly with the corrected scaffolds, which, together with single-nucleotide polymorphism (SNP) markers derived from a high-density genetic linkage map<sup>23</sup>, allowed the construction of the 17 pseudochromosomes (Supplementary Table 2 and Supplementary Note). To estimate the genome size, we calculated different k-mer frequency distributions of the Illumina reads. The estimated GDDH13 genome size of 651 Mb was very close to the 649.7-Mb size in the consensus map.

### Assessment of genome quality

We assessed the quality of the assembly by using two independent sources of data. First, we used the SNP markers that were mapped on the previously mentioned integrated genetic linkage map to validate



**Figure 1** Assembly and validation of the GDDH13 doubled-haploid apple genome. **(a)** k-mer (23 bp) spectra of the doubled-haploid GDDH13 and the heterozygous Golden Delicious<sup>33</sup> genomes. The x axis represents k-mer multiplicity, and the y axis represents the number of k-mers with a given multiplicity in the sequencing data. The green dashed line represents the ideal Poisson distribution fitted on the data of GDDH13. **(b)** Overview of the processing pipeline used for the assembly of the GDDH13 genome (see Supplementary Note for details). **(c)** Graphical representation of the location of SNP markers on the physical map (x axis), as compared to their position on the integrated genetic map (y axis), for Chr11 of the GDDH13 genome. Each marker is depicted as a circle on the plot (1,069 data points). The colors depict the chromosomes as follows: red for Chr01, light green for Chr04, pink for Chr08, blue for Chr10 and violet for Chr11. **(d)** Graphical representation of the mean local recombination rates between successive SNP markers along Chr11 (3-Mb sliding window, 1-Mb shift, threshold 4). The x axis represents the physical positions of the means on Chr11, and the y axis indicates the recombination ratio (centiMorgan (cM)/Mb) in each 3-Mb sliding window. **(e)** Heat map of genotypic linkage disequilibrium (LD;  $r^2$ ) in Chr11 in the 'Old Dessert' INRA apple core collection. Shown are the graphical representation of the location of SNPs on the physical map (top) with correspondence to their order in a regular distribution (bottom) of Chr11 (1,461,195 data points). The color bar indicates the level of LD, from high LD (red) and low LD (blue).



**Table 1 Comparison of the GDDH13 genome with previously published assemblies of the apple genome**

|  | GDDH13                     | Li <i>et al.</i> <sup>33</sup> | Velasco <i>et al.</i> <sup>6</sup> |
|--|----------------------------|--------------------------------|------------------------------------|
| Sequenced genome size (Mb)                       | 643.2*                     | 632.4                          | 603.9                              |
| N50 (kb)   | 5,558                      | 112                            | 16                                 |
| Pearson correlation coefficient with genetic map | 0.897                      | NA                             | 0.667                              |
| TE proportion (%)                                | 57.3 (of BioNano assembly) | NA                             | 42.4                               |
| Annotated protein-coding genes                   | 42,140                     | 53,922                         | 63,141                             |
| Complete BUSCOs                                  | 94.9%                      | 51.5%                          | 86.7%                              |
| Fragmented BUSCOs                                | 2.6%                       | 18.8%                          | 5.6%                               |
| Missing BUSCOs                                   | 2.5%                       | 29.7%                          | 7.7%                               |

\*See **Supplementary Table 2**. NA, not available.

scaffold assembly. Of the 15,417 SNP probe sequences, we identified sequence homology in the GDDH13 genome for 14,732 of them. We then assessed their position on the scaffold assemblies by comparing their location on the integrated genetic linkage map. In total 14,117 of the mapped markers (95.8%) were found to be located at their expected positions (**Supplementary Note**). To visualize these data, we plotted the genetic distance against the physical distance of the genetic markers for each chromosome (**Supplementary Fig. 2**); the data for chromosome (Chr) 11 is shown as an example in **Figure 1c**. This analysis showed that there was very little discrepancy between the physical and genetic maps. For comparison, we plotted these markers to the heterozygous apple genome (v 1.0; **Supplementary Fig. 3**). We also plotted the recombination rates in sliding windows of 3 Mb on this chromosome (**Fig. 1d**) and identified a decrease in recombination frequency toward the middle of Chr11.

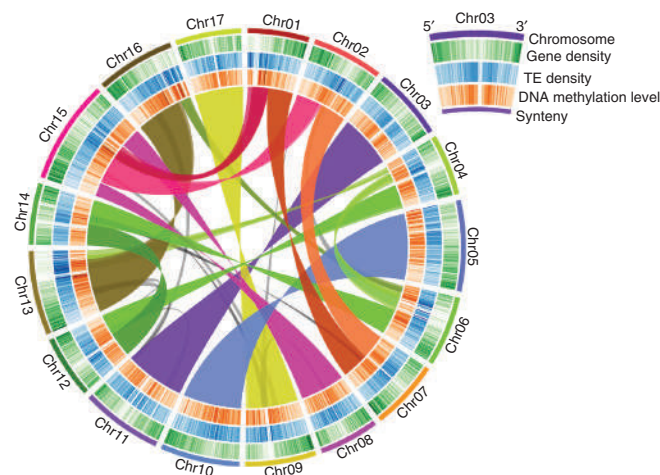
Second, we estimated the level of linkage disequilibrium (LD) using the  $r^2$  parameter between all pairwise SNP comparisons by using marker data that were derived from an apple core collection<sup>27,28</sup>. In the present version of the GDDH13 genome, we did not identify any abrupt jumps in LD, indicating the overall robustness of the assembly (**Fig. 1e** and **Supplementary Fig. 4**). Using previously published genetic data<sup>29</sup>, we generated a haplotype map for GDDH13, which allowed the identification of recombination break-points (**Supplementary Fig. 5**).

Finally, the completeness of the assembly was tested by searching for 248 core eukaryotic genes<sup>30</sup> (CEGs). In total, 237 of 248 CEGs were completely present, and 7 CEGs were partially present, indicating that fewer than 2% of the CEGs could not be detected, which compared very favorably with other assemblies<sup>31</sup>.

### Genome annotation

To obtain a global view of the apple transcriptome, we performed a high-throughput RNA-seq analysis on poly(A)-enriched RNAs from nine libraries that originated from different genotypes and tissues. RNA-seq reads were assembled, and the resulting contigs were mapped to the scaffolds and integrated in the EuGene combiner pipeline<sup>32</sup>. In total, we identified 42,140 protein-coding genes (which represent 23.3% of the genome assembly) and 1,965 non-protein-coding genes (**Supplementary Table 2** and **Supplementary Note**). Evidence of transcription was found for 93% of the annotated genes.

To further evaluate the quality of the annotation, a comparison with annotations of previous apple genome assemblies<sup>6,33</sup> was performed using the BUSCO v2 method, which is based on a benchmark of 1,440 conserved plant genes<sup>34</sup>. The results indicate that our apple genome annotation is the most complete, despite having the lowest number of predicted genes (**Table 1**).



**Figure 2** Synteny and distribution of genomic and epigenomic features of the apple genome. The rings indicate (from outside to inside, as indicated in the inset) chromosomes (Chr), heat maps representing gene density (green), TE density (blue) and DNA methylation levels (orange). A map connecting homologous regions of the apple genome is shown inside the figure. The colored lines link collinearity blocks that represent syntenic regions that were identified by SynMap.

The *de novo* annotated genes were named using the following convention: *MD* (for *Malus domestica*) followed by the chromosome number and gene number on the chromosome (in steps of 100) going from top to bottom according to the linkage map, for example, *MD13G0052100*.

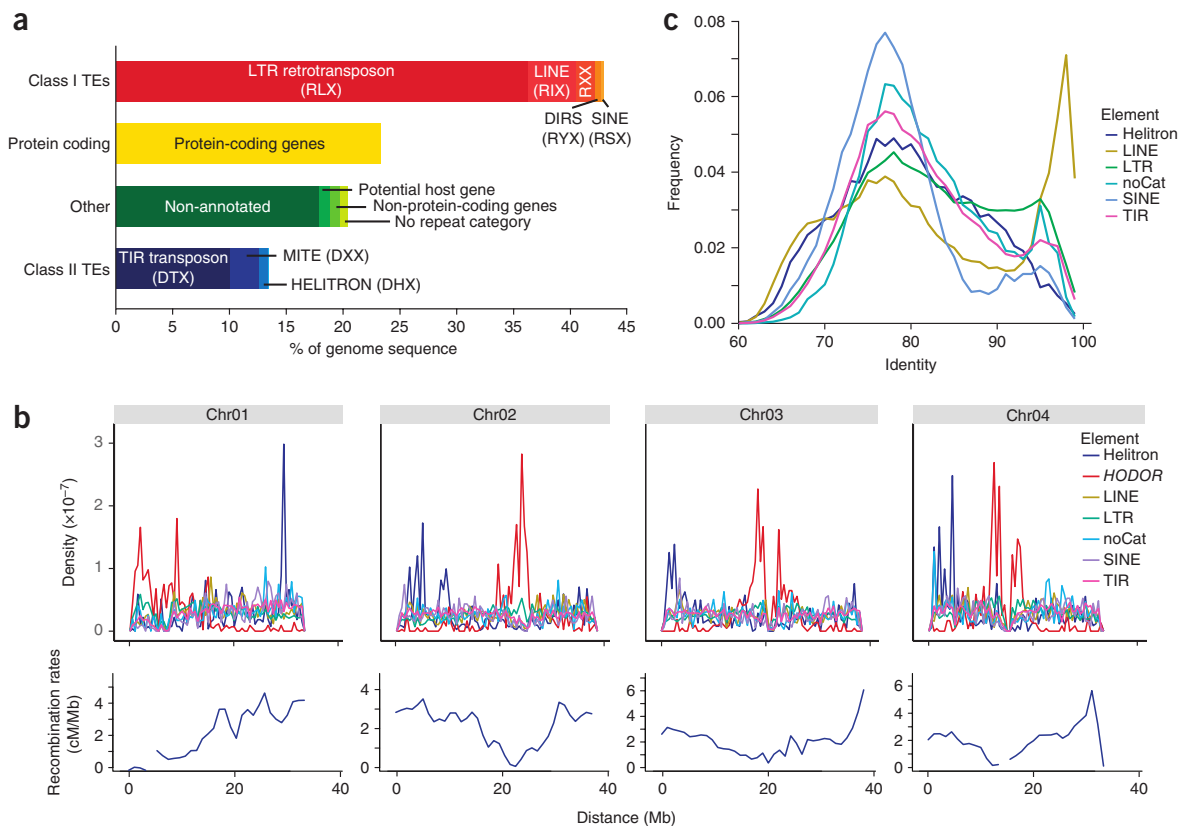
Previously published small RNA (sRNA) data<sup>35</sup> were also mapped to the genome. We found that most 21- and 22-nt-long sRNAs mapped to protein-coding genes, whereas most 24-nt-long sRNAs mapped to TEs. The distribution of 23-nt-long sRNAs was evenly included in both types of genomic features (**Supplementary Fig. 6**).

### Ancestral genome duplication

Intragenomic synteny of GDDH13 was assessed using SynMap (CoGe; <http://www.genomeevolution.org>) and visualized with Circos<sup>36</sup>. Results of this analysis (**Fig. 2**) showed an even clearer genome duplication pattern than has previously been reported<sup>6</sup>. Only very few regions showed no synteny to other parts of the genome (for example, the middle part of Chr04).

### Transposable elements and annotation of repeat sequences

To produce a genome-wide annotation of repetitive sequences, TE consensus sequences (provided by the TE*denovo* detection pipeline<sup>37</sup>) were used to annotate their copies in the whole genome. To refine this annotation, we performed two iterations of the TEannot pipeline. In the GDDH13 genome, TEs represented 372.2 Mb (57.3% of the 649.7 Mb BioNano assembly; **Supplementary Table 2**). Excluding undefined bases (Ns), the TE content of the total nucleotide space in the final annotation was 59.5% of the assembly. The most abundant repeats in this genome are retrotransposons or class I elements (74.8% of TE content, 42.9% of genome assembly), and in particular long terminal repeat retrotransposons (LTR-RTs), which represent 66% of this type of repeat, whereas non-LTR retrotransposons (LINE and SINE) accounted for 7% (**Fig. 3a** and **Supplementary Table 2**). DNA transposons or class II elements (DNA transposons and Helitrons) make up 23% of the TE content (13.4% of the genome assembly; **Fig. 3a** and **Supplementary Table 2**). A complete list of identified TEs, their integrity and copy number can be found in **Supplementary Table 3**.



**Figure 3** Distribution and evolution of transposable elements in the apple genome. **(a)** Percentage of base pairs of the assembled GDDH13 genome that represent genes, pseudo-genes, TEs and non-annotated regions. Retrotransposons (class I) are shown in shades of red, and DNA transposons (class II) are shown in shades of blue. **(b)** Chromosomal density plots of all TE families on Chr01 to Chr04 (top), and the recombination rate for each corresponding chromosome (3-Mb sliding window) (bottom). **(c)** Distribution of sequence identity values between genomic copies and consensus repeats in the GDDH13 assembly (based on 2,198,722 data points). The relative frequencies per percentage of identity of the Helitron, TIR, LTR, LINE, SINE and unclassified TEs (NoCat) are represented in different colors.

We ran the REPET<sup>38</sup> pipeline on the PacBio contigs, which allowed us to identify an additional hyper-repetitive consensus sequence (Genbank entry *KX869746*). This consensus sequence was automatically classified as a 9,716-bp LTR-RT with over 500 full-length copies, and it accounted for 3.6% of the genome assembly (22.3 Mb). We termed this TE consensus sequence *HODOR* (high-copy Golden Delicious repeat). At the chromosomal level, a higher density of *HODOR* copies coincided with particular regions of each chromosome that show reduced recombination levels, whereas the density level of other TEs remained constant or was decreased at these same regions (**Fig. 3b** and **Supplementary Fig. 7**). Even though the retrotransposon consensus sequence has clear 5' and 3' LTRs that are 1.8 kb in size, there are no homologies with typical TE-related sequences encoding a gag protein, a reverse transcriptase or an integrase. However, we found partial sequence similarity to the *Malus domestica* Copia-100 element present in RepBase Update<sup>39</sup>, corresponding to different domains such as gag pre-integrase, RNase H and integrase. These results suggest that *HODOR* is a non-autonomous LTR retrotransposon derivative or LARD (large retrotransposon derivative). We scanned the genome and were able to identify TEs that could contribute to the mobilization of *HODOR* (**Supplementary Table 3** and **Supplementary Note**). Notably, we also found significant (BLASTX *e*-values  $\leq 1 \times 10^{-29}$ ) similarities with sequences encoding three short bacterial proteins of unknown function (**Supplementary Fig. 8a**), and mining of transcriptome data<sup>35</sup> showed *HODOR* to be

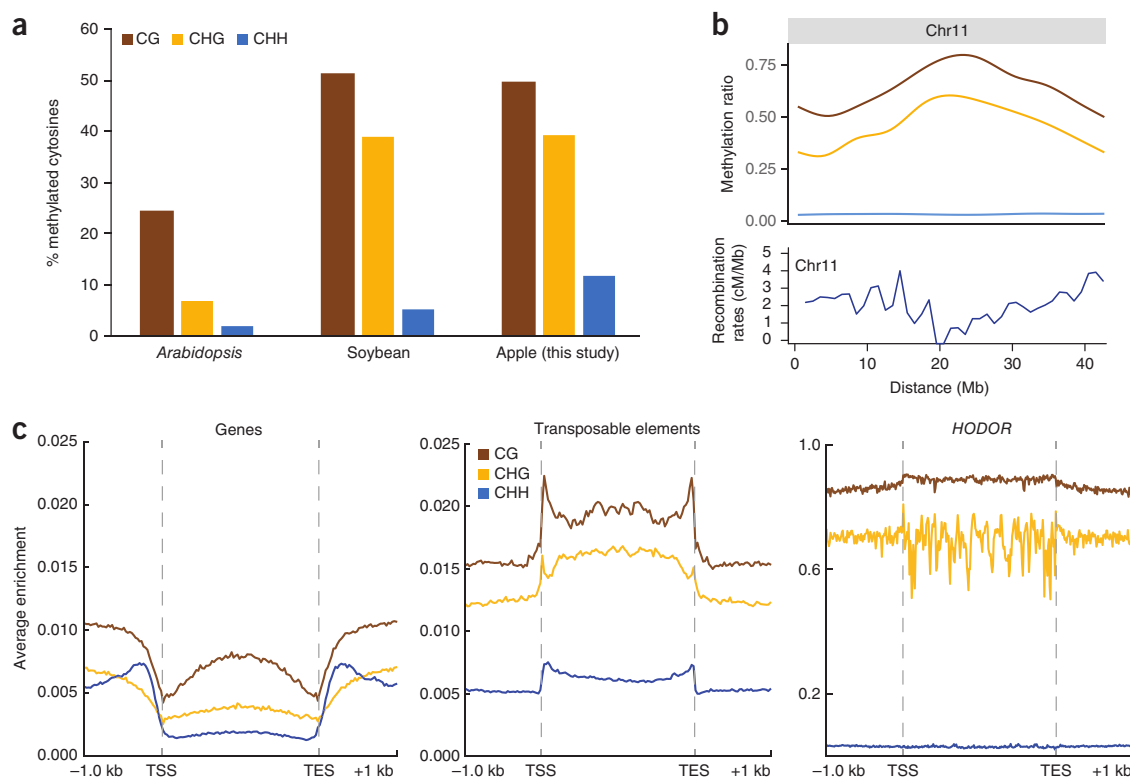
primarily transcribed in the sense and antisense orientations in apple seeds (**Supplementary Fig. 8b**).

To investigate the evolutionary history of TEs in the apple genome, we plotted the distribution of identity values between genomic copies and their consensus sequences (**Fig. 3c**). Distributions for all classes of repeats showed a peak at 77% identity. By considering the mutation rate that has been reported for LTR-RTs in plants ( $1.3 \times 10^{-8}$  base substitutions per site per year<sup>40,41</sup>), we estimated the age of those insertions as described by the International Human Genome Sequencing Consortium<sup>42</sup>. We concluded that the peak at 77% identity corresponded to an insertion age of around 21 million years ago (Mya) (**Fig. 3c**). We also noted a second peak, particularly for LINE elements, at 98% identity that corresponded to a TE burst at  $\sim 1.6$  Mya (**Fig. 3c**).

### The apple methylome

To investigate the apple methylome, we produced genome-wide maps of DNA methylation content at single-base resolution for GDDH13 leaves and young fruits<sup>43,44</sup>.

Globally, in leaves we found DNA methylation levels of 49%, 39% and 12% in the CG, CHG and CHH sequence contexts (where H is adenine, thymine or cytosine), respectively (**Fig. 4a**). DNA methylation was not evenly spread throughout the chromosomes (**Fig. 4b** shows the profile for Chr11; see **Supplementary Fig. 9** for the profiles for all of the chromosomes), and peaks of methylation coincided with recombination cold spots.



**Figure 4** DNA methylation landscape of the GDDH13 genome. **(a)** Percentage of DNA methylation distributions of the three methylation contexts (CG, CHG or CHH) in *Arabidopsis*<sup>44</sup>, soybean<sup>60</sup> and apple. For apple, the percentages were estimated based on the number of cytosines that had a methylation ratio  $\geq 0.75$ . **(b)** Top, chromosomal distribution of the methylation ratios along Chr11. Bottom, the recombination rate plot from **Figure 1d**, for comparison purposes. **(c)** Global distribution of DNA methylation levels at protein-coding genes, TEs and *HODOR*, including a 1-kb window upstream of the TSS and downstream of the transcription end site (TES). In all of the panels, the DNA methylation sequence contexts are color-coded as follows: brown for CG, yellow for CHG and blue for CHH.

As expected<sup>45,46</sup>, there are reduced overall DNA methylation levels in gene sequences, whereas TEs are extensively methylated (**Fig. 4c**). For genes, we identified three major types of DNA methylation patterns. Genes in cluster 1 were characterized by high levels of DNA methylation in the gene body in the CG and CHG contexts, which was concomitant with high DNA methylation in the surrounding regions. Genes in cluster 2 had low CG, and very low CHG and CHH, methylation in the gene itself, yet there were increased levels in the surrounding region. Finally, genes in cluster 3 featured low DNA methylation levels in both the gene body and in the surrounding regions (**Supplementary Fig. 10**). This last cluster contained the largest number of genes (27,179; 64.5% of all genes), showing that in apple, genes are generally depleted for DNA methylation. By mining previously produced large transcriptome data sets for apple<sup>35</sup>, we found that genes covered with very high levels of DNA methylation (cluster 1) showed the lowest expression levels (1.58 median  $\log_2$  value), whereas cluster 2 and cluster 3 genes had higher  $\log_2$  values (3.3 and 2.8, respectively). This result confirmed that the amount of DNA methylation surrounding genes influences their expression level. As one example of TEs, we assessed the DNA methylation levels for *HODOR* and found that *HODOR* was almost completely methylated in the CG (90% methylated) and CHG (65% methylated) contexts but that it had much less methylation in the CHH context (3%) (**Fig. 4c**).

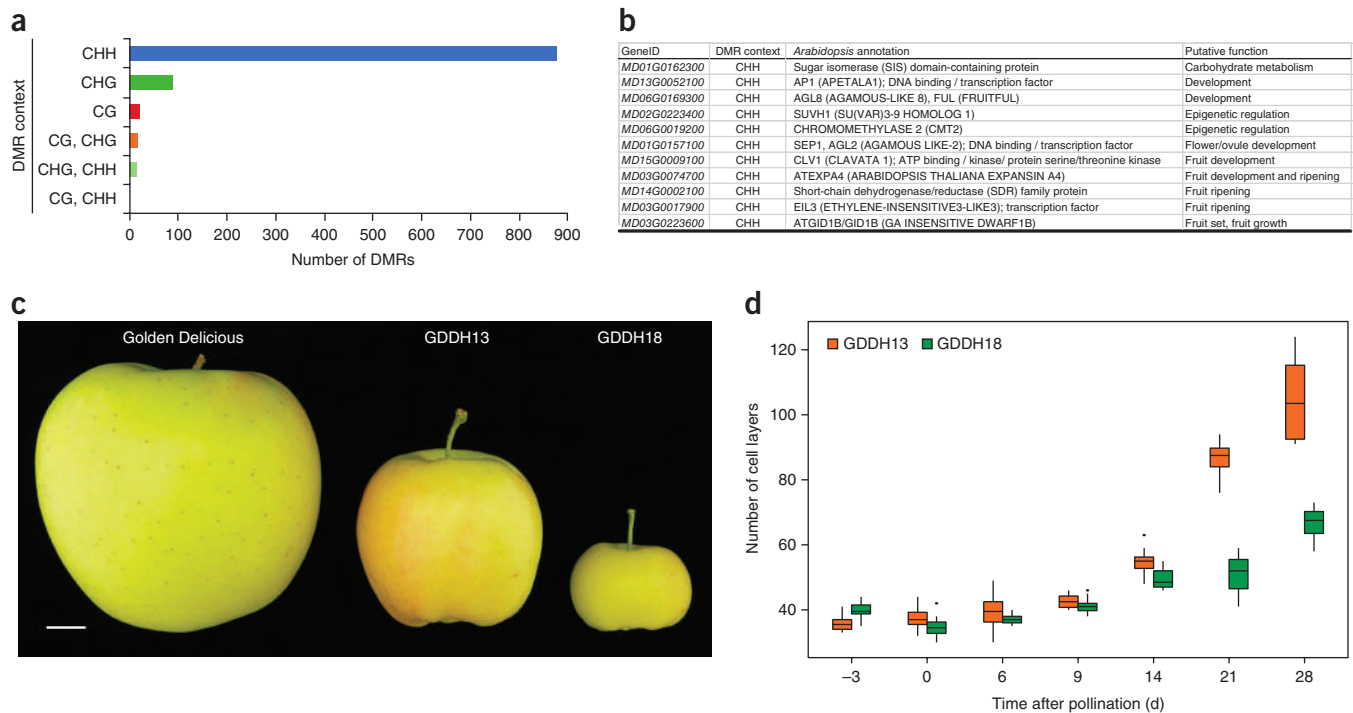
### DNA methylation and fruit development

To assess how DNA methylation contributes to fruit development, we first compared DNA methylation levels between leaves and

fruits. We called differentially methylated regions (DMRs) using a hidden Markov model (HMM)-based approach<sup>47</sup>. In total, we identified 1,017 high-confidence DMRs in all contexts between leaves and fruits, and we observed a very strong bias for DMRs containing methylation changes in the CHH context (875 DMRs; 86.0%) (**Fig. 5a**). We identified 294 genes that contained DMRs in their promoter region—14 DMRs were in the CHG context and showed increased amounts of DNA methylation in leaves, whereas the remaining 280 DMRs were found in the CHH context and showed increased amounts of DNA methylation in fruits. Thus, most methylation differences between leaves and fruits occurred at CHH sites, with a robust increase observed in the developing fruit. Among genes with DMRs that were 2 kb upstream of their transcription start site (TSS), we identified several apple orthologs of *Arabidopsis* genes with important roles in flower and fruit development and in epigenetic regulation (**Fig. 5b**).

Next we wanted to test whether DNA methylation could have a role in the regulation of fruit size. We took advantage of GDDH18, an isogenic line that was obtained from the same haploid that produced GDDH13 (**Supplementary Note**). Whole-genome sequencing showed the presence of 27 homozygous SNPs within genes between the two trees, with nine of these SNPs resulting in amino acid changes (**Supplementary Table 4**). Although the GDDH13 and GDDH18 trees were indistinguishable, the GDDH18 fruits were much smaller (**Fig. 5c**) because of a reduced number of cell layers in the parenchyma (**Fig. 5d**).

To elucidate whether the difference in fruit size could have an epigenetic basis, whole-genome bisulfite sequencing was performed on



**Figure 5** Differentially methylated regions between apple tree leaves and young fruits. **(a)** DMR content in samples of GDDH13 leaves and young fruits (CHH,  $n = 875$  DMRs; CHG,  $n = 88$  DMRs; CG,  $n = 21$  DMRs; CG and CHG,  $n = 17$  DMRs; CHG and CHH,  $n = 14$  DMRs; CG and CHH,  $n = 2$  DMRs). Most of the DMRs (86%) were identified in the CHH context. **(b)** Selection of GDDH13 genes that present a DMR within a region 2 kb upstream of the TSS. The apple gene ID, the methylation context of the DMR, the orthologous *Arabidopsis* gene annotation and the function of the encoded protein are listed. **(c,d)** Representative image comparing the fruit sizes of heterozygous Golden Delicious, GDDH13 and GDDH18 at harvest **(c)** and quantification of the number of cell layers in the parenchyma of GDDH13 (orange) and GDDH18 (green) fruits, as assessed by microscopy ( $n = 12$  data points per box plot) **(d)**. The horizontal line in the box represents the median, the lower and upper hinges correspond to the first and third quartiles, the lower and upper whiskers extend from the hinge to the smallest and largest value (no further than 1.5-fold the inter-quartile range from the hinge), and outlying points are plotted individually. Scale bar, 1 cm.

samples that were collected at 3 d before pollination (or -3 d after pollination (DAP); when fruits have a similar size and number of cell layers) and at 9 DAP (a few days before observing significant phenotypic differences between the fruits). As expected from their common origin, only a limited number of high-confidence DMRs ( $n = 197$ ) could be found between young fruits of GDDH13 and GDDH18 at -3 DAP. Of these, 47 DMRs were located within 2 kb upstream of the TSS of genes. Similarly, we identified a total of 148 high-confidence DMRs between fruits of GDDH13 and GDDH18 at 9 DAP. From this analysis, we found that 53 genes contained DMRs in their promoter region (i.e., within 2 kb upstream of the TSS). At both time points a majority of genes with DMRs showed a decrease in methylation in their promoter region for GDDH18 (**Supplementary Table 5**). Notably, in both comparisons, DMRs in the CG-CHG and CHG contexts were over-represented.

The overlap of DMRs between the two time points analyzed included 22 genes with DMRs in their promoter regions, with most of them ( $n = 17$ ) showing lower methylation in GDDH18 (**Supplementary Table 5**). Several of the 22 genes have orthologs in other species with a role that could explain the observed size difference between the GDDH13 and GDDH18 fruits—including SQUAMOSA PROMOTER-BINDING PROTEIN LIKE 13 (*SPL13*, *MD16G0108400*), 1-AMINO-CYCLOPROPANE-1-CARBOXYLATE SYNTHASE 8 (*ACS8*, *MD15G0127800*) and CYTOCHROME P450 FAMILY 71 SUBFAMILY A POLYPEPTIDE 25 (*CYP71A25*, *MD14G0147300*), which belong to the minority of genes with increased methylation in GDDH18.

## DISCUSSION

As a prerequisite to epigenomic studies in apple, we decided to produce a high-quality reference genome for apple. An advantage for us was the availability of the homozygous GDDH13 doubled-haploid line. Assembling a genome that is both highly heterozygous and recently duplicated into a haploid consensus sequence presents a substantial challenge. This is exemplified by the comparison of our first assembly steps to a recently published report on a heterozygous Golden Delicious apple genome sequence<sup>33</sup>. Following hybrid assembly of PacBio and Illumina reads, Li and colleagues<sup>33</sup> reported a N50 of 112 kb, whereas we obtained a N50 of 620 kb at that same step. These results highlight the power of haploids or doubled haploids in genome sequencing projects<sup>48</sup>, particularly in those for apple, which is not only highly heterozygous but has also undergone a recent whole-genome duplication (ref. 6 and this study). The optical mapping then allowed us to produce scaffolds with a N50 of 5.5 Mb, which, in association with a high-density integrated linkage map, yielded highly contiguous pseudomolecules. In this new apple genome, we followed a newer convention<sup>23</sup> in which the orientation of Chr10 and Chr05 became aligned by the inversion of Chr05. We chose to invert Chr05 because it is the least frequently reported of the two in previous genetic studies on quantitative trait loci (QTL), gene discovery and characterization.

We estimated the genome size of GDDH13 to be 651 Mb (**Supplementary Table 2**), which suggested that the GDDH13 genome may be smaller than that of the heterozygous Golden Delicious line, which was recently estimated to be 701 Mb (ref. 33). Although the GDDH13 tree looks similar to the heterozygous Golden Delicious



counterpart (including tree architecture, flowering time and fruit appearance; **Supplementary Fig. 1**), it is possible that through the consecutive steps of selfing, haploid development and chromosome doubling, some minor parts of the genome might have been lost or re-arranged. Thus, it is possible that some of the genome sequence might be missing in the GDDH13 assembly.

Our gene prediction analysis reduced the estimated number of annotated genes in apple from 63,541 (Genome Database for Rosaceae, see URLs and ref. 6) to 42,140, which is much closer to the 42,812 genes that have been reported for pear<sup>49</sup> and the 45,293 genes that were identified after filtering out overlapping genes from the original apple genome annotation<sup>49</sup> (**Supplementary Note**).

TEs also have an important role in structuring genomes. The in-depth TE annotation we performed showed a major TE burst in apple that we estimated to have happened around 21 Mya. This affected all types of TEs, suggesting that the precursor of the modern apple underwent environmental changes with resulting stresses that led to the activation of these TEs<sup>50</sup>. The observed TE burst corresponds to the Miocene epoch (23 Mya to 5 Mya) and may coincide with two events: the divergence between pear and apple<sup>48</sup> and an uplift event occurring at the Tian Shan mountains<sup>51</sup>, which cover the region where the ancestor of the apple originates from<sup>52</sup>. We hypothesize that these TE bursts, which presumably must have been very different in the predecessor of pear and apple, have contributed to the diversification, and possibly even speciation, of these plants.

Although our analyses using previously reported approaches<sup>53</sup> did not identify any characteristic short centromeric repeat sequence in the apple genome, we can hypothesize the putative localization of centromeres on the GDDH13 chromosomes. We found that the regions in which we observed a decrease in the recombination rate between successive markers of the integrated linkage map coincided with the regions that showed an increase in the estimated level of LD in the core apple collection, as well as an increase in DNA methylation levels. In addition, we identified *HODOR*, the most repetitive consensus sequence in the apple genome, as being over-represented in these same genomic regions. These findings suggest that centromeric regions in the GDDH13 genome may be located within the regions that show an over-representation of *HODOR*. Future studies will show whether *HODOR* has a role in the centromere structure in the apple genome. Blast searches have revealed that the *HODOR* sequence also exists in pear, and because of its origin from potential horizontal gene transfer events, it will be of great interest to investigate when *HODOR* first appeared during the Rosaceae evolution.

The genome-wide distribution of DNA methylation peaked in putative centromeric regions of high LD and high *HODOR* content. As has been observed in *Arabidopsis*<sup>43</sup>, TEs were enriched and genes strongly depleted for DNA methylation. The 10% of genes that possess high levels of DNA methylation (gene body and surrounding region; **Supplementary Fig. 10**), globally showed a very low level of transcription, and these genes may be expressed during very specific developmental stages or tissues. The comparison of the apple leaf and fruit methylomes revealed a noteworthy pattern—the fruit globally had higher CHH DNA methylation levels, which suggested increased activity of the RNA-directed DNA methylation machinery in this organ<sup>54</sup>. Consistent with this observation, it has been shown for *Arabidopsis* that cell-type-specific DNA methylation differences mainly occur at CHH sites<sup>55</sup>. Notably, DNA methylation differences in the CHH context between leaf and fruit tissues occurred next to 294 genes. Several of these were found to be orthologous to genes that are known to be important regulators of flower and fruit development in other species. This suggests that apple fruit development is regulated by epigenetic

processes, which is consistent with data obtained in tomato, demonstrating that DNA methylation is important for fruit ripening<sup>56–58</sup>.

In addition, among the major agronomical traits that contribute to both yield and quality, fruit size is one of the most important for many domesticated crops. Two of the key determinants that are known to alter plant organ size are cell number and cell size<sup>59</sup>. Here we investigated fruit size difference between two isogenic doubled-haploid apple lines. We found that the number of cell layers in the parenchyma of GDDH13 fruits increased more rapidly than those in the parenchyma of the smaller GDDH18 fruits, with significant differences being observed as early as 21 DAP. To identify regulators that contributed to the difference in fruit size between the two doubled-haploid apple lines, we found three genes that potentially contributed to the cell number difference, and these contained DMRs in their promoter regions (**Supplementary Note**).

The identification of potential molecular mechanisms that control cell-division-related processes by DNA methylation provides new insights into the understanding of this important process. However, by comparing the GDDH13 and GDDH18 genomes, we identified nine SNPs that affect protein sequences, and thus we cannot currently exclude a genetic effect.

The high-quality reference apple genome sequence reported here offers unprecedented insights into the genome dynamics of a tree and provides an important basis for future studies, not only in apple but also in other Rosaceae species.

**URLs.** Structural and functional annotations are available through our genome browser: <https://iris.angers.inra.fr/gddh13/>. The Whole-Genome Shotgun project can be found in GenBank under: <https://www.ncbi.nlm.nih.gov/nucleotide/MJAX00000000.1> The REPET package v2.5 used to detect TEs used in this study can be found here: <https://urgi.versailles.inra.fr/Tools/REPET> SynMap- CoGe: <http://www.genomevolution.org> Genome Database for Rosaceae: <http://www.rosaceae.org>.

## METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the [online version of the paper](#).

*Note: Any Supplementary Information and Source Data files are available in the online version of the paper.*

## ACKNOWLEDGMENTS

We thank M. Mirouze and C. Vitte for their support with the TE analysis, and T. Girard and A. Cornille for their insights on apple evolution. We are very grateful to the Horticulture Experimental Unit of the Institut National de la Recherche Agronomique (Pays de la Loire) for taking care of the apple trees used in this study. We thank the IMAC and ANAN platforms from the Structure Fédérative de Recherche 'Qualité et Santé du Végétal' (SFR QUASAV) for their technical support. This research was funded by the EPICENTER ConnecTalent grant of the Pays de la Loire (E.B.) and supported by the Provincia autonoma di Trento (G.L., L.B., D.M., R.V. and M.T.), the EU seventh Framework Programme by the FruitBreedomics project no. 265582: "Integrated approach for increasing breeding efficiency in fruit tree crops (<http://www.fruitbreedomics.com/>)" (F.L., L.B., D.M., R.V. and M.T.), the Max Planck Society (C.B. and D.W.) and the Deutsche Forschungsgemeinschaft (SFB 1101; C.B. and D.W.).

## AUTHOR CONTRIBUTIONS

N.D. and J.-M.C. are joint first authors and contributed equally to the work, and E.B. was the leading investigator. E.B., E.S., J.-M.C., R.V. and H.v.d.G. supported and performed sequencing and genome-mapping experiments; N.D., E.S., J.-M.C., G.L., L.B. and H.v.d.G. performed genome assemblies; J.-M.C. performed field and wet lab work; G.L. and L.B. performed k-mer spectra analysis and genome size estimations; G.L., L.B., D.M., R.V., E.A.D.P., H.M., C.-E.D., F.L., E.v.d.W. and M.T. provided genetic map information and performed quality control experiments;

M.T. and E.A.D.P. made the plots comparing genetic and physical maps and performed recombination rate analyses; H.M. provided LD plots; D.M. and M.T. made the haplotype map plot; N.D., S.A., S.G. and J.G. performed gene annotation; E.B. and N.D. performed genome duplication analysis; N.C., H.Q., E.B. and S.A. annotated and analyzed TEs; C.B. performed bisulfite sequencing; C.B., J.-M.C., N.D., D.W. and E.B. analyzed bisulfite sequencing results; J.-M.C. and D.J.G.R. provided RNA-seq data; Y.L. created the GDDH13 and GDDH18 lines, with support from P.G.; S.G. set up the genome browser; J.-M.C., N.D. and E.B. wrote and edited most of the manuscript; and all authors read and commented on the manuscript.

#### COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Reprints and permissions information is available online at <http://www.nature.com/reprints/index.html>. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



This work is licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

- Veeckman, E., Ruttink, T. & Vandepoele, K. Are we there yet? Reliably estimating the completeness of plant genome sequences. *Plant Cell* **28**, 1759–1768 (2016).
- ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
- Fedoroff, N.V. Transposable elements, epigenetics and genome evolution. *Science* **338**, 758–767 (2012).
- Chénais, B., Caruso, A., Hiard, S. & Casse, N. The impact of transposable elements on eukaryotic genomes: from genome size increase to genetic adaptation to stressful environments. *Gene* **509**, 7–15 (2012).
- Lisch, D. How important are transposons for plant evolution? *Nat. Rev. Genet.* **14**, 49–61 (2013).
- Velasco, R. *et al.* The genome of the domesticated apple (*Malus × domestica* Borkh.). *Nat. Genet.* **42**, 833–839 (2010).
- Khan, M.A., Han, Y., Zhao, Y.F., Troggio, M. & Korban, S.S. A multi-population consensus genetic map reveals inconsistent marker order among maps likely attributed to structural variations in the apple genome. *PLoS One* **7**, e47864 (2012).
- Ansong, W.J. Next-generation DNA sequencing (II): techniques, applications. *Next Generat. Sequenc. & Applic.* **1**, 1–10 (2016).
- Zhang, G. *et al.* Hybrid *de novo* genome assembly of the Chinese herbal plant danshen (*Salvia miltiorrhiza* Bunge). *Gigascience* **4**, 62 (2015).
- VanBuren, R. *et al.* Single-molecule sequencing of the desiccation-tolerant grass *Oropetium thomaeum*. *Nature* **527**, 508–511 (2015).
- Zapata, L. *et al.* Chromosome-level assembly of *Arabidopsis thaliana* Ler reveals the extent of translocation and inversion polymorphisms. *Proc. Natl. Acad. Sci. USA* **113**, E4052–E4060 (2016).
- Redwan, R.M., Saidin, A. & Kumar, S.V. The draft genome of MD-2 pineapple using hybrid error correction of long reads. *DNA Res.* **23**, 427–439 (2016).
- Mahesh, H.B. *et al.* *Indica* rice genome assembly, annotation and mining of blast-disease-resistance genes. *BMC Genomics* **17**, 242 (2016).
- Badouin, H. *et al.* Chaos of rearrangements in the mating-type chromosomes of the anther-smut fungus *Microbotryum lychnidis-dioicae*. *Genetics* **200**, 1275–1284 (2015).
- Cui, L. *et al.* Widespread genome duplications throughout the history of flowering plants. *Genome Res.* **16**, 738–749 (2006).
- Roudier, F., Teixeira, F.K. & Colot, V. Chromatin indexing in *Arabidopsis*: an epigenomic tale of tails and more. *Trends Genet.* **25**, 511–517 (2009).
- He, G., Elling, A.A. & Deng, X.W. The epigenome and plant development. *Annu. Rev. Plant Biol.* **62**, 411–435 (2011).
- Cubas, P., Vincent, C. & Coen, E. An epigenetic mutation responsible for natural variation in floral symmetry. *Nature* **401**, 157–161 (1999).
- Becker, C. *et al.* Spontaneous epigenetic variation in the *Arabidopsis thaliana* methylome. *Nature* **480**, 245–249 (2011).
- Ong-Abdullah, M. *et al.* Loss of *Karma* transposon methylation underlies the mantled somaclonal variant of oil palm. *Nature* **525**, 533–537 (2015).
- El-Sharkawy, I., Liang, D. & Xu, K. Transcriptome analysis of an apple (*Malus × domestica*) yellow fruit somatic mutation identifies a gene network module highly associated with anthocyanin and epigenetic regulation. *J. Exp. Bot.* **66**, 7359–7376 (2015).
- Telias, A. *et al.* Apple skin patterning is associated with differential expression of *MYB10*. *BMC Plant Biol.* **11**, 93 (2011).
- Di Pierro, E.A. *et al.* A high-density, multi-parental SNP genetic map on apple validates a new mapping approach for outcrossing species. *Hortic. Res.* **3**, 16057 (2016).
- Lespinasse, Y., Bouvier, L., Djulbic, M. & Chevreau, E. Haploidy in apple and pear. *Acta Hortic.* **484**, 49–54 (1998).
- Luo, R. *et al.* SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. *Gigascience* **1**, 18 (2012).
- Ye, C., Hill, C.M., Wu, S., Ruan, J. & Ma, Z.S. DBG2OLC: efficient assembly of large genomes using long erroneous reads of the third-generation sequencing technologies. *Sci. Rep.* **6**, 31900 (2016).
- Lassois, L. *et al.* Genetic diversity, population structure, parentage analysis and construction of core collections in the French apple germplasm based on SSR markers. *Plant Mol. Biol. Rep.* **34**, 827–844 (2016).
- Bianco, L. *et al.* Development and validation of the Axiom Apple480K SNP genotyping array. *Plant J.* **86**, 62–74 (2016).
- Falginella, L. *et al.* A major QTL controlling apple skin russetting maps on the linkage group 12 of 'Renetta Grigia di Torriana'. *BMC Plant Biol.* **15**, 150 (2015).
- Parra, G., Bradnam, K. & Korf, I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23**, 1061–1067 (2007).
- Bradnam, K.R. *et al.* Assemblathon 2: evaluating *de novo* methods of genome assembly in three vertebrate species. *Gigascience* **2**, 10 (2013).
- Foissac, S. *et al.* Genome annotation in plants and fungi: EuGene as a model platform. *Curr. Bioinform.* **3**, 87–97 (2008).
- Li, X. *et al.* Improved hybrid *de novo* genome assembly of domesticated apple (*Malus × domestica*). *Gigascience* **5**, 35 (2016).
- Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V. & Zdobnov, E.M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
- Celton, J.M. *et al.* Widespread antisense transcription in apple is correlated with siRNA production and indicates a large potential for transcriptional and/or post-transcriptional control. *New Phytol.* **203**, 287–299 (2014).
- Krzywinski, M. *et al.* CircoS: an information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645 (2009).
- Flutre, T., Duprat, E., Feuillet, C. & Quesneville, H. Considering transposable element diversification in *de novo* annotation approaches. *PLoS One* **6**, e16526 (2011).
- Quesneville, H. *et al.* Combined evidence annotation of transposable elements in genome sequences. *PLoS Comput. Biol.* **1**, 166–175 (2005).
- Jurka, J. *et al.* Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110**, 462–467 (2005).
- Ma, J. & Bennetzen, J.L. Recombination, rearrangement, reshuffling and divergence in a centromeric region of rice. *Proc. Natl. Acad. Sci. USA* **103**, 383–388 (2006).
- Yin, H. *et al.* Genome-wide annotation and comparative analysis of long-terminal-repeat retrotransposons between pear species of *P. bretschneideri* and *P. communis*. *Sci. Rep.* **5**, 17644 (2015).
- Lander, E.S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
- Lister, R. *et al.* Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. *Cell* **133**, 523–536 (2008).
- Cokus, S.J. *et al.* Shotgun bisulphite sequencing of the *Arabidopsis* genome reveals DNA methylation patterning. *Nature* **452**, 215–219 (2008).
- Matzke, M.A. & Mosher, R.A. RNA-directed DNA methylation: an epigenetic pathway of increasing complexity. *Nat. Rev. Genet.* **15**, 394–408 (2014).
- Law, J.A. & Jacobsen, S.E. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nat. Rev. Genet.* **11**, 204–220 (2010).
- Hagmann, J. *et al.* Century-scale methylome stability in a recently diverged *Arabidopsis thaliana* lineage. *PLoS Genet.* **11**, e1004920 (2015).
- Zhang, H. *et al.* Dramatic improvement in genome assembly achieved using doubled-haploid genomes. *Sci. Rep.* **4**, 6780–6785 (2014).
- Wu, J. *et al.* The genome of the pear (*Pyrus bretschneideri* Rehd.). *Genome Res.* **23**, 396–408 (2013).
- Belyayev, A. Bursts of transposable elements as an evolutionary driving force. *J. Evol. Biol.* **27**, 2573–2584 (2014).
- Balukhovskiy, A.N. & Khain, V.E. *Historical Geotectonics—Mesozoic and Cenozoic* (CRC Press, 1997).
- Cornille, A. *et al.* New insight into the history of domesticated apple: secondary contribution of the European wild apple to the genome of cultivated varieties. *PLoS Genet.* **8**, e1002703 (2012).
- Melters, D.P. *et al.* Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. *Genome Biol.* **14**, R10 (2013).
- Matzke, M.A., Kanno, T. & Matzke, A.J.M. RNA-directed DNA methylation: the evolution of a complex epigenetic pathway in flowering plants. *Annu. Rev. Plant Biol.* **66**, 243–267 (2015).
- Kawakatsu, T. *et al.* Unique cell-type-specific patterns of DNA methylation in the root meristem. *Nat. Plants* **2**, 16058 (2016).
- Manning, K. *et al.* A naturally occurring epigenetic mutation in a gene encoding an SBP-box transcription factor inhibits tomato fruit ripening. *Nat. Genet.* **38**, 948–952 (2006).
- Liu, R. *et al.* A DEMETER-like DNA demethylase governs tomato fruit ripening. *Proc. Natl. Acad. Sci. USA* **112**, 10804–10809 (2015).
- Gallusci, P., Hodgman, C., Teyssier, E. & Seymour, G.B. DNA methylation and chromatin regulation during fleshy fruit development and ripening. *Front. Plant Sci.* **7**, 807 (2016).
- Guo, M. & Simmons, C.R. Cell number counts—the *fw2.2* and *CNR* genes and implications for controlling plant fruit and organ size. *Plant Sci.* **181**, 1–7 (2011).
- Schmitz, R.J. *et al.* Epigenome-wide inheritance of cytosine methylation variants in a recombinant inbred population. *Genome Res.* **23**, 1663–1674 (2013).

## ONLINE METHODS

**Genome assembly of GDDH13. Hybrid assembly.** The genome assembly was performed using a combination of sequencing technologies: PacBio RS II reads, Illumina paired-end reads (PE) and Illumina mate-pair reads (MP). First, Illumina PE reads were separately assembled using SOAPdenovo 2.223 (ref. 25). Next, the PacBio reads and Illumina contigs were combined to perform a hybrid assembly using the DBG2OLC pipeline<sup>26</sup>.

**Assembly polishing.** A polishing of the assembly using the Illumina paired-end reads was performed. The 120× Illumina reads were mapped to the contigs using BWA-MEM<sup>61</sup>. This alignment was then used with Pilon 1.17 (ref. 62) to correct the assembly.

**Mate pair scaffolding.** A total of 8.5 Gb of Illumina MP data (approximate sequencing depth = 15×), with an insert size varying between 2 kb and 10 kb, was used to scaffold the assembly. The MP reads were mapped on the corrected contigs using BWA-MEM. The alignments were processed with the BESST<sup>63</sup> software to scaffold the assembly.

**BioNano scaffolding.** A BioNano optical mapping analysis was performed, and data was collected and analyzed with IrisViewer (v2.5). The 397 BioNano maps, with a N50 of 2.649 Mb and a total length of 649.7 Mb, were used in the hybrid assembly step with the scaffolds obtained from the MP scaffolding to assemble the final scaffolds in IrisViewer.

**Scaffold validation and anchoring to the genetic map.** An integrated multiparental genetic linkage map of apple<sup>23</sup> that was composed of 15,417 SNP markers was used to organize and orientate the scaffolds into chromosome-sized sequences. The probe sequences of the 15,417 markers<sup>64</sup> were mapped onto the genome using BWA-MEM. The physical and genetic positions of the mapped markers were used to place and orient the scaffolds and contigs relative to each other. Detailed methodological details describing the assembly processes can be found in the **Supplementary Note**.

**Linkage disequilibrium (LD).** The ‘Old Dessert’ INRA core collection, comprising 278 accessions<sup>27</sup>, was genotyped with the Axiom Apple-480K SNP genotyping array<sup>28</sup>. LD was estimated with the  $r^2$  statistics using the R package snpStats (R package version 1.16.0). Heat maps of pairwise LD between markers were plotted using the R package LDheatmap<sup>65</sup>.

**RNA sequencing (RNA-seq) analysis.** To maximize the number and diversity of genes that were identified by RNA-seq, mRNA was purified from various organs at multiple developmental stages derived from seven cultivars and hybrids. A total of nine libraries were generated (see **Supplementary Note** for more details).

The cDNA sequencing libraries were constructed following the manufacturer’s instructions (Illumina, San Diego, CA, USA), and the Illumina GA processing pipeline Cassava 1.7.0 was used for image analysis and base-calling.

**DNA extraction from leaf and developing fruits, and bisulfite sequencing.** Young leaves from GDDH13 and developing fruits from GDDH13 and GDDH18 (two biological replicates from independently grafted trees) were collected 3 d before pollination (–3 DAP, with petals, sepals, anthers and styles removed) and 9 DAP. DNA was purified using the Macherey-Nagel NucleoSpin plant II DNA extraction kit (Germany), following the manufacturer’s instructions. Bisulfite treatment was applied to determine the cytosine methylation status, using the Epitect bisulfite kit (Qiagen) and 100 ng of genomic DNA.

Whole-genome bisulfite sequencing was performed, and DMRs between leaves and young GDDH13 fruits, and between GDDH13 and GDDH18 fruits, at –3 DAP and 9 DAP were computed according to Hagmann *et al.*<sup>47</sup>. DNA methylation distribution plots and gene clustering analyses by methylation patterns were performed with deepTools<sup>66</sup>.

**Small RNA alignment.** Apple sRNA sequences derived from mature fruit parenchyma<sup>35</sup> were aligned to the Golden Delicious doubled-haploid pseudomolecules using BWA-MEM. Only perfectly mapped sequences were considered further, and reads with identical sequences were allowed to be mapped to two or more loci.

**Genome annotation.** RNA-seq data derived from nine different libraries was *de novo* assembled using Trinity<sup>67</sup> and SOAPdenovo-trans<sup>68</sup>. For each library, the assembly with the highest N50 value was chosen to annotate the genes. 2,033 mRNAs and 326,941 expressed sequence tags (ESTs) extracted from the NCBI nucleotide and EST databases, respectively, were also used for gene prediction.

The structural annotation of coding genes was performed using EuGene<sup>32</sup> by combining Gmap transcript mapping<sup>69</sup>, similarities detected with plant proteomes and Swiss-Prot, and *ab initio* predictions (interpolated Marlov model and weight-array matrix for donor and acceptor splicing sites). Moreover, the EuGene prediction was completed by tRNAscan-SE<sup>70</sup>, RNAMmer<sup>71</sup> and RfamScan<sup>72</sup> to annotate non-protein-coding genes, including those encoding tRNA, rRNA, miRNA and snoRNA, and other regions with proof of transcription but without significant similarities and coding potential (named ncRNA).

Functional annotation of proteins was performed using InterProScan<sup>73</sup>. The functional annotation was then completed by the prediction of targeted signals using the TargetP software<sup>74</sup>.

**Genome synteny.** SynMap (CoGe, see URLs) was used to identify collinearity blocks using homologous coding sequence pairs. Detailed methodological details on the annotation processes can be found in the **Supplementary Note**.

**Comparison of annotation between the heterozygous Golden Delicious and GDDH13 genomes.** *Malus domestica* predicted gene (MDP) sequences obtained from the heterozygous genome annotation<sup>6</sup> were mapped to the GDDH13 genome assembly using the best BLAT<sup>75</sup> hit. Comparison of the two genome annotations was done using Bio++<sup>76</sup>.

**Repeat annotation.** The TEdenovo pipeline<sup>37,77</sup> from the REPET package v2.5 (see URLs) was used to detect TEs in genomic sequences and to provide a consensus sequence for each TE family. Consensus TE sequences were used to annotate the TE copies in the whole genome using the TEannot pipeline<sup>38</sup> from the REPET package v2.5. Consensus sequences that were classified as potential host genes because they contain host gene Pfam domains were kept from this study. The same process was used to identify the HODOR consensus sequence on the PacBio assembly with the REPET pipeline. TE insertion ages were calculated using the adapted  $T = K/r$  formula for nonduplicated LTR sequences, where  $K$  is the sequence divergence, and  $r$  is the substitution rate<sup>78</sup>. The observed sequence divergence was corrected with the Jukes and Cantor model<sup>79</sup>. Additional methodological details on the repeat annotation can be found in the **Supplementary Note**.

**Data availability.** This whole-genome shotgun project has been deposited at GenBank under the accession code [MJAX000000001](https://www.ncbi.nlm.nih.gov/submit/SLI000000001). The raw Illumina mRNA sequences were submitted to the NCBI under BioProject ID [PRJNA191060](https://www.ncbi.nlm.nih.gov/bioproject/PRJNA191060), and the GDDH18 genome reads were deposited under BioProject ID [PRJNA379390](https://www.ncbi.nlm.nih.gov/bioproject/PRJNA379390). DNA methylation data can be accessed on the Gene Expression Omnibus website under accession codes [GSE87014](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE87014) and [GSE93950](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE93950). Structural and functional annotations are available through our genome browser (<https://iris.angers.inra.fr/gddh13/>).

61. Li, H. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics* **30**, 2843–2851 (2014).
62. Walker, B.J. *et al.* Pilon: an integrated tool for comprehensive microbial variant detection and genome-assembly improvement. *PLoS One* **9**, e112963 (2014).
63. Sahlin, K., Vezzi, F., Nystedt, B., Lundeberg, J. & Arvestad, L. BESST—efficient scaffolding of large fragmented assemblies. *BMC Bioinformatics* **15**, 281 (2014).
64. Bianco, L. *et al.* Development and validation of a 20K single-nucleotide polymorphism (SNP) whole-genome genotyping array for apple (*Malus × domestica* Borkh.). *PLoS One* **9**, e110377 (2014).
65. Shin, J.H., Blay, S., McNeney, B. & Graham, J. LDheatmap: an R function for graphical display of pairwise linkage disequilibria between single-nucleotide polymorphisms. *J. Stat. Software* **16**, c03 (2006).
66. Ramírez, F., Dündar, F., Diehl, S., Grüning, B.A. & Manke, T. deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res.* **42**, W187–W191 (2014).
67. Grabherr, M.G. *et al.* Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
68. Xie, Y. *et al.* SOAPdenovo-Trans: *de novo* transcriptome assembly with short RNA-seq reads. *Bioinformatics* **30**, 1660–1666 (2014).
69. Wu, T.D. & Watanabe, C.K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**, 1859–1875 (2005).
70. Lowe, T.M. & Chan, P.P. tRNAscan-SE On-line: integrating search and context for analysis of transfer RNA genes. *Nucleic Acids Res.* **44**, W54–W57 (2016).
71. Lagesen, K. *et al.* RNAMmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.* **35**, 3100–3108 (2007).
72. Nawrocki, E.P. *et al.* Rfam 12.0: updates to the RNA families database. *Nucleic Acids Res.* **43** (W1), D130–D137 (2015).

73. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
74. Emanuelsson, O., Brunak, S., von Heijne, G. & Nielsen, H. Locating proteins in the cell using TargetP, SignalP and related tools. *Nat. Protoc.* **2**, 953–971 (2007).
75. Kent, W.J. BLAT—the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
76. Guéguen, L. *et al.* Bio++: efficient extensible libraries and tools for computational molecular evolution. *Mol. Biol. Evol.* **30**, 1745–1750 (2013).
77. Hoede, C. *et al.* PASTEC: an automatic transposable element classification tool. *PLoS One* **9**, e91929 (2014).
78. de la Chaux, N., Tsuchimatsu, T., Shimizu, K.K. & Wagner, A. The predominantly selfing plant *Arabidopsis thaliana* experienced a recent reduction in transposable element abundance compared to its outcrossing relative *Arabidopsis lyrata*. *Mob. DNA* **3**, 2 (2012).
79. Jukes, T.H. & Cantor, C.R. in *Mammalian Protein Metabolism* (ed. Munro, H.N.) 21–132 (Elsevier, 1969).




RESEARCH

Open Access



# Inhibition of RNA polymerase II allows controlled mobilisation of retrotransposons for plant breeding

Michael Thieme<sup>1</sup>, Sophie Lanciano<sup>2,3</sup>, Sandrine Balzergue<sup>4</sup>, Nicolas Daccord<sup>4</sup>, Marie Mirouze<sup>2,3</sup> and Etienne Bucher<sup>4\*</sup> 

## Abstract

**Background:** Retrotransposons play a central role in plant evolution and could be a powerful endogenous source of genetic and epigenetic variability for crop breeding. To ensure genome integrity several silencing mechanisms have evolved to repress retrotransposon mobility. Even though retrotransposons fully depend on transcriptional activity of the host RNA polymerase II (Pol II) for their mobility, it was so far unclear whether Pol II is directly involved in repressing their activity.

**Results:** Here we show that plants defective in Pol II activity lose DNA methylation at repeat sequences and produce more extrachromosomal retrotransposon DNA upon stress in *Arabidopsis* and rice. We demonstrate that combined inhibition of both DNA methylation and Pol II activity leads to a strong stress-dependent mobilization of the heat responsive *ONSEN* retrotransposon in *Arabidopsis* seedlings. The progenies of these treated plants contain up to 75 new *ONSEN* insertions in their genome which are stably inherited over three generations of selfing. Repeated application of heat stress in progeny plants containing increased numbers of *ONSEN* copies does not result in increased activation of this transposon compared to control lines. Progenies with additional *ONSEN* copies show a broad panel of environment-dependent phenotypic diversity.

**Conclusions:** We demonstrate that Pol II acts at the root of transposon silencing. This is important because it suggests that Pol II can regulate the speed of plant evolution by fine-tuning the amplitude of transposon mobility. Our findings show that it is now possible to study induced transposon bursts in plants and unlock their use to induce epigenetic and genetic diversity for crop breeding.

**Keywords:** Epigenetics, DNA methylation, Genome integrity, Evolution, *Oryza sativa*, *Arabidopsis thaliana*

## Background

Like retroviruses, long terminal repeat (LTR) retrotransposons (class I elements), which represent the most abundant class of transposable elements (TEs) in eukaryotes, transpose via a copy and paste mechanism. This process requires the conversion of a full length RNA polymerase II (Pol II) transcript into extrachromosomal complementary DNA (ecDNA) by reverse transcription [1]. In their life cycle LTR retrotransposons can produce extrachromosomal circular DNA (eccDNA), which is an

indicator for their ongoing activity [2]. In plants, TEs are increasingly seen as a source of genetic and epigenetic variability and thus important drivers of evolution [3–6]. However, plants have evolved several regulatory pathways to retain control over the activity of these potentially harmful mobile genetic elements. Cytosine methylation (<sup>m</sup>C) plays a central role in TE silencing in plants [7]. In addition, plants have evolved two Pol II-related RNA polymerases, Pol IV and Pol V, that are essential to provide specific silencing signals leading to RNA-directed DNA methylation (RdDM) at TEs [8], thereby limiting their mobility [9–11]. More recently, various additional non-canonical Pol IV-independent RdDM pathways have been described [12]. Notably it was found that Pol II

\* Correspondence: etienne.bucher@inra.fr

<sup>4</sup>IRHS, Université d'Angers, INRA, AGROCAMPUS-Ouest, SFR4207 QUASAV, Université Bretagne Loire, 49045 Angers, France

Full list of author information is available at the end of the article

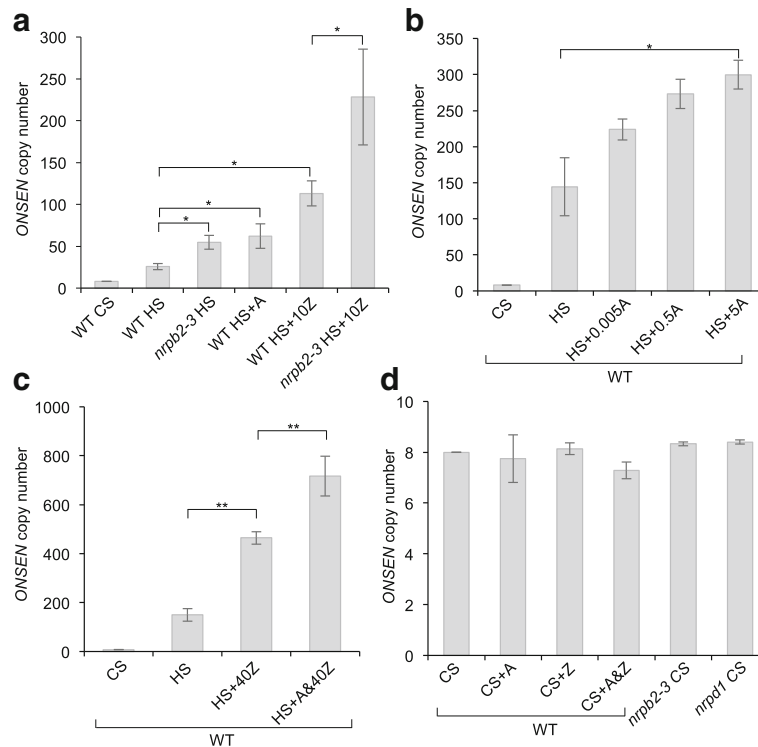


itself also plays an important role in RdDM [13, 14] by feeding template RNAs into downstream factors such as RNA-DEPENDENT RNA POLYMERASE 6 (RDR6), resulting in dicer-dependent or -independent initiation and establishment of TE-specific DNA methylation [15]. Beyond that, recent work suggests a new “non-canonical” branch of RdDM that specializes in targeting transcriptionally active full-length TEs [16]. This pathway functions independently of RDRs via Pol II transcripts that are directly processed by DCL3 into small interfering RNAs (siRNAs).

## Results

Here, we wanted to investigate if Pol II could play a direct role in repressing TE mobility in plants. For this purpose we chose the well-characterized heat-responsive *copia*-like *ONSEN* retrotransposon [11] of *Arabidopsis* and took advantage of the hypomorphic *nrbp2-3* mutant allele that causes reduced NRPB2 (the second-largest component of Pol II) protein levels [14]. Using quantitative real-time PCR (qPCR), we determined that challenging *nrbp2-3* seedlings by heat stress (HS) led to a mild increase

in total *ONSEN* copy number (sum of ecDNA, eccDNA and new genomic insertions) relative to control stress (CS) and compared to the wild type (WT) (Fig. 1a). This result is supported by the observed dose-responsive increase in *ONSEN* copy number after HS and pharmacological inactivation of Pol II with  $\alpha$ -amanitin (A), a potent Pol II inhibitor [17] that does not affect Pol IV or Pol V [18] (Fig. 1b). In order to test the interaction between Pol II-mediated repression of TE activation and DNA methylation, we grew WT and *nrbp2-3* plants on media supplemented with zebularine (Z), an inhibitor of DNA methyltransferases active in plants [19], and subjected them to HS. To ensure the viability of the *nrbp2-3* seedlings we choose a moderate amount of Z (10  $\mu$ M). The presence of Z in the medium during HS generally enhanced the production of *ONSEN* copies. Importantly, this induced increase in *ONSEN* copy number was more distinct in the *nrbp2-3* background (Fig. 1a). This indicated that both DNA methylation and Pol II transcriptional activity contribute to the repression of *ONSEN* ecDNA production. To complete their lifecycle, the reverse transcribed ecDNA of activated retrotransposons

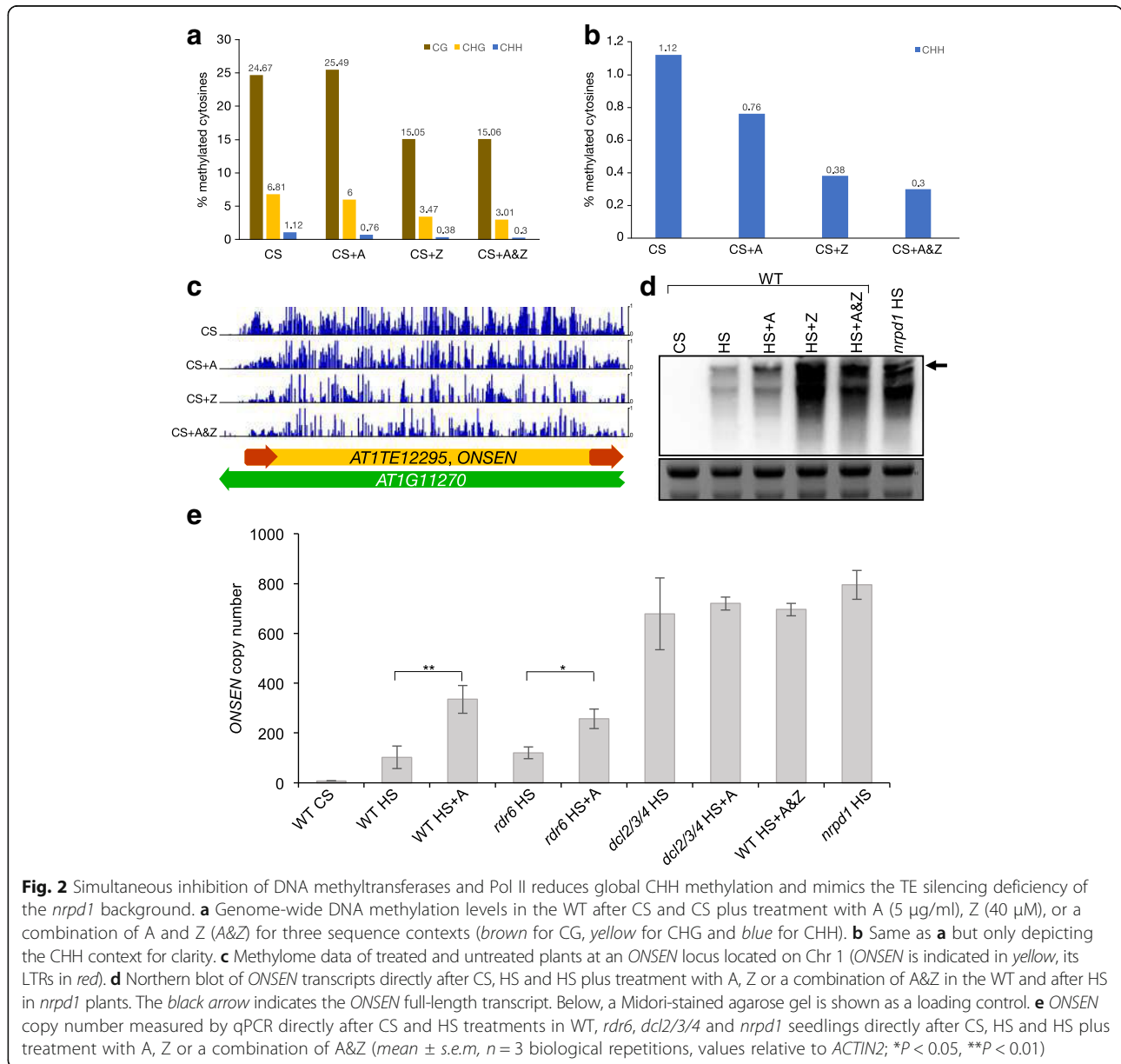


**Fig. 1** Pol II represses the HS-dependent mobility of the *ONSEN* retrotransposon in *Arabidopsis*. *ONSEN* copy number in *Arabidopsis* seedlings measured by qPCR directly after CS and HS treatments. **a** In the WT and the *nrbp2-3* mutant and after HS plus treatments with  $\alpha$ -amanitin (A; 5  $\mu$ g/ml) or zebularine (Z; 10  $\mu$ M) (mean  $\pm$  standard error of the mean (s.e.m.),  $n=6$  biological repetitions). **b** In the WT and after HS plus treatment with A at different concentrations ( $\mu$ g/ml) as specified on the x-axis (mean  $\pm$  s.e.m.,  $n=4$  biological repetitions). **c** In the WT and after HS plus treatment with Z (40  $\mu$ M) or a combination of A (5  $\mu$ g/ml) and Z (A&40Z) (mean  $\pm$  s.e.m.,  $n=3$  biological repetitions). **d** In the WT after chemical treatment with A (5  $\mu$ g/ml), Z (40  $\mu$ M), a combination of A and Z (A&Z) or in the *nrbp2-3* and *nrbp1* backgrounds following CS (mean  $\pm$  s.e.m.,  $n=3$  biological repetitions). All values are relative to *ACTIN2*. \* $P < 0.05$ , \*\* $P < 0.01$

has to integrate back into the genome [1]. Given that we observed a strong increase in *ONSEN* copy number after HS and treatment with moderate amounts of Z in the *nrbp2-3* background, we wanted to address the inheritance of additional *ONSEN* copies by the offspring. For this we compared the average *ONSEN* copy number of pooled S1 seedlings obtained from Z-treated and heat-stressed WT and *nrbp2-3* plants grown under controlled conditions on soil by qPCR. We observed a distinct increase in the overall *ONSEN* copy number exclusively in the *nrbp2-3* background (Additional file 1: Figure S1).

Because both DNA methylation and Pol II can be inhibited by the addition of specific drugs, we wanted to test if treating WT plants with both A and Z at the same

time could strongly activate and even mobilize *ONSEN* after a HS treatment. We grew WT seedlings on MS medium supplemented with Z (40  $\mu$ M) [19] individually or combined with A (5  $\mu$ g/ml, A&Z). Consistent with the strong activation of *ONSEN* in HS and Z-treated *nrbp2-3* seedlings, the combined treatment (A&Z) of the WT gave rise to a very high (Fig. 1c) HS-dependent (Fig. 1d) increase in *ONSEN* copy number, comparable to that in the *nrbp1* background (Fig. 2e). We noted that the overall amplitude of HS-dependent *ONSEN* activation could vary between different waves of stress applications in terms of copy number (Fig. 1a, b). Yet, the observed enhancing effect of Pol II and DNA methyltransferase inhibition with A and Z on *ONSEN* activation was consistent in



**Fig. 2** Simultaneous inhibition of DNA methyltransferases and Pol II reduces global CHH methylation and mimics the TE silencing deficiency of the *nrpd1* background. **a** Genome-wide DNA methylation levels in the WT after CS and CS plus treatment with A (5  $\mu$ g/ml), Z (40  $\mu$ M), or a combination of A and Z (A&Z) for three sequence contexts (brown for CG, yellow for CHG and blue for CHH). **b** Same as **a** but only depicting the CHH context for clarity. **c** Methylation data of treated and untreated plants at an *ONSEN* locus located on Chr 1 (*ONSEN* is indicated in yellow, its LTRs in red). **d** Northern blot of *ONSEN* transcripts directly after CS, HS and HS plus treatment with A, Z or a combination of A&Z in the WT and after HS in *nrpd1* plants. The black arrow indicates the *ONSEN* full-length transcript. Below, a Midori-stained agarose gel is shown as a loading control. **e** *ONSEN* copy number measured by qPCR directly after CS and HS treatments in WT, *rdt6*, *dcl2/3/4* and *nrpd1* seedlings directly after CS, HS and HS plus treatment with A, Z or a combination of A&Z (mean  $\pm$  s.e.m, n = 3 biological repetitions, values relative to *ACTIN2*; \*P < 0.05, \*\*P < 0.01)

independent experiments (Figs. 1a–c and 2e). To detect activated TEs at the genome-wide level we took advantage of the production of eccDNA by active retrotransposons. eccDNA is a byproduct of the LTR retrotransposon life cycle [20]. Using mobilome sequencing, which comprises a specific amplification step of circular DNA followed by high-throughput sequencing to identify eccDNA derived from active LTR retrotransposons [2], we found that only *ONSEN* was activated by HS in combination with A&Z (Additional file 1: Figure S2). Confirming our qPCR data, more *ONSEN*-specific reads were detected in the presence of A and Z in the medium.

To better understand the mechanisms by which the drugs enhanced the activation of *ONSEN* after HS at the DNA level, we assessed how they influenced DNA methylation at the genome-wide level using whole-genome bisulfite sequencing (WGBS) after CS. Overall, we found that all drug treatments affected global DNA methylation levels. While the treatment with Z affected all sequence contexts, we observed that inhibition of Pol II primarily affected cytosine methylation in the CHG and CHH sequence contexts (where H is an A, T or G). The combined A&Z treatment had a slight additive demethylating effect in the CHG and CHH contexts compared to A or Z alone (Fig. 2a, b). DNA methylation levels at one *ONSEN* locus (*AT1TE12295*) is depicted in Fig. 2c. Treatment with A led to a slight decrease in DNA methylation, which was more apparent in Z- and A&Z-treated plants. We then checked by northern blot whether the degree of reduction in DNA methylation would coincide with increased *ONSEN* transcript levels directly after HS. We found that treatment with Z alone resulted in the highest *ONSEN* transcript level after HS (Fig. 2d). Considering the data obtained on *ONSEN* ecDNA (Fig. 1c), we concluded that a substantial proportion of these Z-induced transcripts were not suitable templates for *ONSEN* ecDNA synthesis.

In *Drosophila*, it has been shown that Pol II-mediated antisense transcription results in the production of TE-derived siRNAs in a Dicer-2-dependent manner [21]. In support of this in *Arabidopsis*, a recent publication pointed out the importance of DCL3 in regulating *ONSEN* in the *ddm1* background [16]. To elucidate whether the effect of Pol II inhibition was also dicer-dependent, we grew both *rdr6* and *dcl2/3/4* triple mutant plants on A, applied HS and measured *ONSEN* ecDNA levels. Strikingly, we found that A still enhanced ecDNA accumulation in *rdr6* plants, whereas inhibition of Pol II had no additional effect in the *dcl2/3/4* triple mutant (Fig. 2e).

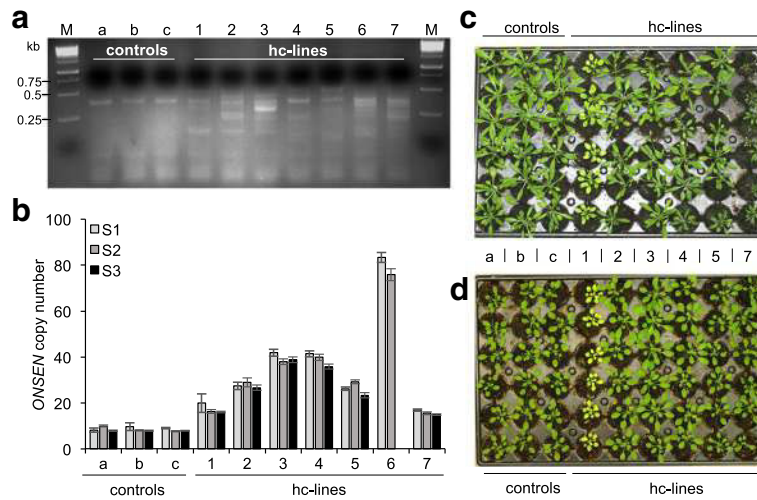
Induced mobilization of endogenous TEs in plants has so far been very inefficient, thus limiting their use in basic research and plant breeding [3]. In the case of *Arabidopsis*, transposition of *ONSEN* in HS-treated WT plants has not been observed [11, 22]. Because the A&Z

drug treatment resulted in high accumulation of *ONSEN* copy numbers—essentially mimicking plants defective in NRDP1 (Fig. 2e)—we wanted to test if the combined drug treatment could lead to efficient *ONSEN* mobilization in WT plants. First, we assessed by qPCR if, and at what frequencies, new *ONSEN* copies could be detected in the progeny of A&Z-treated and heat stressed plants. In fact, we found new *ONSEN* insertions in 29.4% of the tested S1 (selfed first generation) pools (n = 51), with pools having up to 52 insertions (Additional file 1: Figure S3). We then confirmed stable novel *ONSEN* insertions in a subset of independent individual high copy plants by transposon display (Fig. 3a), qPCR (Fig. 3b) and sequencing of 11 insertions in a selected high-copy line (hc line 3; Fig. 4; Additional file 1: Figure S4). Tracking *ONSEN* copy numbers over three generations of selfing indicated that the new insertions were stably inherited (Fig. 3b). Furthermore, the re-application of heat stress and drugs in the S3 generation of two hc lines did not lead to greater accumulation of *ONSEN* copies compared to control lines, but we instead observed stronger silencing in lines with more *ONSEN* copies (Additional file 1: Figure S5).

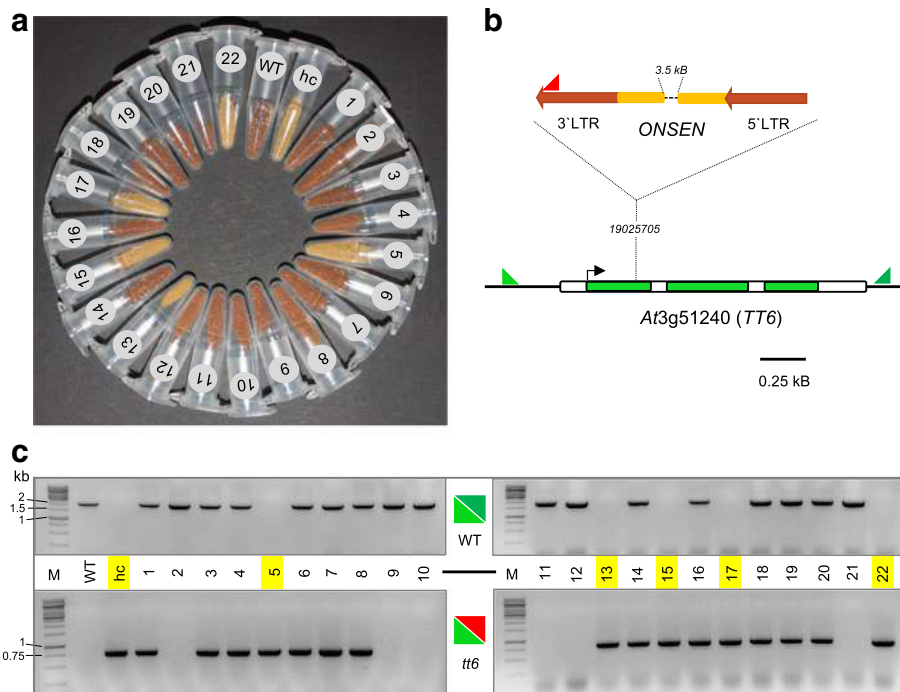
TE insertions can interrupt genes or alter their expression by recruiting epigenetic marks or by stress-dependent readout transcription from the 3' LTR into flanking regions [6]. To test this, we grew the S2 generation of the selected hc lines under long- and short-day conditions. Interestingly, we observed that many hc lines showed clear and homogenous phenotypes in response to the different growth conditions (plant size, chlorophyll content and flowering time; Fig. 3c, d).

To demonstrate that *ONSEN* insertions could directly influence such developmental phenotypes, we closely investigated hc line 3, which produced white seeds (Fig. 4a). Using a candidate gene approach, we found that an *ONSEN* insertion in *transparent testa 6* (*TT6*, *AT3G51240*; Fig. 4b) was responsible for the recessive white seed phenotype [23, 24]. This was confirmed by segregation analysis of the F2 generation of a cross between WT and hc line 3 (Fig. 4a) followed by genotyping (Fig. 4c).

Next, we wanted to test if Pol II plays a more general role in repressing TEs in plants. Due to its significantly different epigenetic and TE landscape compared to *Arabidopsis*, we wanted to test if we could mobilize TEs in rice (*Oryza sativa*) [25], a genetically well-characterized monocotyledonous crop. To capture drug-induced mobilized TEs, we characterized the active mobilome in *O. sativa* seedlings that were grown on MS medium supplemented with no drugs, A only, Z only or a combination of A and Z, using the same approach as we used for *Arabidopsis*. We identified *Houba*, a copia-like retrotransposon [26], as highly activated only when plants were treated with A&Z (Fig. 5a). Bona fide activity of

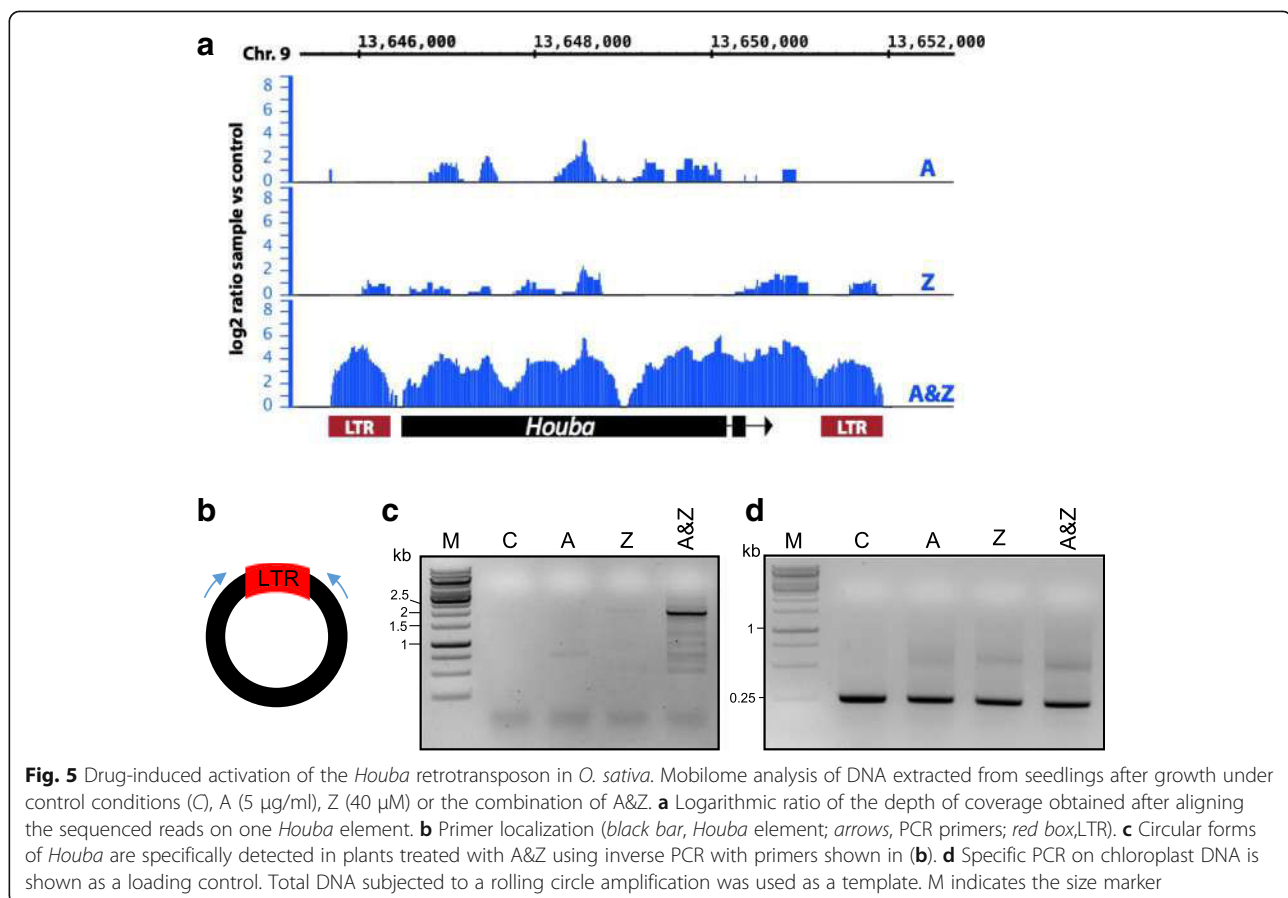


**Fig. 3** Drug-induced mobilization of *ONSEN* in WT *Arabidopsis* plants. **a** Transposon display testing seedlings in the S2 generation of WT plants for novel *ONSEN* insertions: lanes a to c show HS-treated plants; lanes 1 to 7 show hc lines 1–7 treated with HS and A (5 µg/ml) and Z (40 µM), M indicates the size marker. **b** *ONSEN* copy number in the S1, S2 and S3 generations measured by qPCR (mean ± s.e.m, n = 3 technical replicates, values relative to *ACTIN2*). **c, d** Photographs of S2 plants showing both homogeneous and environment-dependent phenotypic variability induced by the *ONSEN* mobilization when grown under long (**c**) and short day (**d**) conditions. qPCR data for the S3 generation of line 6 in **b** as well as pictures of phenotypes in **c** and **d** are missing due to severe infertility and extinction of this line



**Fig. 4** Transparent testa phenotype of hc line 3 co-segregates with an *ONSEN* insertion in *TT6*. Seed phenotypes (**a**) and corresponding genotypes (**c**) of a segregating F2 population (lanes 1–22) obtained from a cross between the WT and hc line 3 (*hc*) are shown. **b** Primers used for genotyping of the *ONSEN* insertion. For the WT-PCR depicted in the upper part of **c** the light (*tt6 fw*) and dark (*tt6 rev*) green primers flanking the *TT6* locus (*AT3G51240*) were used. The *ONSEN* insertion in *TT6* was detected by a combination of the light green primer with the red primer specific to the *ONSEN* LTR (*Copia* 78 3' LTR, red arrow). M indicates the size marker. Primer sequences are given in Additional file 1: Table S1





*Houba* was supported by the detection of eccDNA containing LTR–LTR junctions (Additional file 1: Figure S6). The activation of *Houba* was further confirmed by eccDNA-specific PCR on the *Houba* circles (Fig. 5b–d).

## Discussion

In this study, we show the importance of Pol II in the repression of TE mobility in plants. By choosing the well-characterized heat inducible *ONSEN* retrotransposon, we were able to specifically address the role of Pol II in silencing transcriptionally active endogenous TEs in WT plants. Recent studies propose Pol II as the primary source for the production of TE-silencing signals that can then feed into the RNA silencing and DNA methylation pathways [15]. Our data strongly support these findings at two levels. First, we found that inhibition of Pol II activity reduced the degree of DNA methylation at *ONSEN*, demonstrating its distinct role in this process, and that Pol II also contributes to reinforcing silencing at the genome-wide level, primarily in the CHH but also in the CHG context. Second, our finding that DCL enzymes are sufficient to process the silencing signal produced by Pol II suggest that Pol II acts at very early steps in the TE silencing pathway by providing substrates

to these enzymes. The observation that inhibition of Pol II in the *rdr6* background still further enhanced *ONSEN* accumulation after HS supports the notion that Pol II plays a central role in the previously proposed expression-dependent RdDM pathway [16].

Using mobilome sequencing we confirmed previous findings [2] that this approach is a powerful diagnostic tool to detect mobile retrotransposons: we detected highest levels of eccDNA of *ONSEN* in HS and drug-treated *Arabidopsis* seedlings and found new insertions in successive generations of these plants. Using the same approach on rice we were able to detect production of *Houba* eccDNA after drug treatments, suggesting that the progeny will then contain novel *Houba* insertions. This is still to be confirmed and may be hampered by the already very high *Houba* copy number present in the genome [27].

Our findings may indicate that Pol II is primarily involved in silencing young, recently active retrotransposons and perhaps to a lesser extent other tightly silenced TEs. Indeed, there are indications of very recent natural transposition events for *ONSEN* [28] and *Houba* [29] in the *Arabidopsis* and rice genomes, respectively. For instance, the annual temperature range has and may still contribute to contrasting *ONSEN* mobilization events in different *Arabidopsis* accessions [28]. *Houba* is the most

abundant TE of the *copia* family in rice and has been active in the last 500,000 years [30].

Overall, our findings lead to the question of when plants lower their guard: under what conditions could Pol II be less effective in silencing TEs? Certain stresses that affect the cell cycle have been reported to lead to the inactivation of Pol II [31, 32]; this would provide a window of opportunity for TEs to be mobilized. Therefore, combined stresses that affect the cell cycle and activate TEs may lead to actual TE bursts under natural growth conditions. Interestingly, it has been reported that retrotransposon-derived short interspersed element (SINE) transcripts can inhibit Pol II activity [33]. This strongly suggests the presence of an ongoing arms race between retrotransposons and Pol II. Considering that almost all organisms analyzed so far have TEs [4] and RNA polymerases [34] and the reliance of TEs on host RNA polymerases, it may—from an evolutionary point of view—not come as a surprise that Pol II also has a function as an important regulator of retrotransposon activity. Strikingly, it has been shown in both *Saccharomyces cerevisiae* and *Drosophila melanogaster* that Pol II-dependent intra-element antisense transcription plays an important role in TE silencing [21, 35]. In addition, we observed a discrepancy in *ONSEN* transcript accumulation and measured ecDNA after HS in seedlings that were treated with zebularine only. This substantiates the notion that both the quantity and quality of transcripts affect regulation, reverse transcription and successful integration of retrotransposons. This is well in line with previous observations demonstrating that different TE-derived transcripts have distinct functions in the regulation of TE activity [36]. As a next step it will be of great interest to investigate if Pol II-dependent antisense transcription of TEs and subsequent dicer-dependent processing may be the key to solve “the chicken and the egg problem” of *de novo* silencing functional retrotransposons in eukaryotes.

Finally, our findings will allow future studies on the potential beneficial role TEs play in adaptation to stresses. Indeed, two recent studies point out the adaptive potential of retrotransposon and, more specifically, *ONSEN* copy number variation in natural accessions [28] and RdDM mutant backgrounds of *Arabidopsis* [37]. Upon mobilization, the heat-response elements in the LTRs of *ONSEN* [38] can create new gene regulatory networks responding to heat stress [11]. Therefore, it will now be of great interest to test if the *ONSEN* hc lines obtained in this study are better adapted to heat stress. This will allow us to test if retrotransposon-induced genetic and epigenetic changes more rapidly create beneficial alleles than would occur by random mutagenesis. Furthermore, the observation that HS did not lead to a stronger activation of *ONSEN* in hc lines

compared to WT plants suggests that genome stability is not compromised in these lines. This result can be explained by at least two possible mechanisms: (i) the occurrence of insertions of inverted duplications of *ONSEN*, such as has been observed for the *Mu* killer locus in maize [39]—such insertions will lead to the production of double-stranded RNA feeding into gene silencing and thereby limit the activity of that TE; and (ii) balancing of TE activity and integrated copy number as has been described for *EVADÉ* in *Arabidopsis* [40]. In this case, when a certain TE copy number threshold is reached robust transcriptional gene silencing takes over, thereby limiting TE mobility and ensuring genome stability. The stability of new TE insertions is an important aspect in light of the future use of TEs in crop breeding and trait stability.

## Conclusions

TEs are important contributors to genome evolution. The ability to mobilize them in plants and possibly in other eukaryotes in a controlled manner with straightforward drug application, as shown here, opens the possibility to study their importance in inducing genetic and epigenetic changes resulting from external stimuli. Because the induced transposition of *ONSEN* can efficiently produce developmental changes in *Arabidopsis*, it will be very interesting to test if specific stress-induced TE activation can be used for directed crop breeding for better stress tolerance in the near future.

## Methods

### Plant material

All *Arabidopsis* mutants used in this study (*nrbp2-3* [14], *nrbp1-3* [41], *rdr6* [42], *dcl2/3/4* triple mutant [43]) are in the Col-0 background. For *O. sativa japonica*, the cultivar Nipponbare was used.

### Growth conditions

Prior to germination, *Arabidopsis* seeds were stratified for 2 days at 4 °C. Before and during stress treatments plants were grown under controlled conditions in a Sanyo MLR-350 growth chamber on solid ½ MS medium (1% sucrose, 0.5% Phytagel (Sigma), pH 5.8) under long day conditions (16 h light) at 24 °C (day) and 22 °C (night) (*Arabidopsis*) and 12 h at 28 °C (day) and 27 °C (night) (*O. sativa*).

To analyze successive generations, seedlings were transferred to soil and grown under long day conditions (16 h light) at 24 °C (day) and 22 °C (night) (*Arabidopsis*) in a Sanyo MLR-350 growth chamber until seed maturity.

For phenotyping, *Arabidopsis* plants were grown under long day conditions (16 h light) at 24 °C (day) and 22 °C (night) and short day conditions (10 h light) at 21 °C (day) and 18 °C (night).

### Stress and chemical treatments

Surface sterilized seeds of *Arabidopsis* and *O. sativa* were germinated and grown on solid ½ MS medium that was supplemented with sterile filtered zebularine (Sigma; stock, 5 mg/ml in DMSO),  $\alpha$ -amanitin (Sigma; stock, 1 mg/ml in water) or a combination of both chemicals. Control stresses (6 °C for 24 h followed by control conditions for 24 h, CS) and heat stresses (6 °C for 24 h followed by 37 °C for 24 h, HS) of *Arabidopsis* seedlings were conducted as described previously [11].

### DNA analysis

For qPCR and prior to digestions, total DNA from *Arabidopsis* plants was extracted with the DNeasy Plant Mini Kit (Qiagen) following the manufacturer's recommendations. For the qPCRs to measure the *ONSEN* copy number following HS and chemical treatments the aerial parts of at least ten *Arabidopsis* plants per replicate were pooled prior to DNA extraction. To track *ONSEN* copy numbers in the S1–3 generations of controls (only HS) and hc lines (HS + A&Z treatment) DNA from true leaves was extracted. For the estimation of the *ONSEN* transposition frequency, total DNA of pools consisting of at least eight seedlings of the progeny of HS + A&Z-treated plants was isolated. The DNA concentration was measured with a Qubit Fluorometer (Thermo Fisher Scientific). The copy numbers of *ONSEN* were determined with qPCRs on total DNA using a TaqMan master mix (Life Technologies) in a final volume of 10  $\mu$ l in the Light-Cycler 480 (Roche). *ACTIN2* (*AT3G18780*) was used to normalize DNA levels. Primer sequences are given in Additional file 1: Table S1.

For the mobilome-seq analysis total DNA from the pooled aerial parts of three 10-day-old *O. sativa* seedlings was extracted as previously reported [44]. Genomic DNA (5  $\mu$ g) for each sample was purified using a GeneClean kit (MPBio, USA) according to the manufacturer's instructions. ecDNA was isolated from the GeneClean product using PlasmidSafe DNase (Epicentre, USA) according to the manufacturer's instructions, except that the 37 °C incubation was performed for 17 h. DNA samples were precipitated by adding 0.1 volume of 3 M sodium acetate (pH 5.2), 2.5 volumes of ethanol and 1  $\mu$ l of glycogen (Fisher, USA) and incubating overnight at –20 °C. The precipitated circular DNA was amplified by random rolling circle amplification using the Illustra TempliPhi kit (GE Healthcare, USA) according to the manufacturer's instructions except that the incubation was performed for 65 h at 28 °C. The DNA concentration was determined using the DNA PicoGreen kit (Invitrogen, USA) using a LightCycler480 (Roche, USA). One nanogram of amplified ecDNA from each sample was used to prepare the libraries using the Nextera XT library kit (Illumina, USA) according to the manufacturer's instructions. DNA quality

and concentration were determined using a high sensitivity DNA Bioanalyzer chip (Agilent Technologies, USA). Samples were pooled and loaded onto a MiSeq platform (Illumina, USA) and 2  $\times$  250-nucleotide paired-end sequencing was performed. Quality control of FASTQ files was done using the FastQC tool (version 0.10.1). To remove any read originating from organelle circular genomes, reads were mapped against the mitochondria and chloroplast genomes using the program Bowtie2 version 2.2.2 71 with –sensitive local mapping. Unmapped reads were mapped against the reference genome IRGSP1.0 (<http://rgp.dna.affrc.go.jp/E/IRGSP/Build5/build5.html>) using the following parameters: –sensitive local, -k 1. DNA from both mitochondria and chloroplast genomes integrated in nuclear genomes was masked (1,697,400 bp). The TE-containing regions cover 194,224,800 bp in *O. sativa*. Finally, the bam alignment files were normalized and compared using deeptools [45] and visualized with the Integrative Genomics Viewer (IGV) software (<https://www.broadinstitute.org/igv/>). Data from the mobilome analysis were submitted to GEO (accession number GSE90484).

The presence of circular *Houba* copies was tested by an inverse PCR on 7 ng of the rolling-circle amplified template that was also used for sequencing. A PCR specific to chloroplast DNA served as a loading control. PCR products were separated on a 1% agarose gel that was stained with a Midori Green Nucleic Acid Staining Solution (Nippon Genetics Europe). Primer sequences are given in Additional file 1: Table S1.

### Transposon display

The integration of additional copies of *ONSEN* into the genome of heat stressed and treated plants was ascertained by a simplified transposon display based on the GenomeWalker Universal kit (Clontech Laboratories), as previously described [11] with the following modifications: 300 ng of total DNA from adult plants in the S2 generation of heat stressed and A&Z-treated plants was extracted with a DNeasy Plant Mini Kit (QIAGEN) and digested with blunt cutter restriction enzyme *DraI* (NEB). After purification with a High Pure PCR Product Purification Kit (Roche) digested DNA was ligated to the annealed GenWalkAdapters 1&2. The PCR was performed with the adaptor-specific primer AP1 and the *ONSEN*-specific primer Copia78 3' LTR. The PCR products were separated on a 2% agarose gel that was stained with Midori Green. For primer sequence information, see Additional file 1: Table S1.

### Cloning, sequencing and genotyping of new insertions

To identify the genomic region of new *ONSEN* insertions, the PCR product of the transposon display was purified using a High Pure PCR Product Purification Kit



(Roche), ligated into a pGEM-T vector (Promega) and transformed into *Escherichia coli*. After a blue white selection, positive clones were used for the insert amplification and sequencing (StarSEQ). The obtained sequences were analyzed with Geneious 8.2.1 and blasted against the *Arabidopsis* reference genome. The standard genotyping PCRs to prove novel *ONSEN* insertions were performed with combinations of the *ONSEN*-specific primer Copia78 3' LTR and primers listed in Additional file 1: Table S1.

### RNA analysis and northern blotting

Total RNA from the aerial part of at least ten *Arabidopsis* seedlings was isolated using the TRI Reagent (Sigma) according to the manufacturer's recommendations. RNA concentration was measured (Qubit RNA HS Assay Kit, Thermo Fisher) and 15 µg of RNA was separated on a denaturing 1.5% agarose gel, blotted on a Hybond-N<sup>+</sup> (GE Healthcare) membrane and hybridized with 25 ng of a gel-purified and P<sup>32</sup>-labelled probe (Megaprime DNA Labelling System, GE Healthcare) specific to the full length *ONSEN* transcript (see Additional file 1: Table S1 for primer sequences). Northern blots were repeated in three independent experiments with the same results.

### Whole-genome DNA methylation analysis

Whole-genome bisulfite sequencing library preparation and DNA conversion were performed as previously reported [46]. Bisulphite read mapping and methylation value extraction were done on the *Arabidopsis* TAIR10 genome sequence using BSMAP v2.89 [47]. Following mapping of the reads the fold coverages of the genome for CS, CS + A, CS + Z and CS + A&Z were 13.4, 13.2, 18.4 and 16.3, respectively. Data from the bisulphite sequencing analysis have been submitted to GEO (accession number GSE99396).

### Statistics

Statistical analyses were performed with SigmaPlot (v. 11.0). Depending on the normality of the data, either an H-test or a one-way ANOVA was performed. The Student-Newman-Keuls method was used for multiple comparisons.

### Additional file

**Additional file 1: Table S1.** Table of all primers used in this study. **Figure S1.** Increase in *ONSEN* copy numbers in S1 pools of heat-stressed and Z-treated *nrbp2-3* plants. **Figure S2.** Detection of eccDNAs originating from *ONSEN* loci following heat stress and chemical treatments in *Arabidopsis*. **Figure S3.** Increase in *ONSEN* copy numbers in S1 pools of heat-stressed and A&Z-treated WT plants. **Figure S4.** Summary of confirmed novel *ONSEN* insertions in hc line 3. **Figure S5.** Stress-induced activation of *ONSEN* in the S3 generation after initial HS treatment. **Figure S6.** *Houba* forms LTR-LTR junction eccDNAs after combined A&Z treatment. (PDF 1660 kb)

### Acknowledgements

We wish to thank Emilija Hristova for her support at the beginning of this project. We thank Christel Llauro for technical support on the production of the mobilomes and Todd Blevins for providing the *dc12/3/4* triple mutant line. We thank the IMAC platform from the Structure Fédérative de Recherche 'Qualité et Santé du Végétal' (SFR QUASAV) for their technical support (Illumina sequencing).

### Funding

This work was supported by grants provided by the European Commission (PITN-GA-2013-608422-IDP BRIDGES to MT and ERC grant 725701 BUNGEE to EB) and the region of Pays de la Loire (ConnecTalent EPICENTER project awarded to EB).

### Availability of data and materials

The mobilome sequencing data and whole-genome DNA methylation analysis data are available in the GEO (accession numbers GSE90484 and GSE99396, respectively).

### Authors' contributions

MT and EB conceived the study. MT, SL, SB and MM performed experiments. ND performed methylome analyses. MT and EB wrote the paper with contributions from SL and MM. All authors read and approved the final manuscript.

### Ethics approval and consent to participate

Not applicable.

### Competing interests

MT and EB declare that a patent application based on the presented discoveries has been submitted to the European Patent Office (PCT/EP2016/079276). EB is CEO of epibreed Ltd, a company that has an exclusive use license for this patent.

### Author details

<sup>1</sup>Botanical Institute, Zürich-Basel Plant Science Center, University of Basel, Hebelstrasse 1, 4056 Basel, Switzerland. <sup>2</sup>Institut de Recherche pour le Développement, UMR232 DIADE Diversité Adaptation et Développement des Plantes, Université Montpellier 2, Montpellier, France. <sup>3</sup>University of Perpignan, Laboratory of Plant Genome and Development, 58 Avenue Paul Alduy, 66860 Perpignan, France. <sup>4</sup>IRHS, Université d'Angers, INRA, AGROCAMPUS-Ouest, SFR4207 QUASAV, Université Bretagne Loire, 49045 Angers, France.

Received: 13 February 2017 Accepted: 27 June 2017

Published online: 07 July 2017

### References

- Schulman AH. Retrotransposon replication in plants. *Curr Opin Virol.* 2013;3:604–14.
- Lanciano S, Carpentier M-C, Llauro C, Jobet E, Robakowska-Hyzorek D, Lasserre E, et al. Sequencing the extrachromosomal circular mobilome reveals retrotransposon activity in plants. *PLoS Genet.* 2017;13:e1006630.
- Paszowski J. Controlled activation of retrotransposition for plant breeding. *Curr Opin Biotechnol.* 2015;32C:200–6.
- Huang CRL, Burns KH, Boeke JD. Active transposition in genomes. *Annu Rev Genet.* 2012;46:651–75.
- Belyayev A. Bursts of transposable elements as an evolutionary driving force. *J Evol Biol.* 2014;27:2573–84.
- Lisch D. How important are transposons for plant evolution? *Nat Rev Genet.* 2013;14:49–61.
- Miura A, Yonebayashi S, Watanabe K, Toyama T, Shimada H, Kakutani T. Mobilization of transposons by a mutation abolishing full DNA methylation in *Arabidopsis*. *Nature.* 2001;411:212–4.
- Matzke MA, Moshier RA. RNA-directed DNA methylation: an epigenetic pathway of increasing complexity. *Nat Rev Genet.* 2014;15:394–408.
- Tsukahara S, Kobayashi A, Kawabe A, Mathieu O, Miura A, Kakutani T. Bursts of retrotransposition reproduced in *Arabidopsis*. *Nature.* 2009;461:423–U125.
- Mirouze M, Reinders J, Bucher E, Nishimura T, Schneeberger K, Ossowski S, et al. Selective epigenetic control of retrotransposition in *Arabidopsis*. *Nature.* 2009;461:427–30.

11. Ito H, Gaubert H, Bucher E, Mirouze M, Vaillant I, Paszkowski J. An siRNA pathway prevents transgenerational retrotransposition in plants subjected to stress. *Nature*. 2011;472:115–9.
12. Matzke MA, Kanno T, Matzke AJM. RNA-directed DNA methylation: the evolution of a complex epigenetic pathway in flowering plants. *Annu Rev Plant Biol*. 2015;66:243–67.
13. Gao Z, Liu H-L, Daxinger L, Pontes O, He X, Qian W, et al. An RNA polymerase II- and AGO4-associated protein acts in RNA-directed DNA methylation. *Nature*. 2010;465:106–9.
14. Zheng B, Wang Z, Li S, Yu B, Liu J-Y, Chen X. Intergenic transcription by RNA Polymerase II coordinates Pol IV and Pol V in siRNA-directed transcriptional gene silencing in Arabidopsis. *Genes Dev*. 2009;23:2850–60.
15. Cuerda-Gil D, Slotkin RK. Non-canonical RNA-directed DNA methylation. *Nature Plants*. 2016;2:16163.
16. Panda K, Ji L, Neumann DA, Daron J, Schmitz RJ, Slotkin RK. Full-length autonomous transposable elements are preferentially targeted by expression-dependent forms of RNA-directed DNA methylation. *Genome Biol*. 2016;17:1–19.
17. Lindell TJ, Weinberg F, Morris PW, Roeder RG, Rutter WJ. Specific inhibition of nuclear RNA polymerase II by alpha-Amanitin. *Science*. 1970;170:447–9.
18. Haag JR, Ream TS, Marasco M, Nicora CD, Norbeck AD, Pasa-Tolić L, et al. In vitro transcription activities of Pol IV, Pol V, and RDR2 reveal coupling of Pol IV and RDR2 for dsRNA synthesis in plant RNA silencing. *Mol Cell*. 2012;48:811–8.
19. Baubec T, Pecinka A, Rozhon W, Mittelsten SO. Effective, homogeneous and transient interference with cytosine methylation in plant genomic DNA by zebularine. *Plant J*. 2009;57:542–54.
20. Flavell AJ, Ish-Horowitz D. Extrachromosomal circular copies of the eukaryotic transposable element Copia in cultured *Drosophila* cells. *Nature*. 1981;292:591–5.
21. Russo J, Harrington AW, Steiniger M. Antisense transcription of retrotransposons in *Drosophila*: an origin of endogenous small interfering RNA precursors. *Genetics*. 2016;202:107–21.
22. Matsunaga W, Ohama N, Tanabe N, Masuta Y, Masuda S, Mitani N, et al. A small RNA mediated regulation of a stress-activated retrotransposon and the tissue specific transposition during the reproductive period in Arabidopsis. *Front Plant Sci*. 2015;6:48.
23. Rosso MG, Li Y, Strizhov N, Reiss B, Dekker K, Weisshaar B. An Arabidopsis thaliana T-DNA mutagenized population (GABI-Kat) for flanking sequence tag-based reverse genetics. *Plant Mol Biol*. 2003;53:247–59.
24. Appelhagen I, Thiedig K, Nordholt N, Schmidt N, Huep G, Sagasser M, et al. Update on transparent testa mutants from Arabidopsis thaliana: characterisation of new alleles from an isogenic collection. *Planta*. 2014;240:955–70.
25. Kawahara Y, la Bastide de M, Hamilton JP, Kanamori H, McCombie WR, Ouyang S. Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data. *Rice*. 2013;6:4–10.
26. Panaud O, Vitte C, Hivert J, Muziak S, Talag J, Brar D, et al. Characterization of transposable elements in the genome of rice (*Oryza sativa* L.) using Representational Difference Analysis (RDA). *Mol Genet Genomics*. 2002;268:113–21.
27. Vitte C, Panaud O. LTR retrotransposons and flowering plant genome size: emergence of the increase/decrease model. *Cytogenet Genome Res*. 2005;110:91–107.
28. Quadrona L, Silveira AB, Mayhew GF, LeBlanc C, Martienssen RA, Jeddalo JA, et al. The Arabidopsis thaliana mobilome and its impact at the species level. *elife*. 2016;5:e15716.
29. Vitte C, Panaud O, Quesneville H. LTR retrotransposons in rice (*Oryza sativa*, L.): recent burst amplifications followed by rapid DNA loss. *BMC Genomics*. 2007;8:218–15.
30. Wicker T, Keller B. Genome-wide comparative analysis of copia retrotransposons in Triticeae, rice, and Arabidopsis reveals conserved ancient evolutionary lineages and distinct dynamics of individual copia families. *Genome Res*. 2007;17:1072–81.
31. Oelgeschläger T. Regulation of RNA polymerase II activity by CTD phosphorylation and cell cycle control. *J Cell Physiol*. 2001;190:160–9.
32. Palancade B, Bensaude O. Investigating RNA polymerase II carboxyl-terminal domain (CTD) phosphorylation. *Eur J Biochem*. 2003;270:3859–70.
33. Pai DA, Kaplan CD, Kweon HK, Murakami K, Andrews PC, Engelke DR. RNAs nonspecifically inhibit RNA polymerase II by preventing binding to the DNA template. *RNA*. 2014;20:644–55.
34. Lazzano A, Fastag J, Gariglio P, Ramirez C, Oro J. On the early evolution of RNA-polymerase. *J Mol Evol*. 1988;27:365–76.
35. Berretta J, Pinskaya M, Morillon A. A cryptic unstable transcript mediates transcriptional trans-silencing of the Ty1 retrotransposon in *S. cerevisiae*. *Genes Dev*. 2008;22:615–26.
36. Chang W, Jääskeläinen M, Li S-P, Schulman AH. BARE retrotransposons are translated and replicated via distinct RNA pools. *PLoS One*. 2013;8:e72270–12.
37. Ito H, Kim J-M, Matsunaga W, Saze H, Matsui A, Endo TA, et al. A Stress-activated transposon in Arabidopsis induces transgenerational abscisic acid insensitivity. *Sci Rep*. 2016;6:23181.
38. Pietzenuk B, Markus C, Gaubert H, Bagwan N, Merotto A, Bucher E, et al. Recurrent evolution of heat-responsiveness in Brassicaceae COPIA elements. *Genome Biol*. 2016;17:209.
39. Slotkin R, Freeling M, Lisch D. Heritable transposon silencing initiated by a naturally occurring transposon inverted duplication. *Nat Genet*. 2005;37:641–4.
40. Marí-Ordóñez A, Marchais A, Etcheverry M, Martin A. Reconstructing de novo silencing of an active plant retrotransposon. *Nat Genet*. 2013;45:1029–39.
41. Herr AJ, Jensen MB, Dalmay T, Baulcombe DC. RNA polymerase IV directs silencing of endogenous DNA. *Science*. 2005;308:118–20.
42. Peragine A, Yoshikawa M, Wu G, Albrecht H, Poethig R. SGS3 and SGS2/SDE1/RDR6 are required for juvenile development and the production of trans-acting siRNAs in Arabidopsis. *Genes Dev*. 2004;18:2368–79.
43. Blevins T, Pontes O, Pikaard CS, Meins F. Heterochromatic siRNAs and DDM1 independently silence aberrant 5S rDNA transcripts in Arabidopsis. *PLoS One*. 2009;4:e5932–2.
44. Mette MF, van der Winden J, Matzke MA, Matzke AJ. Production of aberrant promoter transcripts contributes to methylation and silencing of unlinked homologous promoters in trans. *EMBO J*. 1999;18:241–8.
45. Ramírez F, Dündar F, Diehl S, Grüning BA, Manke T. deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res*. 2014;42:W187–91.
46. Becker C, Hagmann J, Müller J, Koenig D, Stegle O, Borgwardt K, et al. Spontaneous epigenetic variation in the Arabidopsis thaliana methylome. *Nature*. 2011;480:245–9.
47. Xi Y, Li W. BSMAP: whole genome bisulfite sequence MAPPING program. *BMC Bioinformatics*. 2009;10:232.

Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at  
[www.biomedcentral.com/submit](http://www.biomedcentral.com/submit)



RESEARCH ARTICLE

Open Access



# A transposable element annotation pipeline and expression analysis reveal potentially active elements in the microalga *Tisochrysis lutea*

Jérémy Berthelier<sup>1,2\*</sup> , Nathalie Casse<sup>2</sup>, Nicolas Daccord<sup>3,5</sup>, Véronique Jamilloux<sup>4</sup>, Bruno Saint-Jean<sup>1</sup> and Grégory Carrier<sup>1</sup>

## Abstract

**Background:** Transposable elements (TEs) are mobile DNA sequences known as drivers of genome evolution. Their impacts have been widely studied in animals, plants and insects, but little is known about them in microalgae. In a previous study, we compared the genetic polymorphisms between strains of the haptophyte microalga *Tisochrysis lutea* and suggested the involvement of active autonomous TEs in their genome evolution.

**Results:** To identify potentially autonomous TEs, we designed a pipeline named PiRATE (Pipeline to Retrieve and Annotate Transposable Elements, download: <https://doi.org/10.17882/51795>), and conducted an accurate TE annotation on a new genome assembly of *T. lutea*. PiRATE is composed of detection, classification and annotation steps. Its detection step combines multiple, existing analysis packages representing all major approaches for TE detection and its classification step was optimized for microalgal genomes. The efficiency of the detection and classification steps was evaluated with data on the model species *Arabidopsis thaliana*. PiRATE detected 81% of the TE families of *A. thaliana* and correctly classified 75% of them. We applied PiRATE to *T. lutea* genomic data and established that its genome contains 15.89% Class I and 4.95% Class II TEs. In these, 3.79 and 17.05% correspond to potentially autonomous and non-autonomous TEs, respectively. Annotation data was combined with transcriptomic and proteomic data to identify potentially active autonomous TEs. We identified 17 expressed TE families and, among these, a TIR/Mariner and a TIR/hAT family were able to synthesize their transposase. Both these TE families were among the three highest expressed genes in a previous transcriptomic study and are composed of highly similar copies throughout the genome of *T. lutea*. This sum of evidence reveals that both these TE families could be capable of transposing or triggering the transposition of potential related MITE elements.

**Conclusion:** This manuscript provides an example of a de novo transposable element annotation of a non-model organism characterized by a fragmented genome assembly and belonging to a poorly studied phylum at genomic level. Integration of multi-omics data enabled the discovery of potential mobile TEs and opens the way for new discoveries on the role of these repeated elements in genomic evolution of microalgae.

**Keywords:** Transposable elements, Genome assembly, Pipeline, Tool, Annotation, Algae, Haptophyte, *Tisochrysis lutea*

\* Correspondence: [berthelier.j@laposte.net](mailto:berthelier.j@laposte.net)

<sup>1</sup>IFREMER, Physiology and Biotechnology of Algae Laboratory, rue de l'Île d'Yeu, 44311 Nantes, France

<sup>2</sup>Mer Molécules Santé, EA 2160 IUML - FR 3473 CNRS, Le Mans University, Le Mans, France

Full list of author information is available at the end of the article



## Background

Transposable Elements (TEs) are defined as DNA sequences able to move and spread within eukaryotic and prokaryotic genomes. These repeated elements constitute a variable fraction of eukaryotic genomes, ranging from 3% in the yeast *Saccharomyces cerevisiae*, 45% in human, to 80% in maize [1–3]. TEs were discovered by Barbara McClintock in the late 1940s, refuting the idea that genomes are stable but are, on the contrary, dynamic entities [4]. TEs are highly diverse and an unified classification system for eukaryotic TEs has been proposed, establishing two TE classes according to their transposition mechanisms, structures and similarities [5]. Class I (Retrotransposons) groups elements moving by a copy-paste mechanism through an RNA that is reversed transcribed. Class I is composed of several TE orders, named LTR, DIRS, PLE, LINE and SINE. Class II (DNA transposons) is composed of TEs using different cut-paste mechanisms to transpose. These elements are grouped into the orders TIR, Crypton, Helitron and Maverick. Although intact retrotransposons and DNA transposons are autonomous elements that can move by themselves, SINE elements are non-autonomous TEs and rely on LINE for their mobility, even though their origin is distinct. Other non-autonomous elements can also be distinguished. LTR elements can degenerate into non-coding structures known as LARD (> 4 kbp) or TRIM (< 4 kbp), and TIR elements can also degenerate into non-coding structures known as MITE. LARD, TRIM and MITE elements have intact termini and can thus move by exploiting the molecular machinery of related autonomous TEs [6]. Genomes also contain highly diverged TE fossils, accumulated over time and having no mobility capacity. Due to their mobility, TEs generate mutations in their host genome through new insertions/deletions and participate in genome evolution by impacting the DNA sequence, genome size [7, 8] and chromosome structure [9]. TE activity is known to be triggered during stressful events and, while the majority of transpositions are neutral or harmful to the organisms, transposition events are recognized to promote beneficial mutations [10]. New TE insertions can impact gene function and gene regulation [11]. They can also create new genes and participate in the rise of new phenotypes. The role of TEs has been widely studied in animals [12], land plants [13] and insects [14, 15], but work on their impact on microalgal genomes is only just beginning [16–19]. Microalgae form a diverse polyphyletic group composed of eukaryotic, unicellular and multicellular, photosynthetic organisms [20]. They live in all aquatic habitats whether these have fresh, brackish or salt water and have colonized different extreme habitats, ranging from hot springs, high altitude streams, ice sheets and desert sand crusts, highlighting their

evolutionary ability to adapt to broad range of ecosystems [21–25]. Currently around 150,000 species of algae have been described (<http://www.algaebase.org>), but the number of non-described species is likely to number hundreds of thousands or millions of species [26]. They are divided among different eukaryotic phyla, in Archaeplastidia (green and red lineage), Rhizaria, Alveolates, Stramenopiles (brown lineage), Cryptophytes, Haptophytes and Excavates [27]. Despite their high number and diversity, few genome-wide TE annotations have been performed for microalgae. For the green lineage, this task was realized for ten Chlorophyte species [28–37]. For the red lineage, TE annotation was only done for the Rhodophyte *Cyanidioschyzon* sp. [38]. TEs were annotated in three diatom genomes (brown lineage) [18, 39–41] and also in five dinoflagellate species [42–46]. In Haptophytes, TE annotation has only been performed for one species [47]. These studies reveal that the TE content of microalgae genomes is diverse and includes both retrotransposons and DNA transposons.

Concerning TE activity in microalgae, a few studies have reported evidence of expression or transposition events. Expression of two LTR/Copia families was identified under nitrate starvation or exposure to diatom-derived reactive aldehydes in the diatom species *Thalassiosira pseudonana* and *Phaeodactylum tricornutum* [18]. Moreover, expression of LTR/Copia or TIR/Mariner elements was also reported under thermal stress in *P. tricornutum*, *Amphora acutiuscula*, *Amphora coffeaeformis* and *Symbiodinium microadriaticum* [16, 48–50]. Evidence of transposition events was only identified for a MITE element in a clone of *Chlamydomonas reinhardtii* in the presence of vitamin B<sub>12</sub>, resulting in a new phenotype [17].

Concerning TE activity in Haptophytes, we previously compared genetic polymorphisms between genomes of several strains of *Tisochrysis lutea* [51]. We identified new insertions/deletions and suggested the implication of autonomous TEs in the genome evolution of this species. In this context, the goal of the present study was to inventory TEs in the *T. lutea* genome and to identify potentially autonomous TEs. This marine microalga is commonly used as a feed in aquaculture [52] and is particularly studied for biotechnological applications such as food and biofuel production [53, 54]. In addition, several domesticated strains of *T. lutea* have been obtained with different processes [55] and a large amount of omics data has been collected [51, 56–60].

In this study, we present a detailed TE annotation of the *T. lutea* genome. To achieve this, we designed a new pipeline named PiRATE (Pipeline to Retrieve and Annotate Transposable Elements). The efficiency of the detection and classification steps of PiRATE was evaluated with data of the model species *Arabidopsis thaliana*.

Moreover, to be as exhaustive as possible about the repeated content of *T. lutea*, a new genome assembly was performed by combining Pacific Bioscience and Illumina data. Finally, available transcriptomic and proteomic data were used to reveal potential active TE families.

**Results**

**PiRATE: Pipeline to Retrieve and Annotate Transposable Elements of non-model organisms**

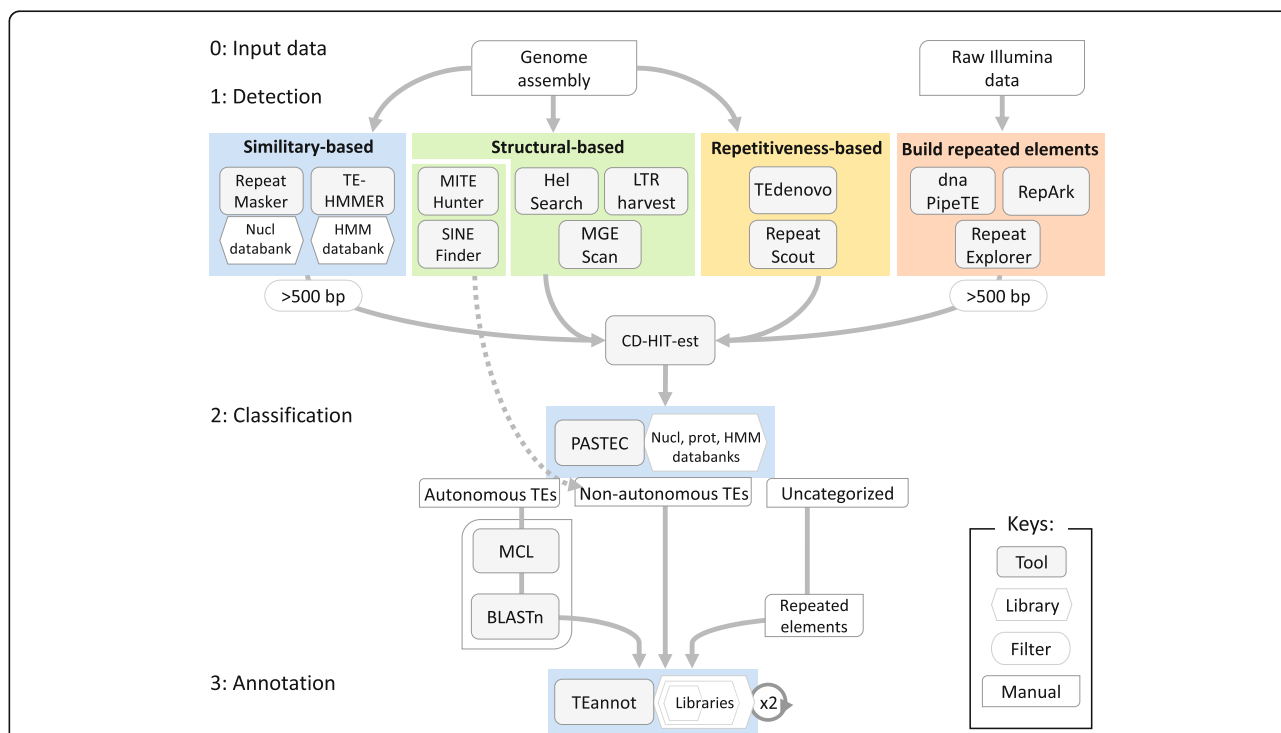
The goal of the present study was to inventory the TE content of the *T. lutea* genome and study the activity of potentially autonomous TEs. Annotation of TEs is a challenging task because of their diversity, their repetitive nature and the complexity of their structures (i.e. GC-rich regions, homopolymers and repeated motifs). Numerous tools have been designed to identify TEs (Additional file 1: Table S1), which can be grouped into four approaches according to their TE detection method: (1) similarity-based detection such as RepeatMasker [61], (2) structure-based detection such as MITE-Hunter [62], (3) repetitiveness-based detection such as

RepeatScout [63], and (4) tools building repeated elements from unassembled data such as dnaPipeTE [64].

Currently, the tool used most frequently to perform a TE annotation is RepeatMasker, which provides a rough estimation of the TE content in a genome assembly [61, 65]. However, this tool compares the genomic sequences with a databank of known TEs to realize the annotation and is therefore not suitable for realizing a de novo TE annotation [65–67]. To perform a de novo TE annotation, pipelines employing repetitiveness-based methods of detection, such as RepeatModeler and REPET, are commonly recommended [66–69]. Here we built PiRATE (Fig. 1) to conduct a de novo TE annotation in the genome of non-model species *T. lutea*. PiRATE is composed of detection, classification and annotation steps.

**Detection of TEs**

To date, genome assembly of non-model organisms has usually not been performed at the level of complete chromosomes but is instead highly fragmented. This fragmentation is recognized to be partly the result of a bad assembly of the TE copies due to their high



**Fig. 1** Overview of the PiRATE pipeline. Step 0: genome assembly and raw Illumina data are used as input data. Step 1: The detection of putative TEs and repeated sequences is performed using 12 tools, combining four detection approaches. Detected sequences from approaches 1 and 4 are filtered according to their length (minimum 500 bp). Detected sequences from the tools MITE-Hunter and SINE-Finder are directly saved as non-autonomous TEs. Other detected sequences are clustered with CD-HIT-est to reduce redundancy. Step 2: Putative TE sequences are automatically classified with PASTEC as potentially autonomous TEs, non-autonomous TEs or uncategorized sequences. The potentially autonomous TEs are manually checked and grouped into TE families. Step 3: Three libraries are manually constructed with a “Russian doll” strategy: 1) a “potentially autonomous TEs library”, a “total TEs library” and a “repeated elements library”. A double-run of TEannot is carried out for each library to select sequences that align with a full-length (FLC) on the genome assembly and finally obtain three independent annotations



repetitive content, which increases the difficulty of their detection [70]. The optimization of the detection step of PiRATE was therefore a priority. We made an overview of tools related to TE detection (Additional file 1: Table S1) and 12 tools were selected according to the specificity and efficiency of their algorithms. These tools represent the four major TE detection approaches (presented above), so as to be as exhaustive as possible. Combining tools is recognised to improve TE detection efficiency [66, 67, 71]. We then applied a clustering method to decrease the redundancy of the detected sequences, by selecting the larger detected sequences of each cluster. The goal of this step was to promote the detection of full-length TE sequences. The detection of complete TE sequences bearing recognizable conserved domains or specific structures and motifs makes the classification step easier. Moreover, a complete TE sequence indicates a potentially autonomous element.

#### Classification of TEs

The classification step of PiRATE is performed by PASTE-TEC [72], which partly uses databanks of known TEs to establish an automated classification of the detected sequences. To improve the classification step of PiRATE, its default databanks were upgraded, by adding 1240 TE sequences from other public databanks, non-inventoried algal TEs and by building 78 new profile HMMs (Hidden Markov Model). Adding non-inventoried data is important for improving the TE classification of species belonging to poorly studied phyla, which often have few described TEs in the databanks. This is common for numerous microalgal phyla (i.e. Haptophyta, Euglenophyta and Dynophyta). In our case, only 17 TE families belonging to the Haptophyte phylum are present in the most frequently used and complete TE databank Repbase [73, 74]. We also estimated that only 2609 TE families are described for microalgal taxa in Repbase. Compared with other taxa, this number is very low, for examples 29,503 TE families are described for Metazoa and 12,620 for Viridiplantae (Repbase, 10/29/2017). The putative TE sequences are classified following the Wicker et al. classification [5] and can be grouped as 1) potentially autonomous TEs, 2) non-autonomous TEs or 3) uncategorized sequences. Because we were interested in potentially autonomous TEs, these sequences were manually checked and grouped into families.

#### Annotation of TEs

For the annotation step, we built three libraries in order to then apply a method that we named “Russian doll”, due to its nesting strategy (Additional file 1: Figure S1). We built a “potentially autonomous TEs library” containing checked potentially autonomous TEs, a “total TEs library” also containing the non-autonomous TEs,

and a “repeated elements library” also containing the uncategorized repeated sequences. These nested libraries made it possible to perform several independent annotations in order to avoid a competition effect among sequences aligning on the same genomic regions.

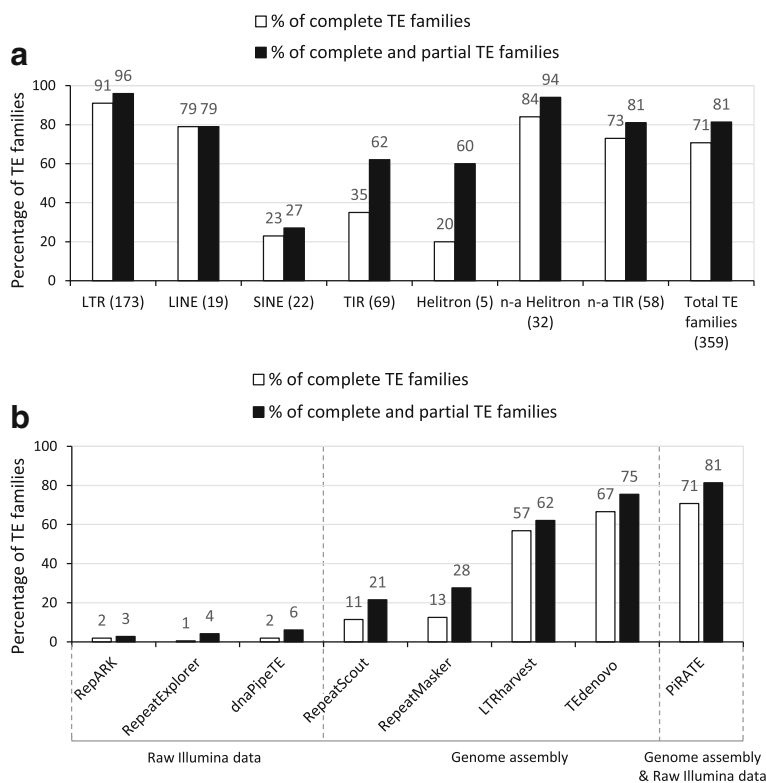
#### Evaluation of PiRATE with *A. thaliana* genomic data

##### Evaluation of the detection step

The detection and classification steps of PiRATE were evaluated to highlight their strengths and weaknesses. This evaluation made it possible to define suitable rules for the manual check step. As a control, we used 359 consensus sequences of the described TE families of *A. thaliana*, available in Repbase. Genomic data of the model plant *A. thaliana* provided a suitable control because of its high quality genome assembly and high TE diversity. Class I and Class II *A. thaliana* TE families are well described for both autonomous and non-autonomous TEs. Detected sequences covering less than 40% of the full-length of a consensus sequence were considered too short to be efficiently classified and were not taken into account. The proportion of TE families detected with a complete length (coverage score  $\geq 70\%$ ) or detected with at least a partial length (coverage score  $\geq 40\%$ ) is given in Fig. 2a. PiRATE detected  $\sim 81\%$  (292/359) of the TE families described in *A. thaliana* genome (Fig. 2a). PiRATE was especially effective for detecting sequences belonging to LTR (96%), LINE (79%), non-autonomous TIR (81%) and non-autonomous Helitron (94%) (Fig. 2a). It had a good efficiency for detecting TIR (62%) and Helitron (60%). However, it had difficulty detecting SINE elements (27%) (Fig. 2a). In addition, we compared the detection step of PiRATE to TEdenovo [68], LTRharvest [75], RepeatScout [63], RepeatMasker [61], dnaPipeTE [64], RepeatExplorer [76] and RepARK [77] (Fig. 2b). Overall, the detection step of PiRATE detected the highest percentage of TE families of *A. thaliana*. Compared to TEdenovo, which displayed the second highest percentage of detected TE families, PiRATE detected 21 additional TE families (+ 6%) (Fig. 2b and Additional file 1: Figure S2). PiRATE was particularly more effective for detecting LINE (+ 32%) and TIR (+ 10%) (Additional file 1: Figure S2).

##### Evaluation of the classification step

To evaluate the classification step of PiRATE, we used the 292 sequences detected by PiRATE during the evaluation of the detection step, which represent the largest detected sequences of the 292 TE families of *A. thaliana*. These 292 sequences were classified with PASTE-TEC using the PiRATE databanks (excluding data from *Arabidopsis* species). To estimate the classification efficiency, we counted the number of detected TEs with correct classification at the order level and the number of sequences



**Fig. 2** Evaluation of the detection step of PiRATE with genomic data of *Arabidopsis thaliana*. **a**) Percentage of TE families detected in *A. thaliana*. For each TE order (x-axis) is indicated the percentage of TE families detected with a complete length (coverage score  $\geq 70\%$ , white bars) or detected with a partial and a complete length (coverage score  $\geq 40\%$ , black bars). The x-axis indicates the number of TE families for each order; “n-a” means non-autonomous. **b**) Comparison of the percentage of TE families of *A. thaliana* detected by PiRATE (Step 1), RepARK, RepeatExplorer, dnaPipeTE, RepeatScout, RepeatMasker, LTRharvest and TEdenovo. For each tool is indicated the percentages of TE families of *A. thaliana* detected with a complete length (coverage score  $\geq 70\%$ , white bars) or detected with a partial and a complete length (coverage score  $\geq 40\%$ , black bars). The x-axis indicates the tools and nature of the input data

that had an incorrect classification or that were uncategorized. We observed that 75% (218/292) of the detected TEs were correctly classified, 7% (21/292) were incorrectly classified and 18% (53/292) were uncategorized. The classification step of PiRATE was therefore efficient at correctly classifying autonomous TEs belonging to LTR (98%), LINE (87%), TIR (91%) and Helitron (100%) but had difficulty correctly classifying SINE (50%), non-autonomous TIR (27%) and non-autonomous Helitron (7%) (Additional file 1: Figure S3). Taking into account all of the above results, PiRATE is efficient enough to detect and correctly classify the majority of the autonomous TE families.

#### A new genome assembly of *T. lutea* to improve the TE annotation

We recently published a draft genome assembly of *T. lutea* obtained with Illumina short-read technology [51]. To obtain an improved genome assembly, the genome of *T. lutea* was re-sequenced with Pacific Bioscience long-read technology. A new genome assembly was

performed from the long reads and was improved with the Illumina short-read data, used to build the draft genome assembly [51]. The new genome assembly of *T. lutea* is composed of 193 contigs and is 82 Mb in size. A gain of around 30 Mb was obtained (+ 34%), compared with the previous 54 Mb genome assembly, which was composed of 7659 contigs [51]. The size of the coding regions increased slightly between these genome versions. While the new genome assembly encodes for 15,972 genes, corresponding to a coding region length of 32 Mb, the gene proportion of the previous draft genome version was 25 Mb, suggesting that the new assembled regions are mostly repeated elements. This new larger version of the genome seems to incorporate more assembled TEs.

#### Effect of genome quality on TE detection approaches

To estimate the contribution of each TE detection approach of PiRATE depending on the level of fragmentation of the genome assembly, the detection step (Fig. 1) of PiRATE was applied with raw Illumina data of *T.*

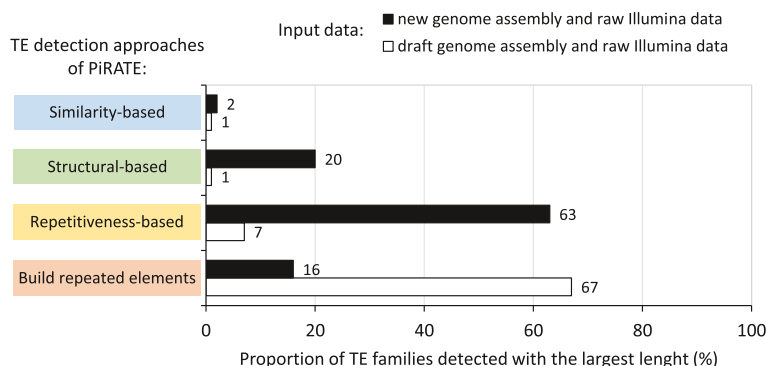
*lutea* and, either the draft genome version of *T. lutea* (7659 contigs) [51] or the new genome assembly of *T. lutea* (193 contigs). In both cases, the detected sequences were compared to the referent sequences of the TE families of *T. lutea* (described below). For each TE detection approach in PiRATE, we counted the number of *T. lutea* TE families detected, with the largest length (i.e. the most complete sequences, having the highest percentage of coverage compared to the reference TE sequences) and divide this number by the total of detected TE families. This provided an estimation ratio of the contribution of each TE detection approach depending on the input data (Fig. 3). With both types of dataset, the similarity-based approach had the weakest percentage and contributed to detecting only 2 or 3% of the *T. lutea* TE families. Using the draft genome assembly and the raw Illumina data, the structural-based approach contributed to detecting 1% of the TEs families of *T. lutea*, but 20% of the TE families of *T. lutea* with the new genome assembly and the raw Illumina data (Fig. 3). The repetitiveness-based approach contributed to detecting 7% of the TE families of *T. lutea* with the draft genome assembly and the raw Illumina data. However, it was the most efficient approach with the new genome and contributed to detecting 63% of the *T. lutea* TE families (Fig. 3). When a draft genome assembly is used as input, the fourth detection approach, using raw Illumina data to build repeated elements, was the most useful approach and contributed to detecting 67% of the TE families (Fig. 3).

**Annotation of the repeated elements content of the *T. lutea* genome**

We applied PiRATE to the new genome assembly of *T. lutea* and raw Illumina data. After the classification step,

we manually curated the sequences as potentially autonomous TEs, non-autonomous TEs or uncategorized repeated elements. Because we were interested in characterizing their activity, the potentially autonomous TEs were manually checked and grouped into families (see Methods). We identified six potentially autonomous families of LTR/Copia and four families of LTR/Gypsy (Table 1). We found 14 potentially autonomous families of LINE elements, similarly close to Tx1 elements, belonging to the L1 superfamily [78, 79]. We identified seven potentially autonomous families of TIR/Harbinger, six families of TIR/PiggyBac and eight families of TIR/Mariner. A high number of potentially autonomous hAT elements were detected. Due to their divergence, they were grouped into 129 putative families.

Three annotations were conducted with three nested libraries (Additional file 1: Figure S1). From the “potentially autonomous TEs library” composed of 240 referent sequences, we estimated that the proportion of the potentially autonomous TEs represent 3.79% of the *T. lutea* genome (Table 1). The annotation of the TE content was performed with the “total TEs library” containing 459 supplementary sequences corresponding to 14 sequences of potential SINE elements, 188 sequences of potential MITE, 240 sequences of potential TRIM and 17 sequences of potential LARD (Table 1). From this annotation, we estimated that the genome of *T. lutea* contains 20.84% of potentially autonomous and non-autonomous TEs (Table 1 and Additional file 1: Figure S4). Class I and Class II TEs represent 15.89 and 4.95%, respectively (Table 1). We found a large quantity of Gypsy (4.65%), LINE (3.87%) and hAT (2.12%) copies, suggesting ancient burst events for these elements (Table 1). We established that the proportion of non-autonomous TEs is 17.05% (Table 1). Then, we performed the annotation



**Fig. 3** Comparison of the contribution of the four TE detection approaches of PiRATE on the detection of the TE families of *Tisochrysis lutea*, depending on the input data. For each TE detection approach, we calculated the number of TE families detected with the largest length and divide this number by the total of detected TE families of *T. lutea*. The input dataset was either the draft genome assembly of *T. lutea* and raw Illumina data of *T. lutea* (white bars) or the new genome assembly of *T. lutea* and raw Illumina data of *T. lutea* (black bars). The similarity-based detection, structural-based detection and the repetitiveness-based detection use a genome assembly as input data. The last approach builds repeated elements from raw Illumina data



**Table 1** Diversity and proportion of transposable element orders and classes in the genome assembly of *Tisochrysis lutea*. The abbreviations “a” and “n-a” indicate autonomous and non-autonomous transposable elements respectively

|                |     | Orders/ Superfamilies | Number of families (f) or detected sequences (s) | Number of potentially autonomous TEs | Proportion of the potentially autonomous TEs (%) | Proportion of total genome (%) |
|----------------|-----|-----------------------|--|--------------------------------------|--|--------------------------------|
| Class I        | a   | LTR/Copia             | 6 f  | 45                                   | 0.37   | 1.09                           |
|                |     | LTR/Gypsy             | 4 f  | 242                                  | 2.56   | 4.65                           |
|                |     | LINE/L1               | 14 f   | 59                                   | 0.25   | 3.87                           |
|                | n-a | SINE                  | 14 s   |                                      |  | 0.04                           |
|                |     | LTR/LARD              | 17 s   |                                      |  | 0.76                           |
|                |     | LTR/TRIM              | 240 s  |                                      |  | 5.48                           |
| Total Class I  |     |                       |  |                                      |  | 15.89                          |
| Class II       | a   | TIR/hAT               | 129 f  | 145                                  | 0.41   | 2.12                           |
|                |     | TIR/Mariner           | 8 f  | 41                                   | 0.11   | 0.19                           |
|                |     | TIR/Harbinger         | 7 f  | 26                                   | 0.05   | 0.34                           |
|                |     | TIR/PiggyBac          | 7 f  | 14                                   | 0.04   | 0.26                           |
|                | n-a | MITE                  | 188 s  |                                      |  | 2.04                           |
| Total Class II |     |                       |  |                                      |  | 4.95                           |
| Total TEs      |     |                       |  | 572                                  | 3.79   | 20.84                          |

of every repeated element by using the “repeated elements library” containing an additional 2680 uncategorized repeated sequences. From this annotation, we estimated that 17.79% of the *T. lutea* genome is represented by uncategorized repeated elements (Additional file 1: Figure S4). To estimate the proportion of the simple tandem repeats, we used the tool RepeatMasker and found that they made up 5.97% of the genome assembly of *T. lutea* (Additional file 1: Figure S4). By adding together the proportions of all the annotated repeats, we estimated that the total proportion of repeated elements in the *T. lutea* genome was 44.6%. Knowing that the coding gene proportion is of 38.49%, we estimated that 16.91% of the genome is non-characterized (Additional file 1: Figure S4).

#### Discovery of potentially active autonomous TEs in the *T. lutea* genome

In this study we chose to focus on the identification of potentially autonomous TEs to reveal potentially active elements. From the annotation obtained with the “potentially autonomous TEs library”, we performed the cartography of the 572 annotated TEs that are potentially autonomous (Fig. 4).

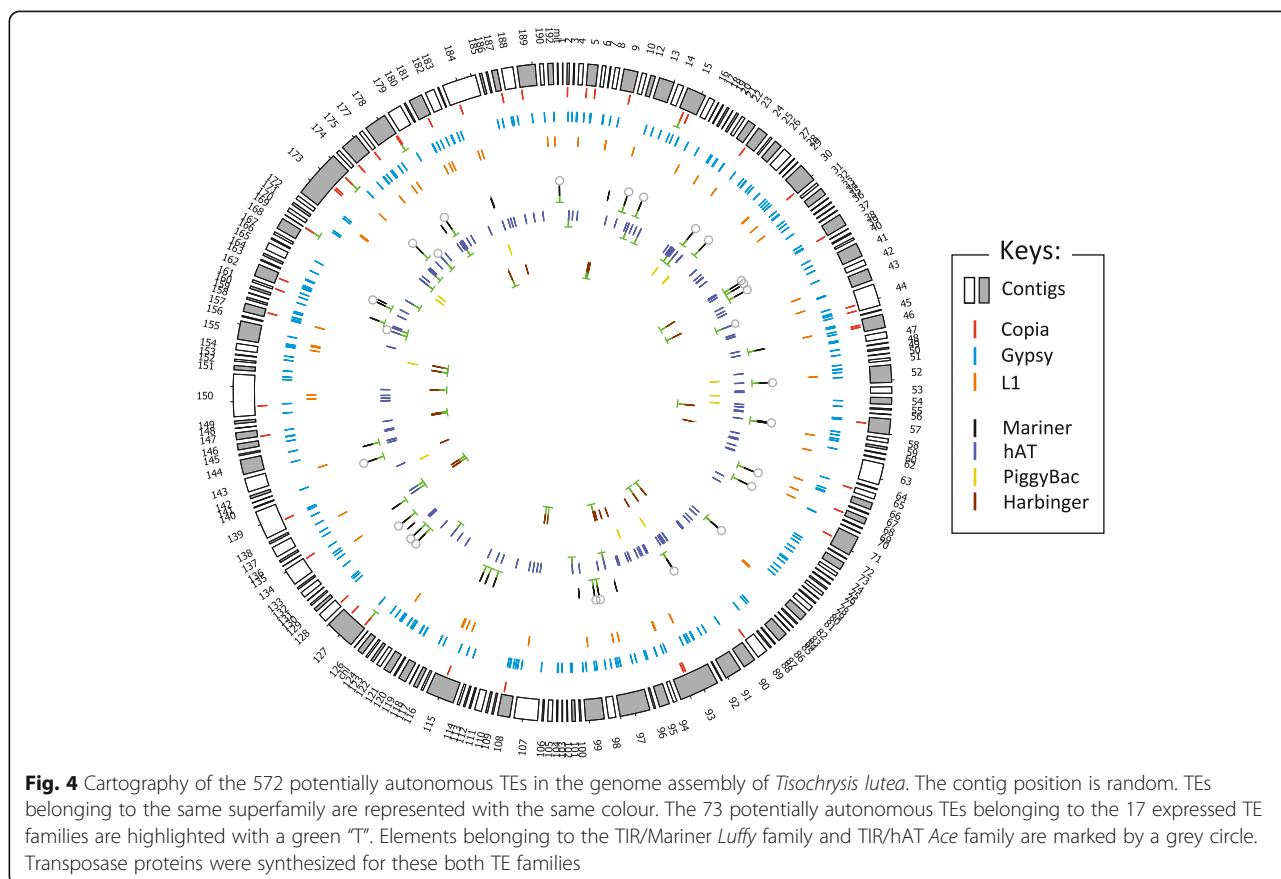
To identify potentially active TEs and have an estimation of the genome dynamic of *T. lutea*, transcriptomic data were mapped on the new genome assembly and crossed with the annotation of the 572 potentially autonomous TEs. Expression was identified for 17 TE families: one LTR/Copia, four TIR/Mariner, four TIR/Harbinger and eight TIR/hAT. These families represent 73 potentially autonomous TEs and their genomic position is illustrated in Fig. 4 and is indicated in

Additional file 2. Putative ancient transpositions were studied by looking for similarities between copies belonging to these 17 expressed TE families (Additional file 2). We identified that the Mariner-3 family is composed of 24 highly similar copies, which share a mean pairwise identity of 99.7%. Among them, 20 copies seem to be complete (Additional file 3). This high number of similar copies suggests that this family was/is active. The hAT-2 family is composed of three highly similar copies that share a mean pairwise identity of 99.8%. Moreover, eight similar copies were identified for the Harbinger-6 family and five similar copies for the Copia-3 family. Other details can be found in Additional file 2. TE copies belonging to these 17 expressed TE families were submitted to BLASTx on proteomic data of *T. lutea*, that we previously obtained under nitrogen limitation [58]. We identified that transposase proteins were synthesized for the Mariner-3 family and the hAT-2 family. The transposases of the Mariner-3 and hAT-2 families match with six and 36 peptides, respectively. The alignments with the matching peptides can be found for both families in Additional file 4. From transcriptomic data of a previous study, we highlight that these families were among the three higher expressed genes [58].

#### Discussion

##### PiRATE: Pipeline to Retrieve and Annotate Transposable Elements of non-model organisms

The goal of the present study was to inventory the TE content in the genome of *T. lutea* genome and study the activity of potentially autonomous TEs. We built PiRATE to counter the lack of knowledge about TEs in Haptophytes and the difficulty of identifying TEs in a



fragmented genome assembly [70]. The detection step of PiRATE has been optimized to promote the detection of full-length TE sequences and its classification step has been improved for algal genomes. The detection step of PiRATE combines multiple, existing analysis packages representing all major approaches for TE detection. The detection step of PiRATE was evaluated with genomic data of *A. thaliana* and compared to TEdenovo [68], LTRharvest [75], RepeatScout [63], RepeatMasker [61], dnaPipeTE [64], RepeatExplorer [76] and RepARK [77] (Fig. 2b). Overall, the detection step of PiRATE detected the highest percentage of TE families (81%) with a partial and complete length compare to the other tools used alone (Fig. 2b). This confirms that the combining of multiple tools, using several approaches improves the detection of different TE families, with complete sequences, as previously indicated [66, 67, 71]. In this comparison, TEdenovo was efficient and displayed the second highest percentage of detected TE families (75%) (Fig. 2b). LTRharvest also showed a good capacity to detect TE families of *A. thaliana* (62%) (Fig. 2b). This is due to the high content of LTR elements in the *A. thaliana* genome and because this tool detected TE families belonging to other TE orders. In this comparison, the least effective tools were RepARK (3%), RepeatExplorer

(4%) and dnaPipeTE (6%), which used raw illumina data as input (Fig. 2b). This is not surprising considering the challenge of building repeated elements from raw Illumina data, compared to the other tools using the complete genome assembly of *A. thaliana*.

**A new genome assembly of *T. lutea* to improve the TE annotation**

We recently published a draft genome assembly of *T. lutea* obtained with Illumina short-read technology [51]. While this technology has a very low sequencing error rate, its use alone often leads to fragmented assemblies, especially in TE-rich genomes, due to the incapacity of short-reads to entirely span repetitive elements [80]. To obtain an improved genome assembly, the genome of *T. lutea* was re-sequenced with Pacific Bioscience long-read technology and the assembly was corrected with short-read Illumina data. Indeed, the use of long-reads leads to a more complete and accurate assembly of long repeated elements such as TEs [81–83]. However, to date, this technology has a high sequencing error rate and its combination with short-read Illumina data has become a common way of partially overcoming this problem [84–86]. Compare to the previous draft genome

assembly, this new genome assembly is larger, less fragmented and seems to incorporate more assembled TEs.

#### Effect of genome quality on TE detection approaches

To estimate the contribution of the four TE detection approaches of PiRATE depending on the level of fragmentation of the genome assembly, the detection step (Fig. 1) of PiRATE was applied with raw Illumina data of *T. lutea* and, either the draft genome version of *T. lutea* (7659 contigs) [51] or the new genome assembly of *T. lutea* (193 contigs). The four TE detection approaches showed different contribution according to the level of fragmentation of the genome assembly (Fig. 3). By gathering these four detection approaches, PiRATE improves the TE detection of organisms having a genome assembly which is highly fragmented.

#### Annotation of the repeated elements content of the *T. lutea* genome

With PiRATE, we established that the total proportion of repeated elements in the *T. lutea* genome is represented by 20.84% of TEs, 17.79% of uncategorized repeated elements and 5.97% of simple tandem repeats (Additional file 1: Figure S4). The high percentage of uncategorized repeated elements could indicate the presence of unknown TEs. A high number of uncategorized sequences (30.9%) was also reported in the *Emiliania huxleyi* genome [40]. Here, we choose to focus on the identification of potentially autonomous TEs to reveal potentially active elements. The proportion of the potentially autonomous TEs represents 3.79% of the *T. lutea* genome, corresponding to 572 annotated TEs (Fig. 4). Interestingly, we found a potentially autonomous TIR/Mariner in the predicted mitochondrial genome and a potentially autonomous LTR/Copia and TIR/hAT in the predicted chloroplast genome.

#### Identification of potentially active TEs in *T. lutea*

Few studies have investigated TE activity in microalgal genomes and their role is poorly known. Regarding Class I TEs, some studies reported expression of LTR elements in dinoflagellate and diatom species under thermal stress or nitrogen limitation [16, 48–50]. Concerning Class II elements, a previous study reported a case of phenotypic evolution for the microalga *Chlamydomonas reinhardtii* caused by the transposition of a MITE in the presence of vitamin B<sub>12</sub> [17]. In the present study, we identified 17 expressed TE families and, among these, a TIR/Mariner *Luffy* and a TIR/hAT *Ace* family were able to synthesize their transposase under nitrogen starvation [58]. We highlight the presence of highly similar copies (Additional file 3) suggesting that these elements are able to transpose or could be able to trigger the transposition of potential derived MITE elements. Although

we cannot draw conclusions about their mobility, the investigation of the TE expression is a good indicator of the potential activity of TEs. Nitrogen limitation has been previously described as a stress condition in the diatom *Phaeodactylum tricornutum*, triggering overexpression of the LTR/Copia family named *Blackbeard* [18]. Although we cannot draw conclusions about de novo insertions, the evidence presented here indicates that these both TEs families are suitable candidates for mobility and could participate in the genome evolution of *T. lutea*.

#### Conclusion

Genome-wide TE annotation has rarely been performed in microalgae compared with animals, insects and land plants. This study opens the way to new searches about the role of TEs in the genome evolution of *Tisochrysis lutea* and their contribution to the microalgal adaptation process. In the present study, we built PiRATE to counter the lack of knowledge about TEs in Haptophytes and the difficulty of identifying TEs in a fragmented genome assembly. With PiRATE, we conducted a genome-wide detection and annotation of the repeated elements in a new genome assembly of *Tisochrysis lutea* and established that it is composed of 3.8 and 15.95% of potentially autonomous and non-autonomous TEs, respectively. The annotation of the potentially autonomous TEs was crossed with transcriptomic and proteomic data and evidence of expression was identified for 17 TE families. Among these, we discovered that transposase proteins were synthesized for both a Mariner (*Luffy*) and a hAT (*Ace*) family. Both these families have several highly similar copies throughout the genome and were among the three highest expressed genes in a previous transcriptomic study. All of this suggests that both these families could be able to transpose themselves or trigger the transposition of potential derived MITE elements.

#### Methods

##### Microalga strain and culture conditions

The *T. lutea* strain was provided by the Culture Collection of Algae and Protozoa (CCAP 927/14). This strain was isolated by Haines in the late 70s and stored in the algae bank. The strain was grown in two 1-L flasks, bubbled with 0.22 mm filtered-air. The culture was maintained at a constant temperature of 21 °C, under a constant irradiance of 50  $\mu\text{mol m}^{-2} \text{s}^{-1}$ .

##### DNA extraction, sequencing, genome assembly and gene annotation

Total DNA was extracted from the *T. lutea* WT-strain using a phenol/chloroform protocol. DNA quality and concentration were assessed with gel electrophoresis and

Qubit Fluorometric Quantitation (ThermoFisher, Massachusetts, USA), respectively. *T. lutea* genome sequencing was performed with a PacBio RSII sequencer (Pacific Bioscience, California, USA) at the Plateforme GeT PlaGe (Toulouse, France); seven SMRT cells were performed. Filtered subreads were assembled using Canu1.3 [82]. The assembly was polished with Quiver (<https://github.com/PacificBiosciences/GenomicConsensus>) and its accuracy was improved with Pilon [87] using previous Illumina Hiseq mate-pair reads of *T. lutea* ([51]; SRA: SRR3156597). The annotation of the coding-gene region was performed with the pipeline MAKER2 [88–91].

### TE annotation in the *T. lutea* genome using PiRATE

#### Step 1: TE detection

The new genome assembly of *T. lutea* and previous raw Illumina data ([51]; SRA: SRR3156597) were used as input. Putative TE sequences were detected using four approaches (Fig. 1). The first approach was represented by two tools using similarity-based detection: RepeatMasker (setting: -s, -no\_low, -lib; with the PiRATE nucleotide databank; [61]) and TE-HMMER (with a homemade profile HMMs databank). TE-HMMER is a homemade tool using HMMER (default setting, [92]) and tBLASTn (setting: -evaluate 10E-300, [93]). The second approach consisted of five tools using structural-based detection: LTRharvest (default setting, [75]), Helsearch (default setting, [94]), MGEScan-nonLTR (default setting, [95]), MITE-Hunter (default setting, [62]) and SINE-Finder (default setting, [96]). The third approach combines tools using repetitiveness-based detection: TEde novo (steps 1 to 4, default setting, [68]) and Repeat Scout (default setting, [63]). These tools cluster repeated sequences from a genome assembly to build consensus sequences. The last approach was composed of tools performing the assembly of repeated sequences from raw Illumina data (fasta or fastq). We used RepARK (default setting, [77]), dnaPipeTE (setting: %coverage: 0.6, [64]) and RepeatExplorer (setting: -paired, [76]). The sequences detected by the first and the last approaches that were below 500 bp in length were removed with a perl script. The sequences detected with SINE-Finder and MITE-Hunter were directly saved for the second step. Other detected sequences were concatenated into a single FASTA file and clustered with CD-HIT-est (settings: -aS 1 -c 1 -r 1 -g 1 -p 0, [97]) to reduce the redundancy. This made it possible to remove shorter sequences that aligned with 100% of identity on a part of the larger sequences.

#### Step 2: TE classification

In the second step, sequences were automatically classified with PASTEC [72], following the Wicker et al.

classification system [5]. This tool was improved with custom databanks (described below). Three libraries were manually constructed with a “Russian doll” strategy in order to perform separate annotations (Additional file 1: Figure S1): a “potentially autonomous TEs library”, a “total TEs library” containing the potentially autonomous TEs and the non-autonomous TEs and a “repeated elements library” also containing the uncategorized repeated sequences. Sequences classified as LTR, LINE and TIR were manually sorted by superfamily (according to the evidence section produced by PASTEC). To facilitate their manual check, sequences belonging to the same putative superfamily were grouped into families with MCL (MCL\_inflation: 1.5; MCL\_coverage: 0). The percentage of identity between sequences belonging to the same family were checked with Blastn (-identity: 80%). We followed the 80–80–80 Wicker rules to form families [5]. Finally, larger sequences from each TE family were checked and selected for the “potentially autonomous TEs library” according to the presence of TE domains or similarities with Pfam (<http://pfam.xfam.org>), NCBI-BLASTx and Censor (<http://www.girinst.org/censor>). We defined as potentially autonomous LTR, sequences bearing at least a reverse transcriptase and an integrase domain and having similarity to known LTR elements. We defined as potentially autonomous LINE, sequences bearing at least a reverse transcriptase domain and sharing similarity to known LINE elements. We defined as potentially autonomous TIR, sequences with evidence of a transposase domain or similarity with known TIR elements.

No manual checks were performed for sequences classified as non-autonomous TEs. Sequences classified as SINE, MITE and TRIM were directly selected for the “total TEs library”. Only sequences classified as LARD, which were obtained with the repetitiveness-based approach of TE detections (TEde novo or Repeat Scout), were selected. Sequences detected by SINE-Finder and MITE-Hunter were also directly selected for the “total TEs library”. Finally, the sequences classified as noCat (uncategorized) and obtained with the repetitiveness-based approach at the TE detection step were selected for the “repeated elements library”.

#### Step 3: TE annotation

Three libraries were built (Additional file 1: Figure S1): 1) a “potentially autonomous TEs library” 2) a “total TEs library” and 3) a “repeated elements library”. A first run of TEannot ([68], default setting, steps 1, 2, 3, 7 and 8) was performed for each library to known sequences matching with a full-length size on the genome (FLC sequences) and to remove potential chimeric data. A second run of TEannot was performed with these FLC sequences for each of the final libraries (default setting,



steps 1, 2, 3, 4, 5, 7 and 8) and three annotations were obtained.

#### Proportion of TEs and repeated elements in *T. lutea*

From the annotation file obtained with the “potentially autonomous TEs library”, we manually selected 572 sequences and calculated their proportion in the genome of *T. lutea*. TEs. The different criteria used are detailed in Additional file 1: Method S1 and Table S2. An illustration of the position of these sequences on the *T. lutea* genome assembly was built with the tool Circos [98]. The annotations obtained with the “total TEs library” and the “repeated elements library” were used to estimate the total proportion of TEs and to calculate the proportion of uncategorized repeated elements in the genome of *T. lutea*. Details on the method are available in Additional file 1: Method S2 and Table S3. The proportion of simple repeats was calculated with the tool RepeatMasker (setting: -s -noint -no\_is, [61]).

#### PiRATE databanks

##### Nucleotide and protein databanks

A nucleotide and a protein databank of TEs were built with sequences from Repbase (REPET version 20.05, <http://www.girinst.org/repbases>), the P-MITE database (<http://pmite.hzau.edu.cn>) and SINE base (<http://sines.eimb.ru>). Because algae originally arose from the predation of a cyanobacterial organism by a eukaryotic heterotrophic organism, cyanobacterial TE sequences were also added from the IS-finder database (<http://www-is.bio-toul.fr>) (Additional file 5). Moreover, we added non-inventoried TEs of microalgae and macroalgae, retrieved from the NCBI database (Additional file 5).

##### Profile HMMs databank

A homemade databank of profile HMMs was built with sequences of the protein databank. Multiple protein alignments were performed with Clustal Omega (<http://www.ebi.ac.uk/Tools/msa/clustalo/>). When possible, TE protein sequences from algae were favoured. 78 profile HMMs were performed with the HMMbuild tool of HMMER [92] for 62 TE categories displayed on the Browse Repbase tool (<http://www.girinst.org/repbases/update/browse.php>). This databank was used with TE-HMMER at the detection step. At the classification step, we combine this databank with the default databank of PASTEC (ProfilesBankForREPET\_Pfam27.0\_GypsyDB.hmm, <https://urgi.versailles.inra.fr/download/repet>).

#### Evaluation of PiRATE

The efficiency of the detection and classification steps of PiRATE were evaluated with genomic data of the model plant *A. thaliana*. We used the genome assembly TAIR10 available on the TAIR project (<https://www>.

[arabidopsis.org/download/index-auto.jsp?dir=%2Fdownload\\_files%2FGenes%2FTAIR10\\_genome\\_release%2FTAIR10\\_chromosome\\_files](http://arabidopsis.org/download/index-auto.jsp?dir=%2Fdownload_files%2FGenes%2FTAIR10_genome_release%2FTAIR10_chromosome_files)) and the raw Illumina data available at the 1001 genome project <http://1001genomes.org/data/SLU/SLUHenning2014/releases/current/strains/Seattle-0>). These data of *A. thaliana* were submitted to the step 1 of PiRATE (RepeatMasker and TE-HMMER were used without data from *Arabidopsis* species in the databanks). The detected sequences were submitted to PASTEC [72] and compared to the 359 TE families described in *A. thaliana* and available on Repbase (<http://www.girinst.org/repbases>). We didn't include the terminal repeated sequences of the LTR TE families and the heterologous TE named DRL1. From the classification file, we selected each of the sequences matching to a TE consensus sequences of *A. thaliana*. Those covering less than 40% of the full-length of a consensus sequences were considered as too short to be efficiently classified and were not taken in account. We considered as a partial or complete detection the detected sequences covering at least 40% or 70% of the full-length of a consensus TE family of *A. thaliana*, respectively. For the comparison of the detection step of PiRATE with TEdenovo [68] (steps 1 to 4, with LTRharvest [75]), LTRharvest [75], RepeatScout [63], RepeatMasker [61], dnaPipeTE [64], RepeatExplorer [76] and RepARK [77], the number of detected TE families was calculated with the same method previously described for the evaluation of the PiRATE detection step. For the evaluation of the classification step of PiRATE, we used the longest detected sequences of the 292 TE families detected by PiRATE during the evaluation of the detection step as a control. These 292 sequences were classified with PASTEC using modified versions of the three PiRATE databanks (Nucleotide, protein and profile HMMs), without data from *Arabidopsis* sp. We calculated the percentages of correct classification, incorrect classification or uncategorized classification. Details on the impact of the genome assembly quality on the efficiency of the TE detection step of PiRATE are available in Additional file 1: Method S3.

#### Transcriptomic and proteomic analyses

The expression analysis was performed using eight sets of previously published transcriptomics data [56, 58]. These data were concatenated and normalized using the tool `insilico_read_normalization.pl` of Trinity [99]. Reads were then mapped on the new genome assembly of *T. lutea* with TopHat [100] and crossed with the annotation of the potentially autonomous TEs. HTseqCount [101] was used to count the number of mapped reads for each potentially autonomous TEs. With a homemade script we retrieved the TE families with transcripts covering at least 90% of the annotated sequences. Sequences

of the TE copies of these expressed TE families were then compared with BLASTx to published proteomic data [58]. Sequence alignments of the peptides of the Mariner (*Luffy*) and hAT (*Ace*) elements on the predicted transposases were performed with ClustalOmega and visualized with Geneious (Additional file 4). With the global-alignment tool of Geneious [102] (setting: free end gaps), a mean pairwise identity was calculated for each expressed TE family having at least three annotated copies (Additional file 3).

### PiRATE is automated through a stand alone Galaxy

All tools used in PiRATE are automated in a standalone Galaxy [103]. The PiRATE-Galaxy is available through a virtual machine at <https://doi.org/10.17882/51795>. A tutorial file can be download.

### Additional files

**Additional file 1:** Additional supporting information. This file contains the additional supporting figures, tables, results and materials and methods. (PDF 594 kb)

**Additional file 2:** Percentage of identity between copies of the expressed TE families. This file lists the percentage of identity between the TE copies of the 17 expressed TE families identified in the genome of *Tisochrysis lutea*. (XLSX 19 kb)

**Additional file 3:** Sequences alignment of TE copies of the TIR/Mariner *Luffy* family. This file contains the sequence alignment of the copies belonging to the TIR/Mariner *Luffy* described in the genome of *Tisochrysis lutea*. (PDF 17427 kb)

**Additional file 4:** Sequences alignment of the peptides matching on the predicted TE proteins. This file contains the alignment of the peptides matching on the predicted proteins of the TIR/Mariner *Luffy* and the TIR/hAT *Ace*. (PDF 996 kb)

**Additional file 5:** List of non-inventoried sequences added to the databanks used by the pipeline PiRATE. This file lists the non-inventoried sequences added to the databanks used by the pipeline PiRATE, they belong to algae and cyanobacteria. (XLSX 23 kb)

### Abbreviations

HMM: Hidden Markov Model; LARD: Large Retrotransposon Derivative; LINE: Long Interspersed Element; LTR: Long Terminal Repeat; MITE: Miniature Inverted-repeat Transposable Element; PiRATE: Pipeline to Retrieve and Annotate Transposable Element; PLE: Penelope; SINE: Short Interspersed Nuclear Element; TE: Transposable element; TIR: Terminal Inverted Repeat; TRIM: Terminal-repeat Retrotransposons In Miniature

### Acknowledgements

This work was supported by the French Region of Pays de la Loire with the Atlantic Microalgae program and the French Research Institute for Exploitation of the Sea (IFREMER). We thank the platform Genotoul GeT-PlaGe for the genome sequencing of *T. lutea*. We also thank the URGI Team for their advice about REPET as well as Jonathan Filée and Etienne Bucher for their advice on this study. We thank Helen McCombie for the proofreading. The authors are grateful to the anonymous reviewers for their critical comments, which have greatly improved the manuscript.

### Funding

This work was supported by the French region of Pays de la Loire and the French Research Institute for Exploitation of the Sea (IFREMER).

### Availability of data and materials

Datasets relating to the identification of TEs, as well as the improved genome assembly of *T. lutea*, are available at <https://doi.org/10.17882/52231>. The virtual machine of the PiRATE-Galaxy and the tutorial are available at <https://doi.org/10.17882/51795>.

### Authors' contributions

JB developed PiRATE, carried out the TE annotation and the expression analyses, participated in the genome assembly and drafted the manuscript. NC participated in the coordination and contributed to draft the manuscript. ND designed the pipeline for the genome assembly, contributed for the analysis of the genome assembly and contributed to draft the manuscript. VJ contributed to parameter the classification and the annotation steps of PiRATE (PASTEC and TEannot), contributed to interpret the classification and annotation results and contributed to draft the manuscript. BSJ participated in the coordination of the study, the expression analyses and contributed to draft the manuscript. GC coordinated the design of PiRATE, the control, the TE annotation and the expression analyses, participated in the genome assembly, realized the gene annotation and helped to draft the manuscript. All authors read and approved the final manuscript.

### Ethics approval and consent to participate

Not applicable.

### Competing interests

The authors declare that they have no competing interests.

### Publisher's Note

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

### Author details

<sup>1</sup>IFREMER, Physiology and Biotechnology of Algae Laboratory, rue de l'Île d'Yeu, 44311 Nantes, France. <sup>2</sup>Mer Molécules Santé, EA 2160 IUML - FR 3473 CNRS, Le Mans University, Le Mans, France. <sup>3</sup>Institut de Recherche en Horticulture et Semences, INRA of Angers, AGROCAMPUS-Ouest, SFR4207 QUASAV, Université d'Angers, Angers, France. <sup>4</sup>Research Unit in Genomics-Info, INRA of Versailles, Versailles, France. <sup>5</sup>Université Bretagne Loire, Angers, France.

Received: 8 December 2017 Accepted: 7 May 2018

Published online: 22 May 2018

### References

- Adams M, Kerlavage A, Fleischmann R, Fuldner R, Bult C, Lee N, et al. Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. *Nature*. 1995;377:3–174.
- Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, et al. The B73 maize genome: complexity, diversity, and dynamics. *Science*. 2009;326:1112.
- Carr M, Bensasson D, Bergman CM. Evolutionary Genomics of Transposable Elements in *Saccharomyces cerevisiae*. Stajich JE, editor. *PLoS ONE*. 2012;7:e50978.
- McClintock B. The origin and behavior of mutable loci in maize. *Proc Natl Acad Sci*. 1950;36:334–55.
- Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, et al. A unified classification system for eukaryotic transposable elements. *Nat Rev Genet*. 2007;8:973–82.
- Bureau TE, Wessler SR. Stowaway: a new family of inverted repeat elements associated with the genes of both monocotyledonous and dicotyledonous plants. *Plant Cell*. 1994;6:907–16.
- Kidwell MG. Transposable elements and the evolution of genome size in eukaryotes. *Genetica*. 2002;115:49–63.
- Chénais B, Caruso A, Hiard S, Casse N. The impact of transposable elements on eukaryotic genomes: from genome size increase to genetic adaptation to stressful environments. *Gene*. 2012;509:7–15.
- Levis RW, Ganesan R, Houtchens K, Tolar LA, Sheen F. Transposons in place of telomeric repeats at a *Drosophila* telomere. *Cell*. 1993;75:1083–93.
- Casacuberta E, González J. The impact of transposable elements in environmental adaptation. *Mol Ecol*. 2013;22:1503–17.
- Kazazian HH. Mobile elements: drivers of genome evolution. *Science*. 2004;303:1626–32.

12. Nekrutenko A, Li W-H. Transposable elements are found in a large number of human protein-coding genes. *Trends Genet.* 2001;17:619–21.
13. Lisch D. How important are transposons for plant evolution? *Nat Rev Genet.* 2012;14:49–61.
14. Darboux I, Charles J-F, Pauchet Y, Warot S, Pauron D. Transposon-mediated resistance to *Bacillus sphaericus* in a field-evolved population of *Culex pipiens* (Diptera: Culicidae). *Cell Microbiol.* 2007;9:2022–9.
15. Maumus F, Fiston-Lavier A-S, Quesneville H. Impact of transposable elements on insect genomes and biology. *Current Opinion in Insect Science.* 2015;7:30–6.
16. Egue F, Chenais B, Tastard E, Marchand J, Hiard S, Gateau H, et al. Expression of the retrotransposons *Surcouf* and *Blackbeard* in the marine diatom *Phaeodactylum tricomutum* under thermal stress. *Phycologia.* 2015;54:617–27.
17. Helliwell KE, Collins S, Kazamia E, Purton S, Wheeler GL, Smith AG. Fundamental shift in vitamin B12 eco-physiology of a model alga demonstrated by experimental evolution. *The ISME journal.* 2015;9:1446–55.
18. Maumus F, Allen AE, Mhiri C, Hu H, Jabbari K, Vardi A, et al. Potential impact of stress activated retrotransposons on genome evolution in a marine diatom. *BMC Genomics.* 2009;10:624.
19. Philippens GS, Avaca-Crusca JS, Araujo APU, DeMarco R. Distribution patterns and impact of transposable elements in genes of green algae. *Gene.* 2016;594:151–9.
20. De Clerck O, Guiry MD, Leliaert F, Samyn V, Verbruggen H. Algal taxonomy: a road to nowhere? *J Phycol.* 2013;49:215–25.
21. Sakai N, Sakamoto Y, Kishimoto N, Chihara M, Karube I. *Chlorella* strains from hot springs tolerant to high temperature and high CO<sub>2</sub>. *Energy Convers Manag.* 1995;36:693–6.
22. Rott E, Cantonati M, Füreder L, Pfister P. Benthic algae in high altitude streams of the alps – a neglected component of the aquatic biota. *Hydrobiologia.* 2006;562:195–216.
23. Anesio AM, Laybourn-Parry J. Glaciers and ice sheets as a biome. *Trends Ecol Evol.* 2012;27:219–25.
24. Treves H, Raanan H, Finkel OM, Berkowicz SM, Keren N, Shotland Y, et al. A newly isolated *Chlorella* sp. from desert sand crusts exhibits a unique resistance to excess light intensity. *FEMS Microbiol Ecol.* 2013;86:373–80.
25. Bertheliet J, Schnitzler CE, Wood-Charlson EM, Poole AZ, Weis VM, Detournay O. Implication of the host TGFβ pathway in the onset of symbiosis between larvae of the coral *Fungia scutaria* and the dinoflagellate *Symbiodinium* sp. (clade C1f). *Coral Reefs.* 2017;36:1263–8.
26. Guiry MD. How many species of algae are there? *J Phycol.* 2012;48:1057–63.
27. Not F, Siano R, Kooistra WHCF, Simon N, Vulot D, Probert I. Diversity and Ecology of Eukaryotic Marine Phytoplankton. *Advances in Botanical Research* [Internet]. Elsevier; 2012 [cited 2015 Oct 29]. p. 1–53. Available from: <http://linkinghub.elsevier.com/retrieve/pii/B9780123914996000013>
28. Merchant SS, Prochnik SE, Vallon O, Harris EH, Karpowicz SJ, Witman GB, et al. The *Chlamydomonas* genome reveals the evolution of key animal and plant functions. *Science.* 2007;318:245–50.
29. Prochnik SE, Umen J, Nedelcu AM, Hallmann A, Miller SM, Nishii I, et al. Genomic analysis of organismal complexity in the multicellular green alga *Volvox carterii*. *Science.* 2010;329:223–6.
30. Blanc G, Duncan G, Agarkova I, Borodovsky M, Gurnon J, Kuo A, et al. The *Chlorella variabilis* NC64A genome reveals adaptation to Photosymbiosis, coevolution with viruses, and cryptic sex. *Plant Cell.* 2010;22:2943–55.
31. Derelle E, Ferraz C, Rombauts S, Rouzé P, Worden AZ, Robbens S, et al. Genome analysis of the smallest free-living eukaryote *Ostreococcus tauri* unveils many unique features. *Proc Natl Acad Sci.* 2006;103:11647–52.
32. Palenik B, Grimwood J, Aerts A, Rouzé P, Salamov A, Putnam N, et al. The tiny eukaryote *Ostreococcus* provides genomic insights into the paradox of plankton speciation. *Proc Natl Acad Sci.* 2007;104:7705–10.
33. Worden AZ, Lee J-H, Mock T, Rouzé P, Simmons MP, Aerts AL, et al. Green evolution and dynamic adaptations revealed by genomes of the marine Picoeukaryotes *Ostreococcus*. *Science.* 2009;324:268.
34. Blanc G, Agarkova I, Grimwood J, Kuo A, Brueggeman A, Dunigan DD, et al. The genome of the polar eukaryotic microalga *Coccomyxa subellipsoidea* reveals traits of cold adaptation. *Genome Biol.* 2012;13:R39.
35. Vieler A, Wu G, Tsai C-H, Bullard B, Cornish AJ, Harvey C, et al. Genome, Functional Gene Annotation, and Nuclear Transformation of the Heterokont Oleaginous Alga *Nannochloropsis oceanica* CCMP1779. *Bhattacharya D, editor. PLoS Genetics.* 2012;8:e1003064.
36. Moreau H, Verhelst B, Couloux A, Derelle E, Rombauts S, Grimsley N, et al. Gene functionalities and genome structure in *Bathycoccus prasinos* reflect cellular specializations at the base of the green lineage. *Genome Biol.* 2012;13:R74.
37. Roth MS, Cokus SJ, Gallaher SD, Walter A, Lopez D, Erickson E, et al. Chromosome-level genome assembly and transcriptome of the green alga *Chromochloris zofingiensis* illuminates astaxanthin production. *Proc Natl Acad Sci.* 2017;114:E4296–305.
38. Nozaki H, Takano H, Misumi O, Terasawa K, Matsuzaki M, Maruyama S, et al. A 100%-complete sequence reveals unusually simple genomic features in the hot-spring red alga *Cyanidioschyzon merolae*. *BMC Biol.* 2007;5:28.
39. Armbrust EV. The genome of the diatom *Thalassiosira Pseudonana*: ecology, evolution, and metabolism. *Science.* 2004;306:79–86.
40. Bowler C, Allen AE, Badger JH, Grimwood J, Jabbari K, Kuo A, et al. The *Phaeodactylum* genome reveals the evolutionary history of diatom genomes. *Nature.* 2008;456:239–44.
41. Tanaka T, Maeda Y, Veluchamy A, Tanaka M, Abida H, Maréchal E, et al. Oil accumulation by the oleaginous diatom *Fistulifera solaris* as revealed by the genome and transcriptome. *The Plant Cell Online.* 2015;27:162–76.
42. McEWAN M, Humayun R, Slamovits CH, Keeling PJ. Nuclear genome sequence survey of the dinoflagellate *Heterocapsa triquetra*. *J Eukaryot Microbiol.* 2008;55:530–5.
43. Jaeckisch N, Yang I, Wohlrab S, Glöckner G, Kroymann J, Vogel H, et al. Comparative Genomic and Transcriptomic Characterization of the Toxicogenic Marine Dinoflagellate *Alexandrium ostenfeldii*. *Moustafa A. PLoS ONE.* 2011;6:e28012.
44. Shoguchi E, Shinzato C, Kawashima T, Gyoja F, Mungpakdee S, Koyanagi R, et al. Draft assembly of the Symbiodinium minutum nuclear genome reveals dinoflagellate gene structure. *Curr Biol.* 2013;23:1399–408.
45. Lin S, Cheng S, Song B, Zhong X, Lin X, Li W, et al. The Symbiodinium kawagutii genome illuminates dinoflagellate gene expression and coral symbiosis. *Science.* 2015;350:691–4.
46. Aranda M, Li Y, Liew YJ, Baumgarten S, Simakov O, Wilson MC, et al. Genomes of coral dinoflagellate symbionts highlight evolutionary adaptations conducive to a symbiotic lifestyle. *Scientific Reports* [Internet]. 2016 [cited 2018 Feb 16];6. Available from: <http://www.nature.com/articles/srep39734>
47. Read BA, Kegel J, Klute MJ, Kuo A, Lefebvre SC, Maumus F, et al. Pan genome of the phytoplankton *Emiliania* underpins its global distribution. *Nature.* 2013;499:209–13.
48. Hermann D. Caractérisation d'éléments transposables de type mariner chez les microalgues marines [Internet]. Université du Maine; 2011 [cited 2015 Nov 23]. Available from: <https://tel.archives-ouvertes.fr/tel-00732952/>
49. Nguyen DH. Caractérisation et expression de nouveaux éléments génétiques transposables de la superfamille Tc1-Mariner chez la microalga marine *Amphora acutiuscula* (Bacillariophyta). 2014;
50. Chen JE, Cui G, Wang X, Liew YJ, Aranda M. Recent expansion of heat-activated retrotransposons in the coral symbiont *Symbiodinium microadriaticum*. *The ISME Journal.* 2017;
51. Carrier G, Barouk C, Rouxel C, Duboscq-Bidot L, Schreiber N, Bougaran G. Draft genomes and phenotypic characterization of *Tisochrysis lutea* strains. Toward the production of domesticated strains with high added value. *Algal Res.* 2018;29:1–11.
52. Liu W, Pearce CM, McKinley RS, Forster IP. Nutritional value of selected species of microalgae for larvae and early post-set juveniles of the Pacific geoduck clam, *Panopea generosa*. *Aquaculture.* 2016;452:326–41.
53. Marchetti J, Bougaran G, Le Dean L, Mégrier C, Lukomska E, Kaas R, et al. Optimizing conditions for the continuous culture of *Isochrysis affinis galbana* relevant to commercial hatcheries. *Aquaculture.* 2012;326–329:106–15.
54. Sánchez Á, Maceiras R, Cancela Á, Pérez A. Culture aspects of *Isochrysis galbana* for biodiesel production. *Appl Energy.* 2013;101:192–7.
55. Bougaran G, Rouxel C, Dubois N, Kaas R, Grouas S, Lukomska E, et al. Enhancement of neutral lipid productivity in the microalga *Isochrysis affinis galbana* (T-Iso) by a mutation-selection procedure. *Biotechnol Bioeng.* 2012; 109:2737–45.
56. Carrier G, Garnier M, Le Cunff L, Bougaran G, Probert I, De Vargas C, et al. Comparative transcriptome of wild type and selected strains of the microalga *Tisochrysis lutea* provides insights into the genetic basis, lipid metabolism and the life cycle. *Abad-Grau MM. PLoS One.* 2014;9:e86889.
57. Charrier A, Bérard J-B, Bougaran G, Carrier G, Lukomska E, Schreiber N, et al. High-affinity nitrate/nitrite transporter genes (*Nrt2*) in *Tisochrysis lutea*: identification and expression analyses reveal some interesting specificities of Haptophyta microalgae. *Physiol Plant.* 2015;154:572–90.

58. Garnier M, Bougaran G, Pavlovic M, Berard J-B, Carrier G, Charrier A, et al. Use of a lipid rich strain reveals mechanisms of nitrogen limitation and carbon partitioning in the haptophyte *Tisochrysis lutea*. *Algal Res.* 2016;20:229–48.
59. Thiriet-Rupert S, Carrier G, Chénais B, Trottier C, Bougaran G, Cadoret J-P, et al. Transcription factors in microalgae: genome-wide prediction and comparative analysis. *BMC Genomics.* 2016;17:282.
60. Thiriet-Rupert S, Carrier G, Trottier C, Eveillard D, Schoefs B, Bougaran G, et al. Identification of transcription factors involved in the phenotype of a domesticated oleaginous microalgae strain of *Tisochrysis lutea*. *Algal Res.* 2018;30:59–72.
61. Smit, A. F., Hubley, R., & Green, P. (1996). RepeatMasker. [Internet]. Available from: <http://www.repeatmasker.org>.
62. Han Y, Wessler SR. MITE-hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic Acids Res.* 2010;38:e199–e199.
63. Price AL, Jones NC, Pevzner PA. De novo identification of repeat families in large genomes. *Bioinformatics.* 2005;21:i351–8.
64. 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 Biology and Evolution.* 2015;7:1192–205.
65. Ragupathy R, You FM, Cloutier S. Arguments for standardizing transposable element annotation in plant genomes. *Trends Plant Sci.* 2013;18:367–76.
66. Hoen DR, Hickey G, Bourque G, Casacuberta J, Cordaux R, Feschotte C, et al. A call for benchmarking transposable element annotation methods. *Mobile DNA* [Internet]. 2015 [cited 2017 Jun 28];6. Available from: <http://www.mobilednajournal.com/content/6/1/13>
67. Arensburg P, Piégu B, Bigot Y. The future of transposable element annotation and their classification in the light of functional genomics - what we can learn from the fables of Jean de la Fontaine? *Mobile Genetic Elements.* 2016;6:e1256852.
68. Flutre T, Duprat E, Feuillet C, Quesneville H. Considering transposable element diversification in De novo annotation approaches. *Xu Y. PLoS One.* 2011;6:e16526.
69. Smit, AF, Hubley, R. RepeatModeler Open-1.0 [Internet]. 2010. Available from: <http://www.repeatmasker.org>.
70. Hoen DR, Hickey G, Bourque G, Casacuberta J, Cordaux R, Feschotte C, et al. A call for benchmarking transposable element annotation methods. *Mob DNA.* 2015;6:13.
71. Kamoun C, Payen T, Hua-Van A, Filée J. Improving prokaryotic transposable elements identification using a combination of de novo and profile HMM methods. *BMC Genomics.* 2013;14:700.
72. Hoede C, Arnoux S, Moisset M, Chaumier T, Inizan O, Jamilloux V, et al. PASTEC: An Automatic Transposable Element Classification Tool. *Cordaux R. PLoS One.* 2014;9:e91929.
73. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. Repbase update, a database of eukaryotic repetitive elements. *Cytogenetic and Genome Research.* 2005;110:462–7.
74. Bao W, Kojima KK, Kohany O. Repbase update, a database of repetitive elements in eukaryotic genomes. *Mob DNA.* 2015;6:11.
75. Ellinghaus D, Kurtz S, Willhoeft U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics.* 2008;9:18.
76. Novak P, Neumann P, Pech J, Steinhaisl J, Macas J. RepeatExplorer: a galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. *Bioinformatics.* 2013;29:792–3.
77. Koch P, Platzer M, Downie BR. RepARK—de novo creation of repeat libraries from whole-genome NGS reads. *Nucleic Acids Res.* 2014;42:e80–e80.
78. Garrett JE, Carroll D. Tx1: a transposable element from *Xenopus laevis* with some unusual properties. *Mol Cell Biol.* 1986;6:933–41.
79. Kapitonov VV, Jurka J. A universal classification of eukaryotic transposable elements implemented in Repbase. *Nat Rev Genet.* 2008;9:411–2.
80. Treangen TJ, Salzberg SL. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews Genetics.* 2011;13:nrg3117.
81. McCoy RC, Taylor RW, Blauwkamp TA, Kelley JL, Kertesz M, Pushkarev D, et al. Illumina TruSeq synthetic long-reads empower De novo assembly and resolve complex, highly-repetitive transposable elements. *Singh N. PLoS One.* 2014;9:e106689.
82. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *bioRxiv.* 2017:071282.
83. Khost DE, Eickbush DG, Larracuenta AM. Single-molecule sequencing resolves the detailed structure of complex satellite DNA loci in *Drosophila melanogaster*. *Genome Res.* 2017;27:709–21.
84. Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, et al. A tale of three next generation sequencing platforms: comparison of ion torrent, Pacific biosciences and Illumina MiSeq sequencers. *BMC Genomics.* 2012;13:341.
85. Phillippy AM. New advances in sequence assembly. *Genome Res.* 2017;27:xi–xiii.
86. Zimin AV, Puiu D, Luo M-C, Zhu T, Koren S, Marçais G, et al. Hybrid assembly of the large and highly repetitive genome of *Aegilops tauschii*, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. *Genome Res.* 2017;27:787–92.
87. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *Wang J. PLoS One.* 2014;9:e112963.
88. Zdobnov EM, Apweiler R. InterProScan—an integration platform for the signature-recognition methods in InterPro. *Bioinformatics.* 2001;17:847–8.
89. Cantarel BL, Korf I, Robb SMC, Parra G, Ross E, Moore B, et al. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.* 2007;18:188–96.
90. Holt C, Yandell M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC bioinformatics.* 2011;12:491.
91. Lomsadze A, Gemayel K, Tang S, Borodovsky M. Improved Prokaryotic Gene Prediction Yields Insights into Transcription and Translation Mechanisms on Whole Genome Scale. *bioRxiv.* 2017;193490.
92. Eddy SR. Others. Multiple alignment using hidden Markov models. *Ismb.* 1995;3:114–20.
93. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol.* 1990;215:403–10.
94. Yang L, Bennetzen JL. Structure-based discovery and description of plant and animal Helitrons. *Proc Natl Acad Sci.* 2009;106:12832–7.
95. Rho M, Tang H. MGEScan-non-LTR: computational identification and classification of autonomous non-LTR retrotransposons in eukaryotic genomes. *Nucleic Acids Res.* 2009;37:e143–e143.
96. Wenke T, Dobel T, Sorensen TR, Junghans H, Weisshaar B, Schmidt T. Targeted identification of short interspersed nuclear element families shows their widespread existence and extreme heterogeneity in plant genomes. *THE PLANT CELL ONLINE.* 2011;23:3117–28.
97. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics.* 2006;22:1658–9.
98. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, et al. Circos: an information aesthetic for comparative genomics. *Genome Res.* 2009;19:1639–45.
99. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol.* 2011;29:644–52.
100. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics.* 2009;25:1105–11.
101. Anders S, Pyl PT, Huber W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics.* 2015;31:166–9.
102. Kearsley M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, et al. Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics.* 2012;28:1647–9.
103. Giardine B, Riemer C, Hardison RC, Burhans R, Eltnitski L, Shah P, et al. Galaxy: a platform for interactive large-scale genome analysis. *Genome Res.* 2005;15:1451–5.



# A high-quality genome sequence of *Rosa chinensis* to elucidate ornamental traits

L. Hibrand Saint-Oyant<sup>1</sup>, T. Ruttink<sup>2</sup>, L. Hamama<sup>1</sup>, I. Kirov<sup>2,3</sup>, D. Lakhwani<sup>1</sup>, N. N. Zhou<sup>1</sup>, P. M. Bourke<sup>4</sup>, N. Daccord<sup>1</sup>, L. Leus<sup>2</sup>, D. Schulz<sup>5</sup>, H. Van de Geest<sup>6</sup>, T. Hesselink<sup>6</sup>, K. Van Laere<sup>2</sup>, K. Debray<sup>1</sup>, S. Balzergue<sup>1</sup>, T. Thouroude<sup>1</sup>, A. Chastellier<sup>1</sup>, J. Jeauffre<sup>1</sup>, L. Voisine<sup>1</sup>, S. Gaillard<sup>1</sup>, T. J. A. Borm<sup>4</sup>, P. Arens<sup>4</sup>, R. E. Voorrips<sup>4</sup>, C. Maliepaard<sup>4</sup>, E. Neu<sup>5</sup>, M. Linde<sup>5</sup>, M. C. Le Paslier<sup>7</sup>, A. Bérard<sup>7</sup>, R. Bounon<sup>7</sup>, J. Clotault<sup>1</sup>, N. Choisne<sup>8</sup>, H. Quesneville<sup>8</sup>, K. Kawamura<sup>9</sup>, S. Aubourg<sup>1</sup>, S. Sakr<sup>1</sup>, M. J. M. Smulders<sup>4</sup>, E. Schijlen<sup>6</sup>, E. Bucher<sup>1</sup>, T. Debener<sup>5</sup>, J. De Riek<sup>2</sup> and F. Foucher<sup>1\*</sup>

**Rose is the world's most important ornamental plant, with economic, cultural and symbolic value. Roses are cultivated worldwide and sold as garden roses, cut flowers and potted plants. Roses are outbred and can have various ploidy levels. Our objectives were to develop a high-quality reference genome sequence for the genus *Rosa* by sequencing a doubled haploid, combining long and short reads, and anchoring to a high-density genetic map, and to study the genome structure and genetic basis of major ornamental traits. We produced a doubled haploid rose line ('HapOB') from *Rosa chinensis* 'Old Blush' and generated a rose genome assembly anchored to seven pseudo-chromosomes (512 Mb with N50 of 3.4 Mb and 564 contigs). The length of 512 Mb represents 90.1–96.1% of the estimated haploid genome size of rose. Of the assembly, 95% is contained in only 196 contigs. The anchoring was validated using high-density diploid and tetraploid genetic maps. We delineated hallmark chromosomal features, including the pericentromeric regions, through annotation of transposable element families and positioned centromeric repeats using fluorescent in situ hybridization. The rose genome displays extensive synteny with the *Fragaria vesca* genome, and we delineated only two major rearrangements. Genetic diversity was analysed using resequencing data of seven diploid and one tetraploid *Rosa* species selected from various sections of the genus. Combining genetic and genomic approaches, we identified potential genetic regulators of key ornamental traits, including prickle density and the number of flower petals. A rose *APETALA2/TOE* homologue is proposed to be the major regulator of petal number in rose. This reference sequence is an important resource for studying polyploidization, meiosis and developmental processes, as we demonstrated for flower and prickle development. It will also accelerate breeding through the development of molecular markers linked to traits, the identification of the genes underlying them and the exploitation of synteny across Rosaceae.**

Rose is the queen of flowers, holding great symbolic and cultural value. Roses appeared as decoration on 5,000-year-old Asian pottery<sup>1</sup>, and Romans cultivated roses for their flowers and essential oil<sup>2</sup>. Today, no ornamental plants have greater economic importance than roses. They are cultivated worldwide and are sold as garden plants, in pots or as cut flowers, the latter accounting for approximately 30% of the market. Roses are also used for scent production and for culinary purposes<sup>3</sup>.

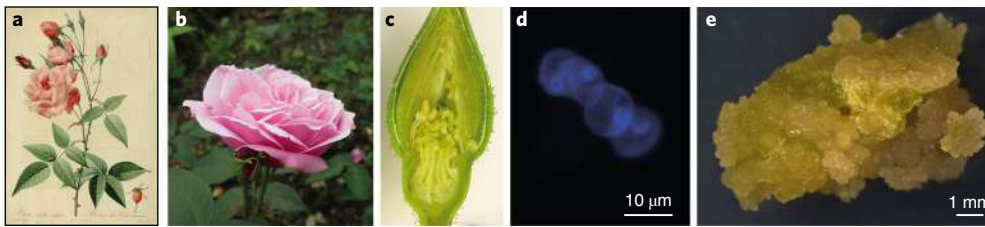
Despite their genetic complexity and lack of biotechnological resources, rose represents a model for ornamental plant species, allowing the investigation of traits such as bloom seasonality or flower morphology. Furthermore, rose displays a range of unique features as a result of its complex evolutionary and breeding history, including interspecific hybridization events and polyploidization<sup>4–6</sup>. Roses belong to the genus *Rosa* (Rosoideae, Rosaceae), which contains more than 150 species<sup>7</sup> of varying ploidy levels, ranging from  $2n=2x$  to  $10x^{8,9}$ . Many modern roses are tetraploid and can be genetically classified as 'segmental' allopolyploids (a mixture between allopolyploidy and autopolyploidy)<sup>10</sup>, whereas

dog-roses display unequal meiosis to maintain pentaploidy<sup>11,12</sup>. Rose breeding has a long and generally unresolved history in Europe and Asia, most likely involving several interspecific hybridization events. Importantly, many very-old varieties are still maintained in private and public rose gardens and are a living historical archive of rose breeding and selection<sup>13</sup>. Large and well-documented herbarium collections, combined with genomic advances, offer excellent opportunities to reconstruct phylogenetic relationships within the species.

Roses have been subject to selection for several traits that are not usually encountered in other crops. In particular, aesthetic criteria have been a principal focus of rose breeding over the past 250 years, next to plant vigour and resistances to biotic and abiotic stresses. Among the aesthetic traits, flower colour and architecture (from 5-petalled 'simple' flowers to 100-petalled 'double' flowers), floral scent and prickle formation on the stem and leaves have been the main targets of the breeders' eyes (and noses). Although these traits can be interpreted as signs of the domestication process, they originally evolved through adaptation to natural conditions.

<sup>1</sup>IRHS, Agrocampus-Ouest, INRA, Université d'Angers, SFR 4207 QuaSaV, Beaucozoué, France. <sup>2</sup>ILVO, Flanders Research Institute for Agriculture, Fisheries and Food, Plant Sciences Unit, Melle, Belgium. <sup>3</sup>Russian State Agrarian University-Moscow Timiryazev Agricultural Academy, Moscow, Russia.

<sup>4</sup>Plant Breeding, Wageningen University & Research, Wageningen, The Netherlands. <sup>5</sup>Leibniz Universität, Hannover, Germany. <sup>6</sup>Wageningen University & Research, Business Unit Bioscience, Wageningen, The Netherlands. <sup>7</sup>INRA, US 1279 EPGV, Université Paris-Saclay, Evry, France. <sup>8</sup>URGI, INRA, Université Paris-Saclay, Versailles, France. <sup>9</sup>Osaka Institute of Technology, Osaka, Japan. \*e-mail: [fabrice.foucher@inra.fr](mailto:fabrice.foucher@inra.fr)



**Fig. 1 | Development of the HapOB haploid line from *R. chinensis* 'Old Blush'.** **a**, The *R. chinensis* variety 'Old Blush' painted by Redouté in 1817. Paul Fearn/Alamy Stock Photo. **b**, A flower from the *R. chinensis* variety 'Old Blush'. **c**, A cross-section of the floral stage used for the anther culture. **d**, DAPI staining on mid-to-late uninucleate microspores. Similar results were observed on more than 15 microspores in one experiment. **e**, The HapOB callus was obtained after the anther culture at the appropriate stage and used for genome sequencing.

The availability of a high-quality reference genome sequence is key to unravelling the genetic basis underlying these evolutionary and developmental processes that accelerate future genetic, genomic, transcriptomic and epigenetic analyses. Recently, a draft reference genome sequence of *Rosa multiflora* has been published<sup>14</sup>. Although completeness measures suggest that the assembly is fairly complete in terms of the gene space covered, it is also highly fragmented (83,189 scaffolds, N50 of 90 kb).

Here, we present an annotated high-quality reference genome sequence for the *Rosa* genus using a haploid rose line derived from an old Chinese *Rosa chinensis* variety 'Old Blush' (Fig. 1a,b). 'Old Blush' (syn. Parsons' Pink China) was brought to Europe and North America in the eighteenth century from China and is one of the most influential genotypes in the history of rose breeding. Among other things, it introduced recurrent flowering into Western germplasm, which is an essential trait for the development of modern rose cultivars<sup>15</sup>. We validated our pseudo-chromosome scale genome assembly of 'Old Blush' using high-density genetic maps of multiple F1 progenies and synteny with *Fragaria vesca*. We delineated hallmark chromosomal features, such as the pericentromeric regions, through annotation of transposable element families and positioning of centromeric repeats using fluorescent in situ hybridization (FISH). This reference genome also allowed us to analyse the genetic diversity within the *Rosa* genus following a resequencing of eight wild species. Using genetic (F1 progeny and diversity panel) and genomic approaches, we were able to identify key potential genetic regulators of important ornamental traits, including continuous flowering, flower development, prickly density and self-incompatibility.

## Results

**Development of a high-quality reference genome sequence.** We developed a haploid callus cell line (HapOB) using an anther culture at the mid-to-late uninucleate microspore developmental stage from the diploid heterozygous 'Old Blush' variety (Fig. 1c–e). The homozygosity of the HapOB line was verified with ten microsatellite markers distributed over the seven linkage groups (Supplementary Table 1). Flow cytometric analysis showed the HapOB callus to be diploid, suggesting that spontaneous genome doubling occurred during in vitro propagation.

A combination of Illumina short-read sequencing and PacBio long-read sequencing technologies was used to assemble the doubled haploid HapOB genome sequence. PacBio sequencing data (Supplementary Table 2) was assembled with CANU<sup>16</sup>, yielding 551 contigs (N50 of 3.4 Mb), representing a total length of 512 Mb. Of the obtained sequence, 95% is contained in only 196 contigs. The PacBio-based assembly was error corrected with Illumina paired-end reads: 37,300 single-nucleotide polymorphisms (SNPs) and 307,700 insertions and deletions (indels) were corrected, representing 341.1 kb (Supplementary Table 2). K-mer spectrum analysis (K=25) suggested a genome size of 532.7 Mb (251.1 Mb of a

unique genome sequence and 279.6 Mb of repetitive sequences), whereas flow cytometric analysis estimated a genome size of  $1C = 568 \pm 9$  Mb. Thus, the assembled sequence represents 96.1% or 90.1%, respectively, of the estimated genome size. No major contamination was detected by screening for the predicted prokaryotic genes (Supplementary Table 3). Furthermore, only four contigs had low Illumina read mapping frequency, all of which were found to most likely encode plant proteins.

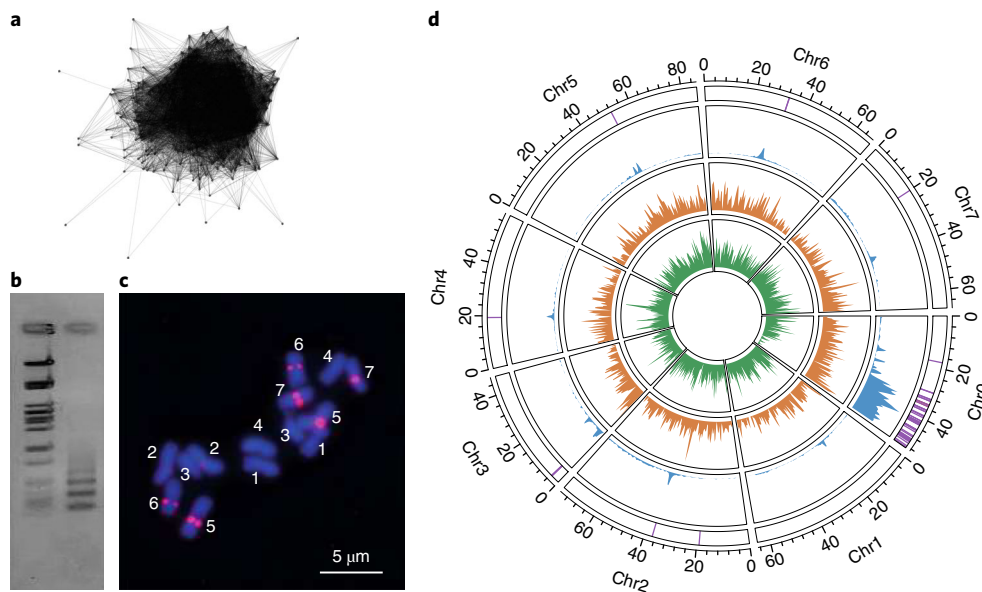
High-density female and male genetic maps were developed from a cross between *R. chinensis* 'Old Blush' and a hybrid of *Rosa wichurana* (OW). F1 progeny from this cross ( $n=151$ ) were genotyped with the WagRhSNP 68K Axiom array<sup>17</sup> (Table 1 and Supplementary Table 4). Thirteen contigs, for which marker order clearly indicated assembly artefacts, were split before anchoring all 564 resulting contigs to the female and male genetic maps using a total of 6,746 SNP markers (Table 1). Of these, 196 contigs were anchored manually onto the seven linkage groups, mostly on both the female and the male genetic maps (174 and 143 contigs, respectively). In total, 466 Mb were therefore anchored onto the genetic maps and assembled into 7 pseudo-chromosomes representing 90% of the assembled contig length (Table 1 and Supplementary Fig. 1a). The remaining 368 contigs (52 Mb) were assigned to chromosome 0 (Chr0). The quality of the assembly of the seven pseudo-chromosomes was assessed using two independent genetic maps: the previously published integrated high-density genetic map (K5) based on 172 tetraploid F1 progeny<sup>10</sup>, and a newly developed high-density map based on 174 diploid F1 progeny from a cross between cultivar 'Yesterday' and *R. wichurana* (YW; see Supplementary Fig. 1b). The co-linearity between the pseudo-chromosomes and both linkage maps is excellent (Supplementary Fig. 2). In addition, the anchoring of the 386 contigs (52 Mb), currently assigned to Chr0, onto the K5 map and the YW map revealed that 39 contigs (total: 28.4 Mb) and 27 contigs (total: 24.1 Mb), respectively, can potentially be positioned onto the 7 linkage groups (Supplementary Fig. 2). However, because these genetic maps were created using independent genotypes that are not related to *R. chinensis* 'Old Blush', we chose a conservative approach by not incorporating these contigs into the pseudo-chromosome sequence of HapOB.

**Positioning centromeres within the genome assembly.** The centromeric regions were identified using both bioinformatic and cytogenetic methods. We discovered a highly abundant tandem repeat (0.06% of the genome with more than 2,000 copies per haploid genome) of monomers (159 bp long) that we call OBC226 ('Old Blush' centromeric repeat from RepeatExplorer cluster 226; Fig. 2a). PCR confirmed the tandem organization of this repeat (Fig. 2b). FISH analysis unambiguously confirmed the location of the repeat in the centromeric regions of four of the seven chromosomes: Chr2, Chr5, Chr6 and Chr7 (Fig. 2c). Mapping of the OBC226 repeat sequence revealed regions with high coverage on all HapOB pseudo-chromosomes except Chr1, which explains why no clear

**Table 1 | Metrics of the alignment of the male and female genetic maps with the HapOB genome assembly**

| Linkage group | Genetic maps (no. of markers) |          | Chr.               | No. of anchored markers used for anchoring |       | No. of anchored contigs |      |                    |     |          | Pseudo-molecules<br>Size (in bp) |
|---------------|-------------------------------|----------|--------------------|--|-------|-------------------------|------|--------------------|-----|----------|----------------------------------|
|               | Female (OB)                   | Male (W) |                    | Female                                     | Male  | Female                  | Male | Manual integration | Cut | Excluded |                                  |
| 1             | 715                           | 195      | 1                  | 587  | 146   | 18                      | 14   | 18                 | 1   | 1        | 64,770,848                       |
| 2             | 1,114                         | 303      | 2                  | 1,001                                      | 249   | 14                      | 18   | 20                 | -   | -        | 75,129,302                       |
| 3             | 528                           | 564      | 3                  | 477  | 498   | 20                      | 25   | 31                 | -   | 1        | 46,843,630                       |
| 4             | 227                           | 404      | 4                  | 191  | 334   | 12                      | 18   | 20                 | -   | -        | 59,004,735                       |
| 5             | 1,031                         | 362      | 5                  | 866  | 275   | 40                      | 29   | 37                 | 2   | 1        | 85,885,663                       |
| 6             | 1,153                         | 254      | 6                  | 1,010                                      | 186   | 43                      | 20   | 43                 | -   | 1        | 67,395,200                       |
| 7             | 863                           | 241      | 7                  | 743  | 183   | 27                      | 19   | 27                 | -   | -        | 67,081,725                       |
| -             | -                             | -        | Total without Chr0 |  |       | 174                     | 143  | 196                | -   | -        | 466,111,103                      |
| -             | -                             | -        | 0                  | -  | -     | 387                     | 418  | 368                | -   | -        | 52,404,850                       |
| Total:        | 5,631                         | 2,323    | -                  | 4,875                                      | 1,871 | 561                     | 561  | 564                | -   | -        | 518,515,953                      |

The genetic maps were developed from a cross between 'Old Blush' (OB; female) and a hybrid of *R. wichurana* (W; male) using an Affymetrix SNP array. The initial size of the genome was 512 Mb and reached a final size of 518.5 Mb owing to the addition of 10,000 N between each contig to create the pseudo-molecules. N, any nucleotide.



**Fig. 2 | Identification of centromeric regions in the HapOB reference genome. a**, The cluster CL226 identified by RepeatExplorer. **b**, Agarose gel electrophoresis of tandem repeat fragments amplified from the genomic DNA of HapOB using OBC226 PCR primers (right lane) along with the lambda-PstI size ladder (left lane). Similar results were obtained in two independent experiments. **c**, FISH with carboxy tetramethylrhodamine (TAMRA)-labelled OBC226 oligo probes on *R. chinensis* metaphase chromosomes. Chromosome numbers are labelled from 1 to 7. Similar results were observed in at least 10 metaphase cells in two independent experiments. **d**, Circos representation of the distribution of OBC226 (purple), the pericentromeric region (blue), Ty3/Gypsy (orange) and Ty1/Copia repeat elements (green) along the seven pseudo-chromosomes and Chr0 (scale in Mb).

centromeric region could be detected on this chromosome (Fig. 2d). On Chr3 and Chr4, the copy number of OBC226 was probably too low to be detected by FISH. Furthermore, the core OBC226 centromeric repeats were flanked by other repetitive sequences, and these were unequally distributed along the chromosomes, with a clearly higher density in the core centromeric regions (Fig. 2d). These centromeric regions were also enriched in Ty3/Gypsy transposable elements. Taken together, these results confirm the position of the centromeric regions on the seven pseudo-chromosomes and reveal the high repeat sequence content, and low gene content, of the scaffolds currently assigned to Chr0.

**Annotation of the sequence. Coding genes.** Based on the mapping of 723,268 transcript sequences (expressed sequence tag/complementary DNA and RNA sequencing (RNA-seq) contigs with a minimum size of 150 bp) onto the HapOB genome assembly, we predicted a total of 44,481 genes covering 21% of the genomic sequence length using Eugene combiner<sup>18</sup>. These include 39,669 protein-coding genes and 4,812 non-coding genes. Evidence of transcription was found for 87.8% of all predicted genes. At least one InterPro domain signature was detected in 86.5% of the protein-coding genes using InterProScan<sup>19</sup>, with 68.0% of the genes assigned to 4,051 PFAM gene families<sup>20</sup>. The quality of the structural annotation

was assessed using the BUSCO v2 method based on a benchmark of 1,440 conserved plant genes<sup>21</sup>, of which 92.5% had complete gene coverage (including 5.3% duplicated ones), 4.1% were fragmented and only 3.4% were missing. This result can be compared to the analysis of the whole-genome assembly, which identified 95% complete genes and 3.6% missing genes. The set of predicted non-coding genes included 186 ribosomal RNA, 751 transfer RNA, 384 small nucleolar RNA, 99 microRNA, 170 small nuclear RNA and 3,222 unclassified genes (annotated as non-coding RNA) with evidence of transcription but no consistent coding sequence.

The number of predicted proteins in *Rosa* (39,669) is higher than the number of predicted proteins in *F. vesca* (28,588 predicted proteins<sup>22</sup>). By BLAST analysis, we identified 6,543 proteins that are rose specific. Among them, 5,867 proteins have no homologue in *Arabidopsis thaliana*. For these proteins, no functional information is available from closely related species and experimental evidence will be required to explore their role in roses. We also looked at whether the difference in the predicted number of proteins was owing to protein family expansion. Such a scenario was detected for some protein families, including nucleotide-binding site leucine-rich repeat (NBS-LRR) and cytochrome P450 (Supplementary Fig. 3).

**Transposable elements.** The REPET package<sup>23</sup> was used to generate a genome-wide annotation of repetitive sequences of the HapOB genome (see Methods for details). Retrotransposons, also called class I elements, represent the largest transposable element genomic fraction (35.1% of the sequenced genome), of which long terminal repeat (LTR) retrotransposons represent 28.3%. *Gypsy* elements are more frequent than *Copia* (Supplementary Table 5a). Non-LTR retrotransposons long interspersed nuclear elements (LINEs) and potential short interspersed nuclear elements (SINEs) represent 5.0% of the sequence genome and class II elements (DNA transposons and Helitrons) represent 11.7% (Supplementary Table 5a). The remaining 15% include unclassified repeats (7.3%), chimaeric consensus sequences (1.9%) and potential repeated host genes (5.8%). We also identified Caulimoviridae copies representing 1.25% of the genome. Interestingly, one particularly abundant *Gypsy* Tat-like family was found in the genome assembly. The total copy coverage represents 3.4% of the genome. Tat-like elements are known to have an open reading frame (ORF) after the polymerase domains, and surprisingly in this case, the ORF corresponds to a class II transposase domain.

In a preliminary comparison between the transposable element annotation in HapOB and the *F. vesca* v2.0.a1 genome assembly (without manual curation) (Supplementary Table 5b), we found that retrotransposon elements represent the largest transposable element genomic fraction in *F. vesca* (13.91%), similar to rose. We found approximately twofold more copies for all transposable element families except SINE and unclassified in *Rosa* than in *Fragaria*. This indicates that the difference in genome size between *Rosa* and *Fragaria* is largely due to an expansion of the transposable element fraction.

**Synteny between *Rosa* and *F. vesca*.** *Rosa* and *Fragaria* both belong to the Rosoideae subfamily of the Rosaceae<sup>24</sup>, having diverged around 50 million years ago<sup>25</sup>. Previous genetic studies have demonstrated that large macrosyntentic blocks are conserved between *Rosa* and *Fragaria*<sup>10,26</sup>. We compared the HapOB genome to the recently updated *F. vesca* genome<sup>22</sup> to analyse the synteny in detail (Supplementary Fig. 4a). *R. chinensis* Chrs 1, 4, 5, 6 and 7 display strong synteny with *F. vesca* Chrs 7, 4, 3, 2 and 5, respectively. Consistent with previous suggestions<sup>10</sup>, a reciprocal translocation was detected between *R. chinensis* Chr 2 and 3 and *F. vesca* Chrs 6 and 1, respectively. Our results clarify the highly conserved synteny between *F. vesca* and *Rosa*, revealing only two major translocation events.

Within the Rosaceae family, the synteny is also well conserved between *Prunus* and *Rosa* (Supplementary Fig. 4c): *Rosa* Chr1 corresponds to *Prunus* Chr2, *Rosa* Chr4 corresponds to the end of *Prunus* Chr1, whereas *Prunus* Chrs 3, 5 and 8 correspond to large parts of *Rosa* Chrs 2, 6 and 7 respectively. Owing to the allopolyploid origin of *Malus*, the overall synteny is less clear, even if large blocks of synteny can be detected (Supplementary Fig. 4b).

**Genetic diversity with the genus *Rosa*.** The 150 or more existing rose species belong to four subgenera. Excluding the subgenus *Rosa*, all subgenera contain only one or two species. We resequenced eight *Rosa* species (Table 2), representing three of the four subgenera (*Hulthemia*: *R. persica*, *Herperhodod*: *R. minutifolia* and *Rosa*). Within *Rosa*, we covered all of the main sections according to the latest phylogenetic analyses<sup>27,28</sup> (Table 2) in the form of *R. chinensis* var. *spontanea*, *R. rugosa*, *R. laevigata*, *R. moschata*, *R. xanthina* *spontanea* and *R. gallica*. All are diploid species except *R. gallica*, which is tetraploid (Table 2). SNPs and indels were identified relative to the HapOB reference sequence (Fig. 3).

The nuclear SNP-based phylogenetic tree of the eight species (Fig. 3a) is consistent with previous molecular analyses<sup>27,28</sup>. The clade, including *R. chinensis*, *R. gallica*, *R. moschata* and *R. laevigata*, fits with the *Synstylae* and allies clade previously found in a chloroplastic analysis<sup>28</sup>. The same is the case for *R. persica* and *R. xanthina*, both belonging to the *Cinnamomeae* and allies clade. However, *R. rugosa* and *R. minutifolia* show an uncertain position. In particular, *R. minutifolia*, which belongs to the *Hesperhodod* subgenus<sup>27,28</sup>, was expected to be closer to *R. persica* and *R. xanthina*. One of the possible explanations is that the resequenced *R. minutifolia* individual is actually the product of an interspecific cross, as it shows unexpected morphological characters, such as few prickles on the young flowering shoots and flowers clustering in inflorescences. Methodologically, the use of only homozygous SNPs may have caused bias, especially in *R. rugosa*, as most of its SNPs were in the heterozygous state (Supplementary Table 6).

The lowest SNP and indel density was found in *R. chinensis* var. *spontanea* (9.9 and 1.6 per kb, respectively). ‘Old Blush’ is described as an interspecific cross between *R. chinensis* var. *spontanea* and *R. x odorata* var. *gigantea*<sup>6</sup>, which is consistent with the relatively low sequence divergence of *R. chinensis* var. *spontanea* compared to the HapOB reference sequence. The highest SNP and indel density was found in *R. gallica* (21.0 and 4.5 per kb, respectively); this could be the result of the (allo)tetraploidy of this species<sup>29</sup>, as shown by its high proportion of heterozygous SNPs (74%; Supplementary Table 6).

As expected, the majority (79.2–89.0%) of the SNPs were located in non-coding regions (Supplementary Table 6). Only 3–7% of the SNPs were located in exons, of which half were synonymous, in line with other species (for example, tomato<sup>30</sup>). The different species displayed varying levels of homozygosity (homozygous SNPs ranging from 79.2% in *R. persica* to 26.0% in tetraploid *R. gallica*; Supplementary Table 6). The number of small indels was higher (between 876,648 and 2,430,123) than *Malus*, with an average of 346,498 indels<sup>31</sup>, suggesting a higher level of diversity within the *Rosa* genus.

**Analysis of the genetic control of important traits.** This new reference sequence is an important tool to help decipher the genetic basis of ornamental traits, such as blooming (including continuous flowering, flower development and the number of petals), prickle density on the stem and self-incompatibility. We studied the genetic determinism in (1) two F1 progenies (151 individuals obtained from the OW progeny and 174 individuals obtained from the YW progeny), and (2) a panel of 96 rose cultivars originating from the nineteenth to the twenty-first century<sup>32,33</sup>. Our data demonstrate that important loci controlling continuous flowering, double flower morphology, self-incompatibility and prickle density were predominantly localized in a single genomic region of Chr3 (Fig. 4a).



**Table 2 | Summary of resequencing and sequence variations (SNP and small indels) identified in eight *Rosa* species**

| Rosa species sequenced                    | Genome size (in Mb) | Classification     |                         | Ploidy | Flower colour | Flower morphology | Blooming seasonality | Geographical origin        | No. of reads (millions) | No. of reads mapped (millions) | HapOB genome covered by the mapping (%) | Depth of coverage (in ×) <sup>a</sup> | No. of SNPs | SNP/density (no. per kb) | No. of small indels | Small indel density (no. per kb) |
|---|---------------------|--------------------|-------------------------|--------|---------------|-------------------|----------------------|----------------------------|-------------------------|--------------------------------|---|---------------------------------------|-------------|--------------------------|---------------------|----------------------------------|
|   |                     | Subgenus           | Section                 |        |               |                   |                      |                            |                         |                                |   |                                       |             |                          |                     |                                  |
| <i>R. chinensis</i> var. <i>spontanea</i> | 562                 | <i>Rosa</i>        | <i>Chinenses</i>        | 2      | Pink          | Single            | Once blooming        | China                      | 110                     | 104                            | 90                                      | 28                                    | 5,564,345   | 9.9                      | 876,648             | 1.6                              |
| <i>R. gallica</i>                         | 538                 | <i>Rosa</i>        | <i>Gallicanae</i>       | 4      | Pink          | Single            | Once blooming        | Europe                     | 231                     | 218                            | 90                                      | 73                                    | 11,280,831  | 21.0                     | 2,430,138           | 4.5                              |
| <i>R. laevigata</i>                       | 562                 | <i>Rosa</i>        | <i>Laevigatae</i>       | 2      | White         | Single            | Occasionally         | China-Taiwan               | 100                     | 92                             | 70                                      | 31                                    | 6,327,292   | 11.3                     | 1,195,164           | 2.1                              |
| <i>R. moschata</i>                        | 554                 | <i>Rosa</i>        | <i>Synstylae</i>        | 2      | White         | Single            | Recurrent blooming   | Asia Minor                 | 92                      | 86                             | 71                                      | 29                                    | 5,862,043   | 10.6                     | 1,417,766           | 2.6                              |
| <i>R. munitifolia</i> <i>alba</i>         | 416                 | <i>Hesperhodos</i> |                         | 2      | White         | Single            | Once blooming        | North America              | 96                      | 89                             | 69                                      | 30                                    | 5,270,249   | 12.7                     | 1,208,933           | 2.9                              |
| <i>R. persica</i>                         | 416                 | <i>Hulthemia</i>   |                         | 2      | Yellow        | Single            | Once blooming        | Central Asia               | 114                     | 100                            | 56                                      | 34                                    | 5,602,086   | 13.5                     | 1,218,337           | 2.9                              |
| <i>R. rugosa</i>                          | 522                 | <i>Rosa</i>        | <i>Cinnamomeae</i>      | 2      | Pink          | Single            |                      | Northern China-Japan-Korea | 125                     | 116                            | 84                                      | 39                                    | 8,270,874   | 15.8                     | 1,703,127           | 3.3                              |
| <i>R. xanthina</i> <i>spontanea</i>       | 391                 | <i>Rosa</i>        | <i>Pimpinellifoliae</i> | 2      | Yellow        | Single            | Once blooming        | Asia                       | 95                      | 85                             | 60                                      | 28                                    | 5,642,595   | 14.4                     | 1,316,384           | 3.4                              |

<sup>a</sup>The depth of coverage is the ratio between the number of mapping base pairs (the number of mapping reads × read size) and the genome size.

**Detection of a new allele controlling continuous flowering in rose.** Most species of roses are 'once flowering'. In rose, continuous flowering is controlled by a homologue of the *TERMINAL FLOWER 1* (*TFL1*) family, *RoKSN*, located on Chr3 (ref. <sup>34</sup>). The continuous flowering phenotype is due to the insertion of a *Copia* retrotransposon element in the *RoKSN* gene. The continuous flowering rose 'Old Blush' was previously proposed to be *RoKSN<sup>Copia</sup>/RoKSN<sup>Copia</sup>* at the *RoKSN* locus<sup>34</sup>. This *Copia* element corresponds to the RLC<sub>denovoHm-B-G10244-Map11</sub> retrotransposon. We identified 34 insertions of this transposable element in the HapOB genome, of which 11 are full length (Supplementary Table 7a). The element is inserted into three genes, all of which are disrupted. The 3' and 5' LTRs of the full-length elements are >99% similar (Supplementary Table 7a), suggesting a recent insertion, as previously proposed for the element inserted in *RoKSN<sup>34</sup>*.

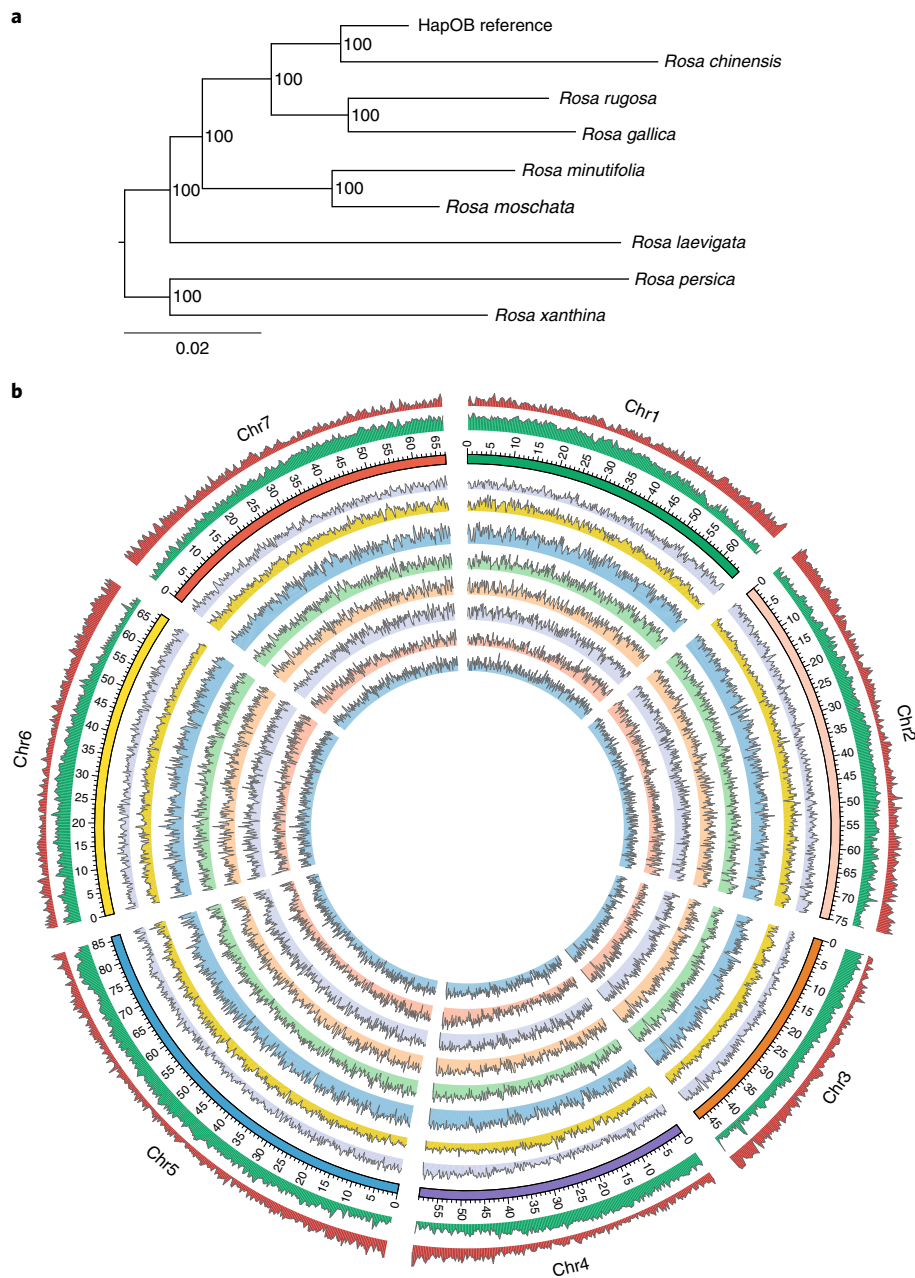
Here, quantitative trait locus (QTL) analysis in the OW progeny identified the *CONTINUOUS FLOWERING* locus on Chr3 (Fig. 4a), as expected, but we were unable to detect the *RoKSN* gene in the annotated HapOB genome. Detailed analysis of *RoKSN* allele segregation in the OW progeny revealed the existence of a null allele, in which *RoKSN* is deleted (see Supplementary Table 8 for further details). The diploid 'Old Blush' parent of the OB mapping population is therefore hemizygous *RoKSN<sup>Copia</sup>/RoKSN<sup>null</sup>*, and the *RoKSN<sup>null</sup>* allele is present in the HapOB genome sequence.

Interesting parallels exist between rose and *F. vesca* because *F. vesca* also exhibits both the once flowering and the continuous flowering phenotypes. In strawberry, a 2-bp deletion in the *TFL1* homologue causes a shift from once flowering to continuous flowering<sup>34,35</sup>. Synteny analysis revealed four orthologous syntenic blocks in the *RoKSN* gene region, here called blocks A–D (Supplementary Fig. 5). We detected a pattern of conserved gene content in combination with genome rearrangements between different *Rosa* species and the published genome sequence of *F. vesca*<sup>36</sup> where the synteny with *F. vesca* is broken at the *FvKSN* location. The *FvKSN* gene is located between the A and B blocks in *F. vesca*. The A block is inverted in the HapOB genome, and the C and D blocks are inserted

between the A and B blocks. In *R. multiflora*<sup>14</sup> and *R. laevigata* (see Methods for the partial *R. laevigata* genome sequence assembly), which are both once flowering, the *RoKSN<sup>WT</sup>* allele is present and synteny is conserved with *F. vesca* (Supplementary Fig. 5). Taken together, these data indicate that the *RoKSN<sup>null</sup>* allele is the result of a large rearrangement at the *CONTINUOUS FLOWERING* locus, leading to the complete deletion of the *RoKSN* gene. The *RoKSN<sup>null</sup>* allele represents a novel allele responsible for continuous flowering, which has not been previously described.

**Double flower.** The number of petals is an important ornamental trait, and roses with higher numbers of petal ('double flower') have traditionally been selected. Through a study of mutant lines (sports), the change in petal number was attributed to a homeotic conversion in organ identity, with stamens converted into petals<sup>37</sup>. The genetic basis of the double flower trait is complex, with a dominant gene (*DOUBLE FLOWER*) controlling simple versus double flower phenotypes and two QTLs controlling the number of petals on double flowers<sup>38</sup>. Here, we combined the genome sequence with segregation data of four different F1 progenies to confine the putative location of the *DOUBLE FLOWER* locus (Supplementary Table 9) to a 293-kb region of Chr3 (between position 33.24 Mb and 33.53 Mb; Fig. 4a). Using a genome-wide association study (GWAS) approach with a panel of 96 cultivated roses, we detected a strong association with simple versus double flowers in the same region (between position 33.08 Mb and 33.94 Mb; Fig. 4b). A second significant peak was located at a distance of 5 Mb, which may correspond to a secondary locus influencing this trait.

The 293-kb region contains 41 annotated genes. Among these, half are expressed during the early stages of floral development (Supplementary Table 10). By excluding genes expressed in later floral stages (with completely open flowers), we retained four candidate genes: an F-box protein (RC3G0245100), a homologue of *APETALA2/TOE* (RC3G0243000), a Ypt/Rab-GAP domain of the gyp1p superfamily protein (RC3G0245000) and a tetratricopeptide repeat-like superfamily protein (RC3G0243500) (Supplementary Table 10).

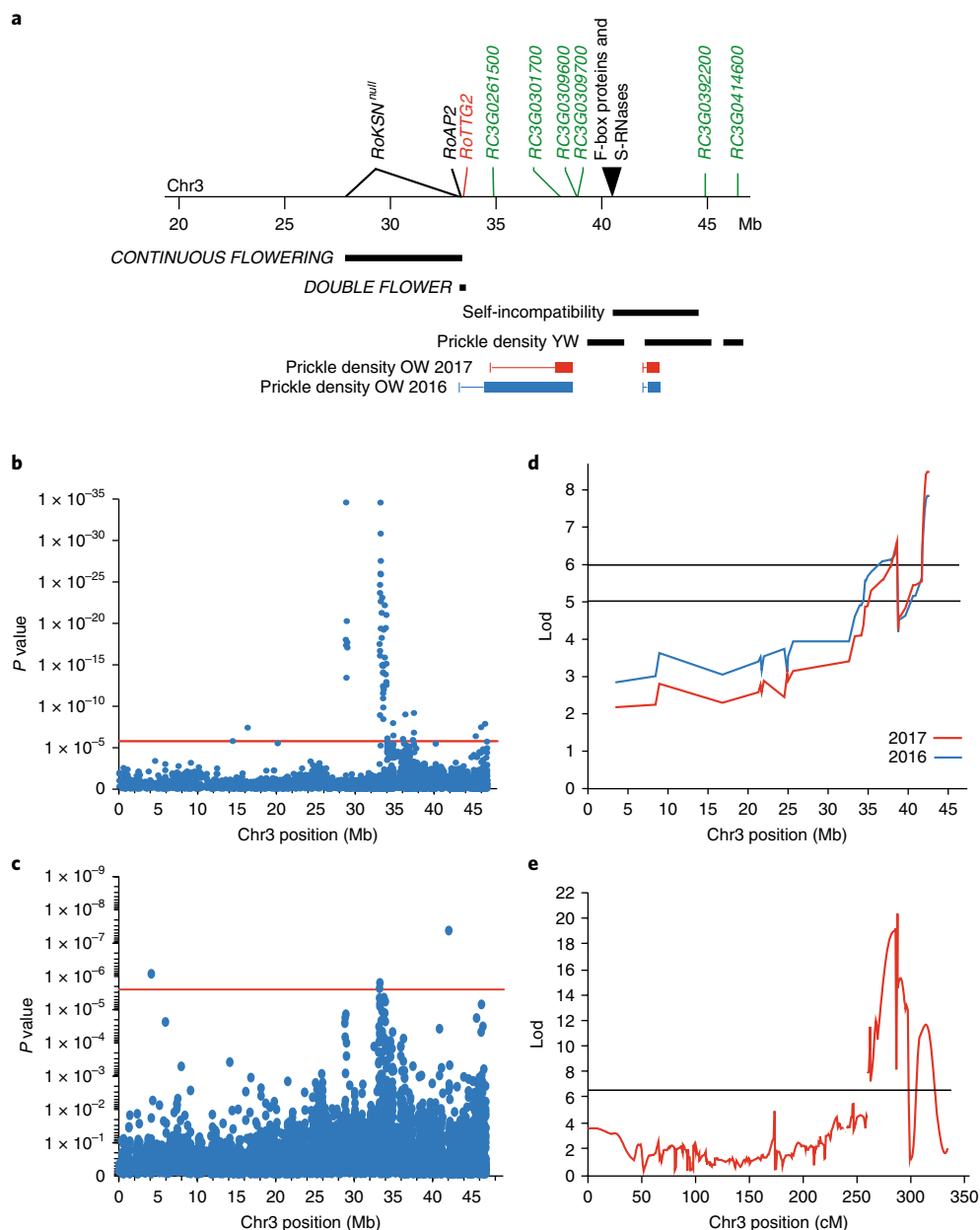


**Fig. 3 | Resequencing of eight *Rosa* species.** **a**, The phylogenetic relationships of the eight sequenced *Rosa* species and the reference genome HapOB, using a genome-wide set of homozygous SNPs. **b**, Analysis of genetic diversity in eight species of the *Rosa* genus along the seven pseudo-chromosomes of the HapOB reference sequence. Circles from outside to inside show: gene density (red), transposable element density (green), SNP density for *R. xanthina* (purple), *R. chinensis* var. *spontanea* (yellow), *R. gallica* (blue), *R. laevigata* (light green), *R. moschata* (light orange), *R. rugosa* (light purple), *R. persica* (light red) and *R. minutifolia alba* (light blue). Scales are in Mb.

Concerning double flowering, ‘Old Blush’ is heterozygous for the *DOUBLE FLOWER* locus. Sequencing both alleles of the four selected candidate genes in ‘Old Blush’ revealed only minor modifications for RC3G0245100, RC3G0245000 and RC3G0243500 (Supplementary Fig. 6a–c, respectively). However, concerning the *APETALA2/TOE* gene (RC3G0243000), we detected a 1,426-bp insertion in intron eight (Supplementary Fig. 7a). The insertion showed high similarity to an unclassified transposable element (annotation noCAT\_denovoHM-B-R7962-Map20; Supplementary Table 5c). This repeat element is present 62 times in the HapOB genome, of which 20 insertions are full length and 4 are located in gene introns (Supplementary Table 7b). Apart from this insertion and a few SNPs, no other differences were detected between the two

alleles. In the OW progeny, all individuals that carry the transposable element insertion allele display the double flower phenotype (see Supplementary Table 11 for further details).

Phylogenetic analysis showed that RC3G0243000 belongs to the *APETALA2/TOE* clade within the *AP2/ERF* subfamily<sup>39</sup> (Supplementary Fig. 7d). Like all members of the AP2 clades, the protein encoded by RC3G0243000 contains two conserved AP2 domains and a conserved putative *miRNA172* binding site (Supplementary Fig. 7b,c). The genomic position, expression analysis, protein sequence data and predicted deleterious effect of the insertion in intron 8 suggest that the *APETALA2/TOE* gene is a good candidate for the *DOUBLE FLOWER* locus. *APETALA2/TOE* has a central role in the



**Fig. 4 | A region at the end of Chr3 controls important ornamental traits.** **a**, Major genes and QTLs that control continuous flowering, double flower, self-incompatibility and prickle density are shown together with candidate genes for each trait. Detailed analyses per locus are described in Supplementary Figs. 5, 7, 9 and 10, respectively. For prickle density in OW progeny (OW2017 and OW2016), the boxes represent the 1-LOD (log of the odds ratio) interval and the lines the 2-LOD interval. **b,c**, GWAS analysis showing the *P* values of the association between SNPs positioned along Chr3 and the number of petals, indicating regions that control the number of petals. The petal number is considered as a qualitative trait (simple versus double flowers; GLM) (**b**) or as a quantitative trait (MLM) (**c**). The horizontal red line shows Bonferroni-corrected significance levels ( $1.78 \times 10^{-6}$ ). Other significant associations detected by GWAS are shown in Supplementary Fig. 12.  $n = 96$  cultivars with 3 flowers scored by cultivar. **d,e**, QTL analysis for prickle density in two F1 progenies using the OW mapping population based on scoring from 2016 and 2017,  $n = 151$  individuals (**d**), and the YW mapping population,  $n = 174$  individuals (**e**). Lod, log likelihood ratio.

establishment of the floral meristem and in the specification of floral organs<sup>40–43</sup>. *APETALA2* was classified as a class A floral homeotic gene, which specifies sepal identity if expressed alone and petal identity if expressed together with class B genes<sup>44</sup>. Furthermore, *AP2/TOE3* repressed *AGAMOUS* expression (a class C gene) in the two outer floral whorls in the floral meristem<sup>42</sup> (reviewed in ref. <sup>45</sup>). In rose, a reduction in the levels of *RhAGAMOUS* transcripts was proposed to be the basis of the double flower phenotype<sup>37</sup>. We hypothesise that misregulation of the rose *APETALA2/TOE* homologue (due to the presence of the

transposable element) is responsible for the *RhAGAMOUS* transcript level reduction, leading to the double flower phenotype.

Interestingly, a GWAS approach for petal number (a quantitative analysis) in a panel of tetraploid and double flower varieties<sup>33</sup> revealed that the most significant QTL is also located at the *DOUBLE FLOWER* locus (Fig. 4c). Several markers in this cluster display significant dose-dependent effects on the number of petals. One of these markers, RhK5\_4359\_382 (at position 33.55 Mb), was analysed via the Kompetitive allele specific PCR (KASP) technology both in the original association panel of 96 cultivars and in an

independent panel of 238 tetraploid varieties and showed the same effect in both populations (Supplementary Fig. 8a,b). Two other markers (RhK5\_14942 and RhMCRND\_760\_1045) were also tested on the 96 cultivars by KASP technology and revealed the same pattern (Supplementary Fig. 8c,d). This demonstrates a dual role of the *DOUBLE FLOWER* locus in rose: it controls both the double flower phenotype (double versus single flowers) and the number of petals. Given that the petal number QTL was detected in several panels of unrelated rose genotypes, it seems that this locus acts independently of the genetic background.

**Self-incompatibility.** As described for other Rosaceae species<sup>46–48</sup>, in some diploid roses, self-incompatibility is caused by a gametophytic SI (self-incompatibility) locus. This locus is most likely composed of genes encoding S-RNases and F-box proteins, which represent the female- and male-specific components, respectively. Previous approaches have failed to characterize the *Rosa* SI-locus genes owing to the low sequence similarity between S-RNase genes across species and the existence of multiple genes for both S-RNases and F-box proteins. A screen for S-RNase and F-box homologues in the HapOB genome sequence identified a region of 100 kb on Chr3 that contains three genes coding for S-RNases and four genes for S-locus F-box proteins (Fig. 4a and Supplementary Fig. 9a). This region is syntenic with the SI locus in *Prunus persica* (Supplementary Fig. 9b). One of the S-RNases (*S-RNase36*) was expressed in pistils of ‘Old-Blush’ flowers. Of the F-box genes, *Fbox38* accumulated in the stamens (Supplementary Fig. 9c,d). Hence, this region fulfils the requirements of a functional S-locus.

This region is consistent with previous data on segregation of the self-incompatibility phenotype in a diploid rose population, in which the self-incompatibility phenotype was analysed by generating a bi-parental progeny and backcrossing individual progeny to both parents<sup>49</sup>. We generated a marker for an orthologue of the S-RNase gene (*SRNase30*) expressed in pistils of ‘Old Blush’ that cosegregates with the S-locus at a distance of 4.2 cM. The large number of recombinants might be explained by incomplete expression of self-incompatibility (leaky phenotypes) in some individuals of the progeny, a phenomenon that is also observed in, for example, *Solanum* populations<sup>50</sup>.

**Prickle density.** We investigated the genetic regulation of prickle density in rose. In two F1 progenies, QTLs were detected on Chr3. In the OW and YW progenies, a large region of significant association was detected between position 31.2 Mb and the end of the chromosome on both male and female maps (Fig. 4d,e, respectively). In both populations, two peaks were clearly detected, which probably correspond to two neighbouring QTLs (Fig. 4a,d,e). Through a GWAS approach, we detected a strong association between SNPs and the presence of prickles between positions 31.0 Mb and 32.4 Mb (Supplementary Fig. 10a). In rose, prickles originate as a deformation of glandular trichomes in combination with cells from the cortex<sup>51</sup>. We have looked for homologues of candidate genes controlling trichome initiation and development identified in *A. thaliana*<sup>52</sup>. Screening the QTL region on Chr3 of HapOW for gene family members of these candidate genes revealed several WRKY transcription factors, of which RC3G0244800 (positioned at 33.40 Mb; Fig. 4a) shows strong similarity with *AtTTG2* (*TESTA TRANSPARENT GLABRA2*), which is involved in trichome development in *Arabidopsis*<sup>53</sup> (Supplementary Fig. 10b). We studied the expression of the rose *TTG2* homologue (*RcTTG2*) in three different individuals of the OW progeny with different prickle densities (absence, medium- and high-density prickles on the stem; Supplementary Fig. 10c). The *RcTTG2* transcript accumulated at higher levels in stems presenting prickles, suggesting that *RcTTG2* is a positive regulator of prickle presence in rose. This *TGG2* homologue represents a good candidate for the control of prickles in rose.

## Discussion

We have produced a high-quality reference rose genome sequence that will represent an essential resource for the rose community but also for rose breeders. Using this new reference sequence, we have analysed important structural features of the genome, including the position of the centromeres (Fig. 2) and SNP and indel frequencies (Fig. 3).

Taking advantage of this new high-quality reference sequence, rose is set to become a model species to study ornamental traits. For example, rose was previously used to study scent emission, leading to the discovery of a new pathway for the synthesis of monoterpenes<sup>54</sup>. Here, using a combination of genomic and genetic approaches (F1 progenies and GWAS diversity panel), we have demonstrated that this new reference sequence can be used to analyse loci controlling ornamental traits, such as continuous flowering, double flower, self-incompatibility and prickle density (Fig. 4). We have identified and characterized candidate genes for these traits. We propose that a rose *APETALA2/TOE* homologue controls the switch from simple to double flower and, unexpectedly, also the number of petals within double flowers. Further analyses are necessary to validate the function of these genes. The analyses were done in diploid roses but also in tetraploid roses, allowing direct implementation in rose breeding materials, with the development of diagnostic markers as we demonstrated for petal number. For this economically crucial trait, we have developed a genetic marker that permits the prediction of petal number, which we validated on a large panel (Supplementary Fig. 8). This represents a good example of how the development and release of the rose genome sequence can accelerate gains in rose breeding.

Cultivated roses have an allopolyploid background but segregate mainly tetrasomically<sup>10,55</sup>. Hence, rose is a unique model for polyploidization and chromosome pairing mechanisms, which can now also be investigated at the molecular level. This reference sequence opens the way to genomic and epigenomic approaches to study important traits, providing an essential bridge between this and other plant species.

## Methods

**Development of haploid ‘Old Blush’ callus.** Young flower buds of ‘Old Blush’ (Fig. 1c) with microspores at a mid-to-late uninucleate developmental stage (Fig. 1d) were collected in a greenhouse, wrapped in aluminium foil and stored in the dark at 4 °C for 25 days. These were then surface sterilized in 70% ethanol for 30 s and in sodium hypochlorite solution (2.9% active chloride) for 15 min followed by rinsing three times in double-distilled sterilized water.

Anthers were aseptically removed using binoculars and ground in starvation B medium<sup>56</sup> with minor modifications (pH 6 and 0.1 M sorbitol) for 2 min using a MSE homogenizer (Measuring & Scientific Equipment) set at 10,000 r.p.m. Anthers were then collected on 50- $\mu$ m mesh filters, covered with a fine layer of fresh modified starvation B medium and incubated for 24 h at 22 °C in darkness. Anthers were transferred on MS medium containing 30 g l<sup>-1</sup> sucrose, 0.5 mg l<sup>-1</sup> BAP (6-benzylaminopurine) and 0.1 mg l<sup>-1</sup> NAA (naphthaleneacetic acid) in 12-well culture plates. Plates were incubated in darkness at 23 °C/19 °C (16 h/8 h), taking care not to move the boxes or expose them to light for 80 days to induce somatic embryo formation. Somatic embryos were isolated from the anthers and transferred on the same medium in petri dishes with filter paper in 4-week intervals until the production of callus (Fig. 1e). Then, callus was multiplied on the same medium in the dark until enough material for DNA extraction was produced. Homozygosity was verified using ten previously described microsatellite markers<sup>57</sup>.

Genome sizes and ploidy levels were analysed on a flow cytometer, PASIII (488-nm, 20-mW laser; Partec). The Cystain absolute PI reagent kit (Sysmex) was used for sample preparation. *Solanum lycopersicum* ‘Stupické polni tyckove rane’ (1,916 Mb/2C) was used as an internal standard.

**Genome sequencing and assembly.** *DNA extraction for PacBio and Illumina sequencing.* Callus tissues of the haploid ‘Old Blush’ HapOB line was kept in the dark for 3 days prior to DNA extraction to reduce chloroplast DNA contamination. DNA extraction was performed on 1 g HapOB callus tissue as described previously<sup>58</sup>. In total, approximately 30 mg genomic DNA was obtained in several batches for the preparation of three independent single-molecule real-time (SMRT) bell libraries. For the first library, genomic DNA was sheared by a Megaruptor (Diagenode) device with 30-kb settings. Sheared DNA was purified and



concentrated with AMPureXP beads (Agencourt) and further used for SMRTbell preparation according to the manufacturer's protocol (Pacific Biosciences; 20-kb template preparation using BluePippin (SageScience) size selection system with a 15-kb cut-off). Two additional libraries were made excluding the DNA shearing step, but with an additional initial damage repair. Size-selected and -isolated SMRTbell fractions were purified using AMPureXP beads and finally used for primer and polymerase (P6) binding according to the manufacturer's binding calculator (Pacific Biosciences). Three library DNA-polymerase complexes were used for Magbead binding and loaded at 0.16, 0.25 and 0.20 nM on-plate concentrations, using 12, 7 and 8 SMRT cells, respectively. Final sequencing was done on a PacBio RS-II platform, with a 345- or 360-min movie time, 1 cell per well protocol and C4 sequencing chemistry. Raw sequence data were imported and further processed on a SMRT Analysis Server v2.3.0.

For Illumina sequencing, approximately 200 ng genomic DNA was sheared in a 55- $\mu$ l volume using a Covaris E210 device to approximately 500–600 bp. One library with an insert size of 720 bp was made using Illumina TruSeq Nano DNA Library Preparation Kit according to the manufacturer's guidelines. The final library was quantified by Qubit fluorescence spectrophotometry (Invitrogen) and the library fragment size range was assessed by Bioanalyzer High Sensitivity DNA assay (Agilent). The library was used for clustering as part of two lanes of a paired-end flow cell v4 using a Cbot device and subsequent  $2 \times 125$  paired-end sequencing on a HiSeq2500 system (Illumina). De-multiplexing was carried out using Casava 1.8 software.

**Genome assembly, polishing and contamination assessment.** All sequence data generated that were derived from 27 SMRT cells containing 19.2 Gb of reads larger than 500 bp were assembled with CANU hierarchical assembler v1.4 (ref. 10) (version release r8046). In general, default settings were used except 'corMinCoverage', which was changed from 4 to 3, 'minOverlapLength', which was increased from 500 to 1,000, and 'errorRate', which was adjusted to 0.015. The assembly was completed on the Dutch National e-Infrastructure with the support of SURF Cooperative using 2,024 CPU hours (Intel Xeon Haswell 2.6 GHz) for the complete CANU process. Illumina paired-end ( $2 \times 125$  bp) reads were mapped onto the genome assembly using Burrow-Wheeler aligner maximum exact match (BWA-MEM)<sup>59</sup>. Pilon<sup>60</sup> was then used to error correct the assembly. This procedure was repeated three times iteratively.

For contamination assessment, prokaryotic genes were predicted on the contigs using MetaGeneAnnotator<sup>61</sup>. The number of genes per nucleotide was computed for every contig. Furthermore, Illumina reads were mapped on the contigs using BWA-MEM<sup>62</sup>. The number of mapped reads per nucleotide was computed for every contig. Contigs with a low Illumina read mapping frequency were aligned against the GenBank non-redundant protein database using BLASTX.

#### Development of high-density genetic maps and GWAS analysis. *Plant material.*

A diploid F1 population of 151 individuals (OW) was obtained by crossing *R. chinensis* 'Old Blush' and a hybrid of *R. wichurana* obtained from Jardin de Bagatelle (Paris, France). This population was planted at the INRA Experimental Unit Horti (Beaucouzé, France).

A diploid F1 population of 174 individuals (YW) was obtained from a cross between 'Yesterday' and *R. wichurana* (the extended population as used in ref. 63). This population was planted at the ILVO (Melle, Belgium).

The tetraploid K5 cut rose mapping population consisted of 172 individuals obtained from a cross between P540 and P867. It was planted in Wageningen, the Netherlands, and was previously used in various QTL studies<sup>64,65</sup>.

The association panel comprised 96 cultivars, of which 87 were tetraploid, 8 were triploid and 1 was diploid, selected to reduce the genetic relatedness between genotypes<sup>33</sup>. Plants were cultivated in a randomized block design, with three blocks comprising one clone of each genotype both in the greenhouse and at an experimental field location at Leibniz Universität Hannover, Germany. For marker validation, an independent population of 238 tetraploid varieties was used that was cultivated in a field plot of the Federal Plant Variety Office in Hannover, Germany. Plants of the association panel and the phenotypic data are described in Supplementary Table 12.

**Genetic map construction.** The construction of the different genetic maps from F1 progenies (OW, YW and K5), the KASP assay for SNP validation and the development of a sequence characterized amplified region (SCAR) marker for the SI locus are described in Supplementary Methods.

**GWAS analysis.** The GWAS analyses for petal numbers and prickles density were performed in TASSEL 3.0 (ref. 66) as described previously<sup>33</sup>. Trait marker association for petal number was analysed using the mixed linear model (MLM) and 39,831 markers (petal as a quantitative trait with the Q + K model), including a fixed effect as the population structure matrix (Q) and random effect as the kinship matrix (K). Significance thresholds were corrected for multiple testing by the Bonferroni method using the number of contigs (19,083) as a correction factor, resulting in a significance threshold of  $1.78 \times 10^{-6}$ . The kinship matrix used in the MLM was calculated for 10,000 SNP markers with the software SPAGeDi 1.5 (Zitai) as described previously<sup>33</sup>. For the GWAS analysis of prickles and petals

with the general linear model (GLM) in TASSEL 3.0 (ref. 66), 63,000 markers were analysed. Petals and prickles were set as qualitative traits (1 and 0 to indicate presence or absence, respectively), and the analysis was performed without any correction for population (Q + K). Significance thresholds in the GLM were corrected by the number of contigs (28,054) to  $1.78 \times 10^{-6}$ .

**Alignment of the HapOB rose genome with the OW genetic maps.** The alignment of the genetic and physical maps was done in two steps. First, the HapOB sequence was aligned to the integrated genetic maps to detect problems of assembly (contigs that are present on two linkage groups). Second, to precisely order and orient the contigs on each linkage group, the alignment was done separately on the male and female maps and manually integrated.

During the first step, 7,822 out of a total of 7,840 SNP markers were positioned by mapping the corresponding 70-bp probes onto the HapOB genome sequence using Blat v.35 (ref. 67). Markers with more than one best hit were eliminated. Out of the 7,360 remaining markers, 6,808 passed the mapping quality filter ( $\geq 95\%$  match and  $\leq 4\%$  mismatch). Of these, 6,746 markers belonging to the most common linkage group on their respective contigs were conserved and described as 'concordant' markers. Only contigs with more than one of these markers were retained.

During the second step, the mapping and anchoring were done independently on the male and female maps (Table 1). The procedure and conditions were the same as for the first mapping. Only concordant markers were kept (4,875 (87%) and 1,871 (81%) for the female and male map, respectively). We positioned and oriented the different contigs manually (Supplementary Table 13). When a contig spanned several loci, its order and position were clear. However, for some contigs, genetic maps did not resolve the orientation problems. In these situations, we used the synteny between *Rosa* and *F. vesca*<sup>10</sup>. The strategy used to position and orient contigs is described in Supplementary Fig. 11. The position and orientation of the contigs are listed in Supplementary Table 13.

Concerning the K5 integrated genetic map, among the 25,695 SNP markers present, 20,706 SNPs (80.6%) could be positioned on the HapOB genome sequence by BLAST of the SNP-flanking marker sequences (Supplementary Fig. 2a).

**Centromere region identification and FISH.** Three complementary tools were used to identify centromeric tandem repeats and to estimate their abundance in the *R. chinensis* 'Old Blush' genome: Tandem Repeat Finder (TRF)<sup>68</sup>, TAREAN<sup>69</sup> and RepeatExplorer<sup>70</sup>, each with default settings, and the output was parsed using custom python scripts. All tandem repeats identified by TRF were subjected to all-against-all BLAST to cluster similar repeats and to estimate abundance (the total number of tandem repeat cluster copies) in the genome. Paired reads were quality filtered and trimmed to 120 bp for analysis by RepeatExplorer (0.5 M read pairs) and TAREAN (1.3 M read pairs). RepeatExplorer cluster CL226 had the globular-like shape specific for tandem repeats. The corresponding monomer repeat sequence was identified by analysing the contigs of this cluster with TRF. The identical tandem repeat was also identified by TAREAN and TRF. To determine the location of the CL226 tandem repeat cluster in the genome assembly, 275 M paired-end genomic reads of 'Old Blush' were mapped onto the contigs from RepeatExplorer cluster CL226, using Bowtie2 (ref. 71) with parameter -k 1 to select read pairs with high similarity to the CL226 repeat. Selected read pairs were then split into two groups: reads that matched the CL226 repeat sequence itself and reads that matched the flanking genome sequence. Both groups of reads were separately mapped onto the genomic scaffolds using Bowtie with parameters -a 1 and -N 1. The distribution of the two sets of CL226 reads was visualized using the circlize package<sup>72</sup> of R Bioconductor<sup>73</sup>. Mitotic chromosome slides were prepared with the 'SteamDrop' method<sup>74</sup> using young root meristems of *R. chinensis* 'Old Blush'. Two oligonucleotide probes (5'-TTGCGTTGTCTAGTGACATTCA-TAMRA-3'; 5'-ACCCTAGAAGCGAGAAGTTTGG-TAMRA-3') were used for FISH, as previously described<sup>75</sup>. DRAWID<sup>76</sup> was used for chromosome and signal analysis.

**Annotation of the rose genome.** Gene and transposable element annotations are described in Supplementary Methods.

**Diversity analysis.** The plant material originated from 'Loubert Nursery' in Rosier-sur-Loire, France (*R. persica*), from 'Rose Loubert' rose garden in Rosier-sur-Loire, France (*R. moschata*, *R. xanthina spontanea* and *R. gallica*) and from 'Rosaerie du Val de Marne', Haÿ-Les-Roses, France (*R. chinensis* var. *spontanea*, *R. rugosa*, *R. laevigata* and *R. minutifolia alba*).

Illumina paired-end shotgun indexed libraries were prepared from 3  $\mu$ g DNA per accession, using the TruSeq DNA PCR-Free LT Kit (Illumina). Briefly, indexed library preparation was performed with low-sample protocol with a special development to reach an insert size of 1–1.5 kb. DNA fragmentation was performed by AFA (Adaptive Focused Acoustics) technology on the focused ultrasonicator E210 (Covaris). All enzymatic steps and clean up were done according to the manufacturer's instructions, apart from the fragmentation and sizing steps. Paired-end sequencing using  $2 \times 150$  sequencing-by-synthesis cycles was performed on a HiSeq 2000/2500, Rapid TruSeq V2 chemistry (Illumina) running in rapid mode using on-board cluster generation (according to the

manufacturer's instructions). For some read sets, a low enrichment of libraries with five PCR amplification cycles was performed.

Cutadapt and FASTX toolkit software were used for quality control ( $Q > 30$ ), and adapter trimming and high-quality reads were considered for further analysis. To identify the SNPs and indels in each species, filtered paired-end reads were mapped against the HapOB reference using BWA with default parameters<sup>77</sup>. The BWA software produced highly accurate alignment compared to other software. Unmapped and duplicated reads were removed using SAMtools and the Picard package, respectively<sup>78</sup>. Furthermore, reliable mapped reads were used for base quality score recalibration and indel realignment using the Genome Analysis Toolkit (GATK) software<sup>79</sup>. We then called variants individually on each sample using the HaplotypeCaller/GATK. The identified SNPs and indels were filtered out on the bases of a minimum read depth of 20 and SNP quality ( $Q$ )  $\geq 40$ . The genomic distribution of SNPs and indels was analysed by calculating their frequency over each 200-kb interval on each HapOB chromosome. Circos was used to visualize the distribution of SNPs and indels on each HapOB chromosome. SnpEff and SnpSift<sup>80,81</sup> were used to annotate the effects of SNPs and identify the potential functional effects of amino acid substitution on corresponding proteins, respectively.

To infer phylogenetic relationships between *Rosa* species, homozygous SNPs from each VCF file were merged using GATK CombineVariants and parsed to build a SNP alignment using VCFtools and our own scripts. A maximum likelihood analysis was performed using RAxML v8.1.5 with 100 bootstrap replicates<sup>82</sup>. As the SNP alignment contains only variable sites, an ascertainment bias correction was applied to the GTRGAMMA model of substitution<sup>83</sup>. The resulting phylogenetic tree was rooted on *R. persica*, which was purported to be the most divergent *Rosa* species<sup>84</sup>.

To conduct the synteny analysis between the HapOB reference sequence and *F. vesca*, orthologous genes were identified using reciprocal BLAST with an  $e$ -value of  $1 \times 10^5$  (ref. <sup>85</sup>),  $v = 5$  and  $b = 5$ . The protein sequences and annotation for *F. vesca* (v2.0.a1) were downloaded from the GDR database (<https://www.rosaceae.org/>). The output of the BLAST tool was used in the McSCANX tool to identify syntenic regions between the genomes<sup>86</sup>. The circos software<sup>87</sup> was used to visualize the syntenic regions between two genomes. In addition, an analysis of microsynteny was performed between *R. chinensis* 'Old blush' and *F. vesca* for Chr3 to see the conserved region near the *RoKSN* locus using Symap software<sup>85</sup>.

Good-quality and pre-processed Illumina reads of *R. laevigata* were used for assembly. Genomic sequence reads were assembled using SPAdes (v3.11.1) with a  $k$ -mer value of 63 (ref. <sup>88</sup>).

**Morphological traits.** *Petal number.* For the OW and YW populations (151 and 174 individuals, respectively), the number of petals per flower was counted using 5 or up to 10 independent flowers, respectively. In roses, single flowers typically have five petals. Flowers with fewer than eight petals were considered as simple flowers, whereas those with eight or more petals were considered as 'double' flowers.

For the GWAS panel, the number of petals was counted for three flowers on each of the three clones from greenhouse-grown plants, and the arithmetic means were calculated for each genotype.

*Prickle number.* In the OW and YW populations, the length of a stem part with four internodes was measured in the middle of a stem (between the fifth and seventh internodes). Prickles were counted on four internodes. The prickle density was expressed as the number of prickles per internode. For each genotype, three stems were measured and counted.

For the GWAS panel, prickle density was calculated as the arithmetic mean of the number of prickles between the third and fourth node of newly developed shoots. For each genotype, three shoots were counted from three replicates in a randomized block design.

*Expression analysis.* For *TTG2* expression analysis, three individuals of the OW progeny were selected according to prickle density: OW9068 (no prickles), OW9155 (low density) and OW9106 (high density). The terminal part of young stems was harvested in spring 2016 from field-grown plants (two biological replicates). RNA extraction, cDNA synthesis, qPCR (three technical replicates) and relative quantifications were performed as previously described<sup>89</sup>. Calibration was done using *TCTP* and *UBC* genes. The following primers were used to amplify *TTG2* (RcTTG2-1-F: CCTCAAAACCCAGGAGCATC and RcTTG2-1-R: CAACAGCTTGATCCCTGAGAG).

Organ-specific expression of candidate self-incompatibility genes were tested using RNA extracted from the stamens and pistils of three flower buds and five open flowers and the terminal leaflets of three young leaves, sampled from an individual of 'Old Blush' in August 2017. RNA extraction was carried out according to previously published protocols<sup>37</sup>. cDNA synthesis and RT-PCR were performed with the PrimeScript RT reagent Kit with genomic DNA Eraser and EmeraldAmp PCR Master Mix (TaKaRa) according to the manufacturer's protocols. The following primers (5' to 3') were used to amplify seven candidate genes and a house-keeping gene: *SRNase26* (F1: TGCAGCAACACATACGATT and R1: GCAAGAAGATCGGCGTAGTC), *SRNase30* (F1: TGTTCACAAATGGCCGATAA and

R1: TGCACATAAGCGAAGGAGTG), *SRNase36* (F1: TGTGGTAACAGCTGCAAAGC and R1: TCAACCACGTTTTTGGCATA), *Fbox29* (F2: TGACTATTTTCTATGCGCTTGAG and R1: CACCACAAAAAGGATAACAAGAC), *Fbox31* (F1: TTTGCTATGAAAATGATAACAACAG and R1: AACCCCATGGTTTCATTAAGTA), *Fbox38* (F1: GACTACTCTCTTTGGCCTGAA and R1: CTACAGCTGCAGAATCATTGAC), *Fbox40* (F1: CGTCCAATATCTCTACTCAATGGT and R1: CCTCTTCTGGTGAGTCTGAAAT) and *RoTCTP* (F2: AAGAAGCAGTTTGTACATGG and R2: TCTTAGCACTTGACCTCTTCA).

**Reporting Summary.** Further information on experimental design is available in the Nature Research Reporting Summary linked to this article.

**Code availability.** The R code used for pairwise maximum likelihood recombination and lod score calculations is available through CRAN (<https://CRAN.R-project.org/package=polymapR>). The R code used to infer phylogenetic relationships is available on request from the corresponding author<sup>90</sup>. The python scripts used for centromeric region identification are available on request from the corresponding author.

**Data availability.** All the genome data have been made available on a genome browser (<https://iris.angers.inra.fr/obh/>) and in the public GDR database ([https://www.rosaceae.org/species/rosa/chinensis/genome\\_v1.0](https://www.rosaceae.org/species/rosa/chinensis/genome_v1.0))<sup>91</sup>. FASTA files of chromosomes and genes (mRNA, proteins and non-coding RNA) and gff files for gene models and structural features (transposable element) can be downloaded from both the previously mentioned websites. Raw data (PacBio and Illumina reads) are available under the accession number PRJNA445774. RNA-seq data used for genome annotation are available under the following SRA accession numbers: SRP128461 for 91/100-5 leaves infected with blackspot and SRP133785 for *R. wichurana* and 'Yesterday' leaves infected with two powdery mildew pathotypes. Raw data of resequencing of the eight wild *Rosa* species are available under the SRA accession number SRP143586.

Received: 30 January 2018; Accepted: 1 May 2018;  
Published online: 11 June 2018

## References

- Wang, G. A study on the history of Chinese roses from ancient works and images. *Acta Hort.* **751**, 347–356 (2007).
- Pliny (2013) *Pine L'Antienne: Histoire naturelle* (Schmit, S., Trans) *Bibliothèque de la Pléiade* No. 593 (Gallimard, Paris, 2013).
- Nybohm, H. & Werlemark, G. Realizing the potential of health-promoting rosehips from dogroses (*Rosa* sect. *Caninae*). *Curr. Bioact. Compd.* **13**, 3–17 (2017).
- Zhang, J. et al. The diploid origins of allopolyploid rose species studied using single nucleotide polymorphism haplotypes flanking a microsatellite repeat. *J. Hort. Sci. Biotechnol.* **88**, 85–92 (2013).
- Ritz, C. M. & Wisseman, V. Microsatellite analyses of artificial and spontaneous dogroses hybrids reveal the hybridogenic origin of *Rosa micrantha* by the contribution of unreduced gametes. *J. Hered.* **102**, 2117–2127 (2011).
- Meng, J., Fougère-Danezan, M., Zhang, L.-B., Li, D.-Z. & Yi, T.-S. Untangling the hybrid origin of the Chinese tea roses: evidence from DNA sequences of single-copy nuclear and chloroplast genes. *Plant Syst. Evol.* **297**, 157–170 (2011).
- Wisseman, V. & Ritz, C. M. The genus *Rosa* (Rosoideae, Rosaceae) revisited: molecular analysis of nrITS-1 and *atpB-rbcL* intergenic spacer (IGS) versus conventional taxonomy. *Bot. J. Linn. Soc.* **147**, 275–290 (2005).
- Jian, H. et al. Decaploidy in *Rosa praelucens* Byhouwer (Rosaceae) endemic to Zhongdian Plateau, Yunnan, China. *Caryologia.* **63**, 162–167 (2012).
- Robert, A. V., Gladis, T. & Brumme, H. DNA amounts of roses (*Rosa* L.) and their use in attributing ploidy levels. *Plant Cell Rep.* **28**, 61–71 (2009).
- Bourke, P. M. et al. Partial preferential chromosome pairing is genotype dependent in tetraploid rose. *Plant J.* **90**, 330–343 (2017).
- Herklotz, V. & Ritz, C. M. Multiple and asymmetrical origin of polyploid dog rose hybrids (*Rosa* L. sect. *Caninae* (DC.) Ser.) involving unreduced gametes. *Ann. Bot.* **120**, 209–220 (2017).
- Ritz, C. M., Köhnen, I., Groth, M., Theissen, G. & Wisseman, V. To be or not to be the odd one out—allele-specific transcription in pentaploid dogroses (*Rosa* L. sect. *Caninae* (DC.) Ser.). *BMC Plant Biol.* **11**, 37 (2011).
- Liorzou, M. et al. Nineteenth century French rose (*Rosa* sp.) germplasm shows a shift over time from a European to an Asian genetic background. *J. Exp. Bot.* **67**, 4711–4725 (2016).
- Nakamura, N. et al. Genome structure of *Rosa multiflora*, a wild ancestor of cultivated roses. *DNA Res.* **25**, 113–121 (2018).

15. Wylie, A. P. The history of garden roses. *J. R. Horticult. Soc.* **79**, 555–571 (1954).
16. Koren, S. et al. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* **27**, 722–736 (2017).
17. Koning-Boucoiran, C. F. et al. Using RNA-seq to assemble a rose transcriptome with more than 13,000 full-length expressed genes and to develop the WagRhSNP 68k Axiom SNP array for rose (*Rosa* L.). *Front. Plant Sci.* **6**, 249 (2015).
18. Foissac, S. et al. Genome annotation in plants and fungi: EuGene as a model platform. *Curr. Bioinform.* **3**, 87–97 (2008).
19. Jones, P. et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
20. Finn, R. D. et al. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* **44**, D279–D285 (2016).
21. Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
22. Edger, P. P. et al. Single-molecule sequencing and optical mapping yields an improved genome of woodland strawberry (*Fragaria vesca*) with chromosome-scale contiguity. *GigaScience* **7**, 1–7 (2018).
23. Flutre, T., Duprat, E., Feuillet, C. & Quesneville, H. Considering transposable element diversification in de novo annotation approaches. *PLoS ONE* **6**, e16526 (2011).
24. Potter, D. et al. Phylogeny and classification of Rosaceae. *Plant Syst. Evol.* **266**, 5–43 (2007).
25. Xiang, Y. et al. Evolution of Rosaceae fruit types based on nuclear phylogeny in the context of geological times and genome duplication. *Mol. Biol. Evol.* **34**, 262–281 (2017).
26. Gar, O. et al. An autotetraploid linkage map of rose (*Rosa hybrida*) validated using the strawberry (*Fragaria vesca*) genome sequence. *PLoS ONE* **6**, e20463 (2011).
27. Bruneau, A., Starr, J. R. & Joly, S. Phylogenetic relationships in the genus *Rosa*: new evidence from chloroplast DNA sequences and an appraisal of current knowledge. *Syst. Bot.* **32**, 366–378 (2007).
28. Fougère-Danezan, M., Joly, S., Bruneau, A., Gao, X.-F. & Zhang, L.-B. Phylogeny and biogeography of wild roses with specific attention to polyploids. *Ann. Bot.* **115**, 275–291 (2015).
29. Fernández-Romero, M. D., Torres, A. M., Millán, T., Cubero, J. I. & Cabrera, A. Physical mapping of ribosomal DNA on several species of the subgenus *Rosa*. *Theor. Appl. Genet.* **103**, 835–838 (2001).
30. The 100 Tomato Genome Sequencing Consortium et al. Exploring genetic variation in the tomato (*Solanum section Lycopersicon*) clade by whole-genome sequencing. *Plant J.* **80**, 136–148 (2014).
31. Duan, N. et al. Genome re-sequencing reveals the history of apple and supports a two-stage model for fruit enlargement. *Nat. Commun.* **8**, 249 (2017).
32. Nguyen, T. H. N., Schulz, D., Winkelmann, T. & Debener, T. Genetic dissection of adventitious shoot regeneration in roses by employing genome-wide association studies. *Plant Cell Rep.* **36**, 1493–1505 (2017).
33. Schulz, D. F. et al. Genome-wide association analysis of the anthocyanin and carotenoid contents of rose petals. *Front. Plant Sci.* **7**, 1798 (2016).
34. Iwata, H. et al. The *TFL1* homologue *KSN* is a regulator of continuous flowering in rose and strawberry. *Plant J.* **69**, 116–125 (2012).
35. Koskela, E. A. et al. Mutation in *TERMINAL FLOWER1* reverses the photoperiodic requirement for flowering in the wild strawberry *Fragaria vesca*. *Plant Physiol.* **159**, 1043–1054 (2012).
36. Shulaev, V. et al. The genome of woodland strawberry (*Fragaria vesca*). *Nat. Genet.* **43**, 109–116 (2010).
37. Dubois, A. et al. Tinkering with the C-function: a molecular frame for the selection of double flowers in cultivated roses. *PLoS ONE* **5**, e9288 (2010).
38. Roman, H. et al. Genetic analysis of the flowering date and number of petals in rose. *Tree Genet. Genomes* **11**, 85 (2015).
39. Shigyo, M., Hasebe, M. & Ito, M. Molecular evolution of the AP2 subfamily. *Gene* **366**, 256–265 (2006).
40. Bowman, J. L., Alvarez, J., Weigel, D., Meyerowitz, E. M. & Smyth, D. R. Control of flower development in *Arabidopsis thaliana* by *APETALA1* and interacting genes. *Development* **119**, 721–743 (1993).
41. Bowman, J. L., Smyth, D. R. & Meyerowitz, E. M. Genes directing flower development in *Arabidopsis*. *Plant Cell* **1**, 37–52 (1989).
42. Jung, J.-H., Lee, S., Yun, J., Lee, M. & Park, C.-M. The miR172 target TOE3 represses *AGAMOUS* expression during *Arabidopsis* floral patterning. *Plant Sci.* **215–216**, 29–38 (2014).
43. Zhang, B., Wang, L., Zeng, L., Zhang, C. & Ma, H. *Arabidopsis* TOE proteins convey a photoperiodic signal to antagonize *CONSTANS* and regulate flowering time. *Genes Dev.* **29**, 975–987 (2015).
44. Bowman, J. L., Smyth, D. R. & Meyerowitz, E. M. Genetic interactions among floral homeotic genes of *Arabidopsis*. *Development* **112**, 1–20 (1991).
45. ÓMaoiléidigh, D. S., Graciet, E. & Wellmer, F. Gene networks controlling *Arabidopsis thaliana* flower development. *New Phytol.* **201**, 16–30 (2014).
46. Ashkani, J. & Rees, D. J. G. A comprehensive study of molecular evolution at the self-incompatibility locus of Rosaceae. *J. Mol. Evol.* **82**, 128–145 (2016).
47. Charlesworth, D., Vekemans, X., Castric, V. & Glemin, S. Plant self-incompatibility systems: a molecular evolutionary perspective. *New Phytol.* **168**, 61–69 (2005).
48. McClure, B., Cruz-García, F. & Romero, C. Compatibility and incompatibility in S-RNase-based systems. *Ann. Bot.* **108**, 647–658 (2011).
49. Debener, T. et al. Genetic and molecular analysis of key loci involved in self-incompatibility and floral scent in roses. *Acta Horticult.* **870**, 183–190 (2010).
50. Mena-Ali, J. I. & Stephenson, A. G. Segregation analyses of partial self-incompatibility in self and cross progeny of *Solanum carolinense* reveal a leaky S-allele. *Genetics* **177**, 501–510 (2007).
51. Kellogg, A. A., Branaman, T. J., Jones, N. M., Little, C. Z. & Swanson, J. D. Morphological studies of developing *Rubus* prickles suggest that they are modified glandular trichomes. *Botany* **89**, 217–226 (2011).
52. Pattanaik, S., Patra, B., Singh, S. K. & Yuan, L. An overview of the gene regulatory network controlling trichome development in the model plant, *Arabidopsis*. *Front. Plant Sci.* **5**, 259 (2014).
53. Johnson, C. S., Kolevski, B. & Smyth, D. R. *TRANSPARENT TESTA GLABRA2*, a trichome and seed coat development gene of *Arabidopsis*, encodes a WRKY transcription factor. *Plant Cell* **14**, 1359–1375 (2002).
54. Magnard, J.-L. et al. Biosynthesis of monoterpene scent compounds in roses. *Science* **349**, 81–83 (2015).
55. Koning-Boucoiran, C. F. S. et al. The mode of inheritance in tetraploid cut roses. *Theor. Appl. Genet.* **125**, 591–607 (2012).
56. Kyo, M. & Harada, H. Control of the developmental pathway of tobacco pollen in vitro. *Planta* **168**, 427–432 (1986).
57. Hibrand-Saint Oyant, L., Crespel, L., Rajapakse, S., Zhang, L. & Foucher, F. Genetic linkage maps of rose constructed with new microsatellite markers and locating QTL controlling flowering traits. *Tree Genet. Genomes* **4**, 11–23 (2008).
58. Daccord, N. et al. High-quality de novo assembly of the apple genome and methylome dynamics of early fruit development. *Nat. Genet.* **49**, 1099–1106 (2017).
59. Li, H. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics* **30**, 2843–2851 (2014).
60. Walker, B. J. et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE* **9**, e112963 (2014).
61. Noguchi, H., Taniguchi, T. & Itoh, T. MetaGeneAnnotator: detecting species-specific patterns of ribosomal binding site for precise gene prediction in anonymous prokaryotic and phage genomes. *DNA Res.* **15**, 387–396 (2008).
62. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv* **1303**, 3997 (2013).
63. Hosseini Moghaddam, H., Leus, L., De Riek, J., Van Huylenbroeck, J. & Van Bockstaele, E. Construction of a genetic linkage map with SSR, AFLP and morphological markers to locate QTLs controlling pathotype-specific powdery mildew resistance in diploid roses. *Euphytica* **184**, 413–427 (2012).
64. Gitonga, V. W. et al. Inheritance and QTL analysis of the determinants of flower color in tetraploid cut roses. *Mol. Breed.* **36**, 143 (2016).
65. Yan, Z., Dolstra, O., Prins, T. W., Stam, P. & Visser, P. B. Assessment of partial resistance to powdery mildew (*Podospheera pannosa*) in a tetraploid rose population using a spore-suspension inoculation method. *Eur. J. Plant Pathol.* **114**, 301–308 (2006).
66. Bradbury, P. J. et al. TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* **23**, 2633–2635 (2007).
67. Kent, W. J. BLAT—the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
68. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
69. Novák, P. et al. TAREAN: a computational tool for identification and characterization of satellite DNA from unassembled short reads. *Nucleic Acids Res.* **45**, e111 (2017).
70. Novak, P., Neumann, P., Pech, J., Steinhaisl, J. & Macas, J. RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. *Bioinformatics* **29**, 792–793 (2013).
71. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
72. Gu, Z., Gu, L., Eils, R., Schlesner, M. & Brors, B. Circlize implements and enhances circular visualization in R. *Bioinformatics* **30**, 2811–2812 (2014).
73. Gentleman, R. C. et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **5**, R80 (2004).
74. Kirov, I., Divashuk, M., Van Laere, K., Soloviev, A. & Khrustaleva, L. An easy “SteamDrop” method for high quality plant chromosome preparation. *Mol. Cytogenet.* **7**, 21 (2014).
75. Kirov, I. V., Van Laere, K., Van Roy, N. & Khrustaleva, L. I. Towards a FISH-based karyotype of *Rosa* L. (Rosaceae). *Comp. Cytogenet.* **10**, 543–554 (2016).



76. Kirov, I. V. et al. DRAWID: user-friendly java software for chromosome measurements and idiogram drawing. *Comp. Cytogenet.* **11**, 747–757 (2017).
77. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
78. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).
79. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
80. Cingolani, P. et al. Using *Drosophila melanogaster* as a model for genotoxic chemical mutational studies with a new program, SnpSift. *Front. Genet.* **3**, 35 (2012).
81. Cingolani, P. et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w(1118); iso-2; iso-3. *Fly* **6**, 80–92 (2012).
82. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
83. Lewis, P. O. A likelihood approach to estimating phylogeny from discrete morphological character data. *Syst. Biol.* **50**, 913–925 (2001).
84. Du Mortier, B. C. *Notice sur un Nouveau Genre de Plantes: Hulthemia; Précédée d'un Aperçu sur la Classification des Roses* (Casterman, J., 1824).
85. Lyons, E. et al. Finding and comparing syntenic regions among *Arabidopsis* and the outgroups papaya, poplar, and grape: CoGe with rosids. *Plant Physiol.* **148**, 1772–1781 (2008).
86. Wang, Y. et al. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* **40**, e49 (2012).
87. Krzywinski, M. et al. Circos: an information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645 (2009).
88. Bankevich, A. et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).
89. Randoux, M. et al. Gibberellins regulate the transcription of the continuous flowering regulator, RoKSN, a rose TFL1 homologue. *J. Exp. Bot.* **63**, 6543–6554 (2012).
90. Jung, S. et al. The Genome Database for Rosaceae (GDR): year 10 update. *Nucleic Acids Res.* **42**, D1237–D1244 (2014).

## Acknowledgements

We thank the ImHorPhen team of IRHS and the experimental unit (UE Horti) for their technical assistance in plant management. We thank the PTM ANAN (M. Bahut) of the SFR Quasav and the Gentyane platforms (especially C. Poncet) for the SSR and SNP analyses, respectively. We acknowledge A. Chauveau and I. Le Clainche for libraries preparation and E. Marquand and A. Canaguier for data processing. This work was supported by CEA-IG/CNG, by conducting the DNA quality control and by providing access to the INRA-EPGV group for their Illumina Sequencing Platform. We acknowledge J.-L. Gaignard (from the communication service of the INRA) for his help to fund the project. We thank the GDR team, and particularly P. Zheng, S. Jung and D. Main, for management of the genome sequence at the GDR database. We thank 'Région Pays de la Loire' for funding the sequencing of HapOB (Rose Genome Project), the resequencing of eight wild species (Genorose project in the framework of RFI 'Objectif Végétal') and for the EPICENTER ConnecTalent grant of the Pays de la Loire (N.D. and E.B.). F.F. and L.H.S.-O. thank the ANR for funding the genetic determinism of flower development (ANR-13-BSV7-0014). K.K. thanks the JSPS for funding the analysis of the S-locus (JSPS KAKENHI no.17H04616). T.D. thanks the German Ministry of Economic Affairs for funding the GWAS analysis (Aif programme ZI) and the Deutsche Forschungsgemeinschaft for the RNA-seq data generation (DFG program GRK1798). The development of the high-density SNP maps was partly funded by TTI Green Genetics and by the TKI Polyploids projects (BO-26.03-002-001 and BO-50-002-022).

## Author contributions

L.H.S.-O. developed the OW genetic map, analysed the haploid and performed the genetic determinism studies on the OW progeny. I.K. performed and interpreted the analyses of the centromeric regions. K.V.L. performed the FISH analysis. L.L. performed the cytometric analysis of the HapOB line. T.R., L.L. and J.D.R. developed the YW genetic map. J.D.R. and T.R. aligned the YW genetic map to the HapOB reference sequence. J.D.R. and L.L. performed the QTL analyses on prickles and flower traits in YW. T.R. analysed the candidate genes in QTLs. L.H. developed the haploid line. D.L. performed the synteny and diversity analyses. K.D. performed the phylogenetic analysis. P.M.B. developed and aligned the K5 map to the HapOB reference sequence and analysed Chr0. N.N.Z. analysed the genetic basis of prickle density and studied the *TTG2* candidate gene. N.D. performed sequence polishing and anchoring of the reference sequence to the OW genetic map. D.S., E.N. and M.L. contributed to the GWAS approach and developed KASP markers. E.N. generated part of the RNA-seq data. S.B. produced the haploid DNA for sequencing. T.T. developed and maintained the F1 OW individuals. A.C. analysed the SNP data of the OW progeny. J.J. analysed the candidate genes for the double flower. L.V. contributed to the production of the haploid. S.G. developed the genome browser. T.J.A.B. and P.A. contributed to the development of the K5 genetic map and its alignment to the reference sequence. R.E.V. and C.M. contributed to the K5 and OW genetic maps. H.V.d.G., T.H. and E.S. performed the rose genome sequencing and assembly. M.C.L.P., A.B. and R.B. performed the wild species resequencing. J.C. coordinated the diversity analysis. N.C. and H.Q. performed the transposable element annotation. S.A. performed the gene annotation. K.K. performed the SI locus analysis. S.S. contributed to financial support and discussion for the haploid line development. M.J.M.S. contributed to the K5 analysis and to the management of the project. T.D. developed the GWAS approach and some of the RNA-seq experiments, contributed to the genetic determinism analysis (*DOUBLE FLOWER* and SI loci) and to the management of the project. E.B. managed the haploid sequencing. F.F. performed the AP2 analysis and genome anchoring to the OW genetic map, coordinated the project and the writing of the manuscript. F.F., L.H.S.-O., T.R., P.M.B., M.J.M.S., T.D. and J.D.R. were major contributors to the writing of the manuscript. All authors read and approved the final manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41477-018-0166-1>.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Correspondence and requests for materials** should be addressed to F.F.

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.



# Genome-Wide Association Mapping of Flowering and Ripening Periods in Apple

## OPEN ACCESS

### Edited by:

Marion S. Röder,  
Leibniz-Institut für Pflanzengenetik und  
Kulturpflanzenforschung (IPK),  
Germany

### Reviewed by:

Ahmad M. Alqudah,  
Leibniz-Institut für Pflanzengenetik und  
Kulturpflanzenforschung (IPK),  
Germany

Robert Henry,

The University of Queensland,  
Australia

Raquel Sánchez-Pérez,

University of Copenhagen, Denmark

### \*Correspondence:

Jorge Urrestarazu  
jorge.urrestarazu@unavarra.es  
Charles-Eric Durel  
charles-eric.durel@inra.fr

† These authors have contributed  
equally to this work.

### Specialty section:

This article was submitted to  
Plant Genetics and Genomics,  
a section of the journal  
Frontiers in Plant Science

Received: 31 July 2017

Accepted: 24 October 2017

Published: 10 November 2017

### Citation:

Urrestarazu J, Muranty H, Denancé C,  
Leforestier D, Ravon E, Guyader A,  
Guisnel R, Feugey L, Aubourg S,  
Celton J-M, Daccord N, Dondini L,  
Gregori R, Lateur M, Houben P,  
Ordidge M, Paprstein F, Sedlak J,  
Nybom H, Garkava-Gustavsson L,  
Troggio M, Bianco L, Velasco R,  
Poncet C, Théron A, Moriya S,  
Bink MCAM, Laurens F, Tartarini S and  
Durel C-E (2017) Genome-Wide  
Association Mapping of Flowering and  
Ripening Periods in Apple.  
Front. Plant Sci. 8:1923.  
doi: 10.3389/fpls.2017.01923

Jorge Urrestarazu<sup>1,2,3\*†</sup>, Hélène Muranty<sup>1†</sup>, Caroline Denancé<sup>1</sup>, Diane Leforestier<sup>1</sup>,  
Elisa Ravon<sup>1</sup>, Arnaud Guyader<sup>1</sup>, Rémi Guisnel<sup>1</sup>, Laurence Feugey<sup>1</sup>, Sébastien Aubourg<sup>1</sup>,  
Jean-Marc Celton<sup>1</sup>, Nicolas Daccord<sup>1</sup>, Luca Dondini<sup>2</sup>, Roberto Gregori<sup>2</sup>, Marc Lateur<sup>4</sup>,  
Patrick Houben<sup>4</sup>, Matthew Ordidge<sup>5</sup>, Frantisek Paprstein<sup>6</sup>, Jiri Sedlak<sup>6</sup>, Hilde Nybom<sup>7</sup>,  
Larisa Garkava-Gustavsson<sup>8</sup>, Michela Troggio<sup>9</sup>, Luca Bianco<sup>9</sup>, Riccardo Velasco<sup>9</sup>,  
Charles Poncet<sup>10</sup>, Anthony Théron<sup>10</sup>, Shigeki Moriya<sup>1,11</sup>, Marco C. A. M. Bink<sup>12,13</sup>,  
François Laurens<sup>1</sup>, Stefano Tartarini<sup>2</sup> and Charles-Eric Durel<sup>1\*</sup>

<sup>1</sup> Institut de Recherche en Horticulture et Semences UMR 1345, INRA, SFR 4207 QUASAV, Beaucouzé, France,

<sup>2</sup> Department of Agricultural Sciences, University of Bologna, Bologna, Italy, <sup>3</sup> Department of Agricultural Sciences, Public

University of Navarre, Pamplona, Spain, <sup>4</sup> Plant Breeding and Biodiversity, Centre Wallon de Recherches Agronomiques,

Gembloux, Belgium, <sup>5</sup> School of Agriculture, Policy and Development, University of Reading, Reading, United Kingdom,

<sup>6</sup> Research and Breeding Institute of Pomology Holovousy Ltd., Horice, Czechia, <sup>7</sup> Department of Plant Breeding, Swedish

University of Agricultural Sciences, Kristianstad, Sweden, <sup>8</sup> Department of Plant Breeding, Swedish University of Agricultural

Sciences, Alnarp, Sweden, <sup>9</sup> Fondazione Edmund Mach, San Michele all'Adige, Italy, <sup>10</sup> Plateforme Gentyane, INRA, UMR

1095 Genetics, Diversity and Ecophysiology of Cereals, Clermont-Ferrand, France, <sup>11</sup> Apple Research Station, Institute of

Fruit Tree and Tea Science, National Agriculture and Food Research Organization (NARO), Morioka, Japan, <sup>12</sup> Wageningen

UR, Biometris, Wageningen, Netherlands, <sup>13</sup> Hendrix Genetics, Boxmeer, Netherlands

Deciphering the genetic control of flowering and ripening periods in apple is essential for breeding cultivars adapted to their growing environments. We implemented a large Genome-Wide Association Study (GWAS) at the European level using an association panel of 1,168 different apple genotypes distributed over six locations and phenotyped for these phenological traits. The panel was genotyped at a high-density of SNPs using the Axiom<sup>®</sup> Apple 480 K SNP array. We ran GWAS with a multi-locus mixed model (MLMM), which handles the putatively confounding effect of significant SNPs elsewhere on the genome. Genomic regions were further investigated to reveal candidate genes responsible for the phenotypic variation. At the whole population level, GWAS retained two SNPs as cofactors on chromosome 9 for flowering period, and six for ripening period (four on chromosome 3, one on chromosome 10 and one on chromosome 16) which, together accounted for 8.9 and 17.2% of the phenotypic variance, respectively. For both traits, SNPs in weak linkage disequilibrium were detected nearby, thus suggesting the existence of allelic heterogeneity. The geographic origins and relationships of apple cultivars accounted for large parts of the phenotypic variation. Variation in genotypic frequency of the SNPs associated with the two traits was connected to the geographic origin of the genotypes (grouped as North+East, West and South Europe), and indicated differential selection in different growing environments. Genes encoding transcription factors containing either NAC or MADS domains were identified as major candidates within the small confidence intervals computed for the associated genomic regions. A

strong microsynteny between apple and peach was revealed in all the four confidence interval regions. This study shows how association genetics can unravel the genetic control of important horticultural traits in apple, as well as reduce the confidence intervals of the associated regions identified by linkage mapping approaches. Our findings can be used for the improvement of apple through marker-assisted breeding strategies that take advantage of the accumulating additive effects of the identified SNPs.

**Keywords:** adaptive traits, association genetics, germplasm collection, GWAS, *Malus × domestica* Borkh., microsynteny, quantitative trait loci, SNP

## INTRODUCTION

Flowering time in temperate plants is influenced by multiple environmental factors related to temperature and day length at different periods of the year (Wilczek et al., 2009; Cook et al., 2012; Abbott et al., 2015). For crop cultivation, floral timing is of utmost importance, because it is a major yield determinant (Jung et al., 2017). Temperate fruit trees use bud dormancy for adaptation to seasonality (Campoy et al., 2011; Sánchez-Pérez et al., 2014; Ionescu et al., 2017): flowering occurs uniformly when the chilling and heating requirements associated with winter and spring have been fulfilled. In the context of global climate change, increasing temperatures tend to result in an acceleration of springtime phenological events (Hänninen and Tanino, 2011; Cook et al., 2012), with implications for both the risk of frost damage (Cannell and Smith, 1986; Vitasse et al., 2014) and the photosynthetic capacity of the trees (Ensminger et al., 2008). Moreover, this advance is responsible for several morphological disorders/abnormalities, including bud burst delay, low burst rate, irregular floral or leaf budbreak and poor fruit set (Erez, 2000; Celton et al., 2011; Dirlwanger et al., 2012; Abbott et al., 2015). Disruptions in synchronization of flowering may disturb pollination for self-incompatible cultivars, while modifications of fruit harvesting periods can cause problems with orchard management and fruit marketing (Dirlwanger et al., 2012). Breeding programs mainly focus on improvement of yield and fruit quality, but additional objectives like climate change adaptation receive increased attention. Genetic control of flowering and ripening periods plays a crucial role, since adaptation to different growing environments affects fruit quality (Chagné et al., 2014; Jung et al., 2017).

Flowering time is regulated by an intricate signaling network of multiple genes that integrates both endogenous and exogenous stimuli to induce flowering under the most favorable conditions

(Boss et al., 2004; Amasino, 2005). Fruit ripening control involves coordinated regulation of many metabolic pathways (Johnston et al., 2009; Pirona et al., 2013; Chagné et al., 2014), resulting in the conversion of starch to sugars, reduced acidity, reduced flesh firmness, changes in color and an increase in aroma/flavor volatile compounds. Both traits are quantitatively inherited in most fruit tree species (Celton et al., 2011; Pirona et al., 2013; Castède et al., 2014; Chagné et al., 2014).

Association mapping exploits the linkage disequilibrium (LD) present among individuals from natural populations or germplasm collections to dissect the genetic basis of complex trait variation (Neale and Savolainen, 2004; Aranzana et al., 2005; Balding, 2006; Myles et al., 2009). Germplasm collections generally contain more genetic diversity than segregating progenies and, since association mapping exploits all the recombination events that have occurred in the evolutionary history of the association panel, a much higher mapping resolution is expected (Zhu et al., 2008; Myles et al., 2009; Ingvarsson and Street, 2011). In addition, the number of QTLs that can be mapped for a given phenotype is not limited to the segregation products in a specific cross, but rather by the number of QTLs underlying the trait and the degree to which the studied population captures the genetic species-wide diversity (Zhu et al., 2008; Yano et al., 2016). Association mapping has recently been applied to fruit tree species such as peach (Micheletti et al., 2015), apricot (Mariette et al., 2016) and apple (Leforestier et al., 2015; Migicovsky et al., 2016; Di Guardo et al., 2017; Farneti et al., 2017), especially after the release of high-density SNP arrays with uniform coverage of the whole genome (Chagné et al., 2012; Verde et al., 2012; Bianco et al., 2014) or Genotyping-by-Sequencing (Gardner et al., 2014).

The high density Axiom® Apple 480 K SNP array (Bianco et al., 2016) developed within the EU-FruitBreedomics project (Laurens et al., 2012; <http://www.fruitbreedomics.com>) was used for the first time in the present study to perform GWAS. Here, we focused on the analysis of the genetic control of flowering and ripening periods in a panel of almost 1,200 different genotypes distributed over six apple collections managed by six European institutes. We identified one genomic region associated with flowering period and three with ripening period. Co-variation between the genotypic frequencies at the significant SNPs and three major geographic groupings of genotypes was explored, and candidate genes were identified in the detected genomic

**Abbreviations:** CRA-W, Centre Wallon de Recherche Agronomique [Gembloux (Belgium)]; EBIC, Extended Bayesian Information Criterion; GDDH13 genome, version (v1.1) released for the apple genome based on the doubled haploid GDDH13; GWAS, Genome-Wide Association Study; INRA, Institut National de la Recherche Agronomique [Angers (France)]; LD, linkage disequilibrium; MLM, multi-locus mixed model; NFC, University of Reading [Brogdale (United Kingdom)]; PCA, Principal Component Analysis; QTL, Quantitative Trait Loci; RBIPH, Research and Breeding Institute of Pomology Holovousy [Holovousy (Czech Republic)]; SLU, Swedish University of Agricultural Sciences [Alnarp (Sweden)]; SNP, Single Nucleotide Polymorphism; UNIBO, University of Bologna [Bologna (Italy)].

regions. To our knowledge, this is the largest association study ever performed in a fruit tree species considering both population size and SNP marker density.

## MATERIALS AND METHODS

### Plant Material

The association panel consisted of 1,168 different diploid apple genotypes corresponding to accessions preserved in six European germplasm collections (Table S1). The uniqueness of these genotypes was confirmed with SSR markers in a previous analysis (Urrestarazu et al., 2016). Some accessions corresponding to genotypes present at multiple locations were maintained in order to adjust phenotypic data between collections. Especially, ten standard genotypes (“Alkmene,” “Ananas Reinette,” “Discovery,” “Golden Delicious,” “Ingrid Marie,” “James Grieve,” “Jonathan,” “Reine des Reinettes” (= “King of the Pippins”), “Reinette de Champagne” and “Winter Banana”) were included from almost all collections. The association panel comprised mainly genotypes corresponding to old local/national cultivars, and the majority could be classified into three geographic groups according to their area of origin in Europe [North+East (141 different genotypes), South (148) and West (775)]; the remaining 104 corresponded to recent cultivars, germplasm originating from other worldwide regions, or were of unknown origin (Urrestarazu et al., 2016).

### Phenotypic Data Analysis

Phenotypic data for flowering and ripening periods were scored on an ordinal scale from 1 to 9, and consisted of both historical data in germplasm databases and of new data acquired in recent years (2012–2014) using the same scoring scales. Flowering period was assessed by recording dates of Fleckinger phenological flower stages F or F2 (Fleckinger, 1964), and then assigning a score on the ordinal scale by comparison to reference cultivars. Assessments for flowering period were performed over a period of 3–19 years except for NFC where only a single average value (assessed over 10 years) was available (Table 1). Ripening period was determined by observing pre-ripening drop of healthy fruits, ground- and over-color of fruits, taste of fruits and/or iodine starch index. It was recorded over 3–13 years (Table 1).

Genotypic means obtained for each genotype by adjusting for year and site effects were used as phenotypes for association analysis. When analyzing individual collections, the genotypic means were estimated using a linear model taking into account the year effect (Equation 1), while we considered the combined effect of site and year (Equation 2) for the whole analysis, i.e., all the collections were combined into a single analysis:

$$P_{ik} = \mu + Y_i + g_k + e_{ik} \quad (1)$$

$$P_{ijk} = \mu + (Y_i \times S_j) + g_k + e_{ijk} \quad (2)$$

where for (Equation 1),  $P_{ik}$  is the phenotypic value of the  $k$ th genotype in the  $i$ th year;  $\mu$  is the mean value of the trait;  $Y_i$  is the fixed effect of the  $i$ th year on the trait;  $g_k$  is the random genotypic effect of genotype  $k$ ; and  $e_{ik}$  is the residual term of the

model. For (Equation 2),  $\mu$ ,  $Y_i$ , and  $g_k$  have the same meanings as in (Equation 1);  $P_{ijk}$  refers to the phenotypic value of the  $k$ th genotype in the  $i$ th year in the  $j$ th site;  $S_j$  is the fixed effect of the  $j$ th site; and  $e_{ijk}$  is the residual term of the model. Heritability of genotypic means ( $h^2$ , here called broad-sense heritability) was estimated for each individual collection as:

$$h^2 = \frac{\sigma_G^2}{\sigma_G^2 + \frac{\sigma_\epsilon^2}{n}} \quad (3)$$

where  $\sigma_G^2$  is the variance of genotype effect,  $\sigma_\epsilon^2$  is the variance of the residual term, and  $n$  is the mean number of observations per genotype. These analyses were performed using “R” software (R Core Team, 2014), in particular the packages effects (Fox, 2003), lme4 (Bates and Sarkar, 2007) and FactoMineR (Lê et al., 2008).

### SNP Genotyping

The 1,168 apple genotypes were genotyped with the Axiom® Apple 480 K array containing 487,249 SNPs evenly distributed over the 17 apple chromosomes (Bianco et al., 2016). Bianco et al. (2016) applied stringent filters that resulted in a set of 275,223 robust SNPs for GWAS. Further details on the development of the SNP array, genotyping process, and the filtering pipeline procedure can be found in Bianco et al. (2016). All presented results use the SNP positions on the latest version (v1.1) released for the apple genome based on the doubled haploid GDDH13 (hereafter, GDDH13 genome; Daccord et al., 2017; see also <https://iris.angers.inra.fr/gddh13/> for the genome browser).

### Kinship, Population Structure and Linkage Disequilibrium Estimates

GEMMA software (Zhou and Stephens, 2012) was used to estimate the standardized relatedness matrix (**K**) between the genotypes at the whole population level and within each collection. A Principal Component Analysis (PCA) of the SNP data was performed using PLINK (Purcell et al., 2007) and the ten largest Eigenvalues were used to control for population structure (**Q**). Matrix **Q** was constructed for the whole population as well as for each collection separately.

LD was studied between sets of SNPs spanning regions of 10 kb randomly sampled along the genome. These sets were obtained by a random choice of 50 contigs larger than 10 kb on each chromosome, followed by the selection of SNPs spanning a random region of 10 kb in each of these contigs. LD was estimated as squared allele frequency correlations ( $r^2$ ) and as  $r^2$  corrected for population structure and relatedness ( $r_{vs}^2$ ) using the R-package LDcorSV (Mangin et al., 2012). In addition, local LD ( $r^2$ ) was assessed for chromosomal regions of 1 Mb surrounding the SNPs retained as cofactors in the GWAS (see next section for details) and displayed in LD maps and network plots using “LDheatmap” and “network” R-packages, respectively.

### Genome-Wide Association Study (GWAS)

The GWAS method was applied both at the whole population level and for each collection independently. GWAS were conducted with correction for population structure (**Q**) and



**TABLE 1** | Averages and ranges for the genotypic means for flowering and ripening periods.

| Population              | Phenotypic assessments |           |                   | Genotypic adjusted means |      |           |
|-------------------------|------------------------|-----------|-------------------|--------------------------|------|-----------|
|                         | No. years (range)      | data/cvr. | bs_h <sup>2</sup> | Mean                     | SD   | Range     |
| <b>FLOWERING PERIOD</b> |                        |           |                   |                          |      |           |
| Whole population        | 29 (1985–2014)         | 5.45      | 0.82              | 4.79                     | 1.15 | 1.73–9.24 |
| INRA                    | 4 (2009–2012)          | 3.00      | 0.88              | 5.58                     | 1.45 | 2.56–9.52 |
| UNIBO                   | 6 (1987–1992)          | 7.60      | 0.84              | 5.26                     | 0.99 | 2.57–7.81 |
| CRA-W                   | 19 (1985–2007)         | 4.90      | 0.88              | 3.97                     | 1.03 | 1.25–7.25 |
| RBIPH                   | 13 (1995–2010)         | 5.00      | 0.85              | 4.42                     | 0.83 | 3.03–8.61 |
| NFC                     | 1 <sup>a</sup>         | –         | –                 | 4.91                     | 0.92 | 2.00–9.00 |
| SLU                     | 3 (2012–2014)          | 3.00      | 0.81              | 3.66                     | 0.85 | 1.99–6.56 |
| <b>RIPENING PERIOD</b>  |                        |           |                   |                          |      |           |
| Whole population        | 22 (1987–2014)         | 5.37      | 0.95              | 5.43                     | 2.05 | 0.54–9.95 |
| INRA                    | 10 (2002–2014)         | 4.86      | 0.95              | 6.89                     | 1.77 | 1.62–9.26 |
| UNIBO                   | 13 (1987–2014)         | 7.83      | 0.96              | 6.51                     | 1.87 | 0.98–9.19 |
| CRA-W                   | 10 (1987–2008)         | 4.37      | 0.87              | 4.90                     | 1.15 | 1.12–8.38 |
| RBIPH                   | 5 (2006–2010)          | 5.07      | 0.92              | 5.00                     | 1.56 | 1.00–7.60 |
| NFC                     | 3 (1999–2013)          | 2.93      | 0.87              | 5.94                     | 1.77 | 2.00–8.33 |
| SLU                     | 3 (2012–2014)          | 2.91      | 0.98              | 3.75                     | 1.44 | 1.00–7.00 |

<sup>a</sup>A single average value was available at NFC, assessed over 10 years (different years according to the cultivars).

modeling phenotypic covariance with the kinship matrix (**K**) implemented in a modified version of the multi-locus mixed model (MLMM) proposed by Segura et al. (2012). The Extended Bayesian Information Criterion (EBIC, Chen and Chen, 2008) was used to select the model that best fitted our data. A genome-wide significance threshold was determined using a Bonferroni correction at 5%. MLMM uses a stepwise mixed-model regression with forward inclusion and backward elimination of SNPs re-estimating the variance components of the model at each step. MLMM divides the phenotypic variance into genetic variance (explained by structure, by kinship, and by SNPs included as cofactors in the model), and unexplained variance (residual variance), suggesting a natural stopping criterion (genetic variance = 0) for including cofactors, and allowing to estimate the explained and unexplained heritable variance for each trait. The causal-variant heritability tagged by all possible genotyped SNPs, was quantified for each trait at the step 0 of MLMM, i.e., when the structure, kinship and residual variances were estimated with no SNP included as cofactors in the model. The part of variance explained (PVE) by the significant SNP(s) as well as the part of variance due to population structure and due to kinship, were estimated at the optimal step of the MLMM (i.e., stopping criterion).

To establish 95% confidence intervals for the significant SNPs retained as cofactors in the whole population, we conducted a re-sampling approach as proposed by Hayes (2013). The full set of individuals with phenotypic data was randomly split into two subsets with equal size; this procedure was repeated 50 times for each trait, and then a GWAS was run on each subset as explained above. The standard error ( $se(x)$ ) of the position of an underlying association was estimated as the median absolute deviation of the positions of the SNPs retained as cofactors on

each chromosome over all subsets. Then, the 95% confidence interval was calculated as the position of the most significant SNPs retained as cofactors in the analysis of the whole population  $\pm 1.96 se(x)$ .

### Effects of the SNPs Identified as Cofactors

The SNPs identified as cofactors were analyzed toward mode and size of allelic effects. The mode of gene action at each SNP was estimated for the whole population and (when possible) for each of the three geographic groups using the ratio of dominance ( $d$ ) to additive ( $a$ ) effects calculated from the mean of the genotypic means for each genotypic class. The dominance effect was calculated as the difference between the mean observed within the heterozygous class and the mean across both homozygous classes ( $d = G_{AB} - 0.5(G_{AA} + G_{BB})$ ), where  $G_{ij}$  is the trait mean in the  $ij$ th genotypic class). To classify the mode, we used the following ranges, similar to Wegrzyn et al. (2010). No dominance was defined for small absolute values, i.e.,  $|d/a| \leq 0.50$ ; partial or complete dominance was defined as values in the range  $0.50 < |d/a| < 1.25$ ; and over- or under-dominance pertained to values of  $|d/a| > 1.25$ .

To assess the joint effect of the different allelic combinations (i.e., genetic variants) defined by the SNPs identified for each trait, mean and statistical significance among the most frequent genetic variants were calculated by the Tukey-Kramer test ( $\alpha = 0.05$ ).

Dominance and epistatic effects among the identified SNPs were tested in a model including their additive effects with correction for population structure (**Q**) and modeling phenotypic covariance with the kinship matrix (**K**). Percentages of variance explained by additive plus dominance effects, and by additive



plus dominance and epistatic effects were estimated in a hierarchical sequence, using a cumulative  $R^2$  metric.

### In Silico Candidate Gene Research

Chromosomal regions corresponding to approximate 95% confidence intervals for the position of SNPs retained as cofactors in the whole population for flowering and ripening periods were investigated for *in silico* candidate gene identification using GDDH13 genome (Daccord et al., 2017). The annotations of protein-coding and non-protein-coding genes of the regions of interest were identified using GDDH13 genome v1.1 browser (<https://iris.angers.inra.fr/gddh13/>). Annotations regarding the biochemical function of genes (mainly provided by InterproScan) were enriched by the biological functions inferred from the putative orthologs identified in *Arabidopsis thaliana*, *Solanum lycopersicum*, and *Prunus persica* genomes. Furthermore, structures of predicted genes and intergenic regions were systematically investigated to detect eventual mis/not-annotated genes and pseudogenes (stop codons and/or frameshifts in their CDS) in the regions of interest.

## RESULTS

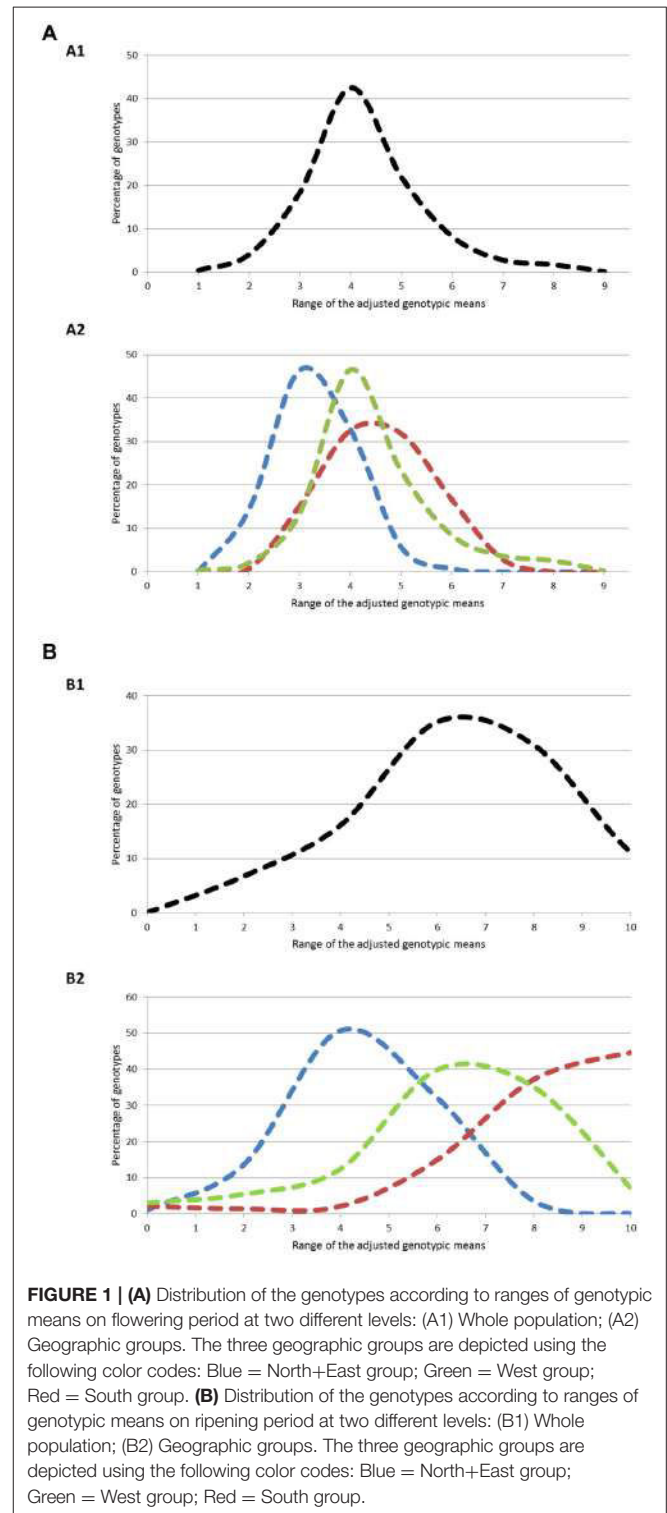
### Phenotypic Variation

Large phenotypic variation was observed at both whole population and collection level (Table 1). In the individual collections, average flowering period varied in genotypic means from 3.66 to 5.58 (on the 1–9 scale), while average ripening period varied from 3.75 to 6.89. Heritability was consistently high ( $>0.80$ ). The two traits were significantly correlated when calculated across all genotypes in the whole population ( $r = 0.44$ ;  $p\text{-value} = 2.2e^{-16}$ ), see Figure S1.

The three geographical groups differed considerably for both traits (Figure 1). For flowering period, 94% of the genotypes had genotypic means between 2 and 5 (mean value = 3.77) in the North+East group, while 96 and 83% of the genotypes varied between 3 and 7 and between 3 and 6 for the South and West groups respectively, with almost identical mean values (South: 5.03; West: 4.99). For ripening period, phenotypic variation was even higher: 83% of the genotypes in the North+East group had a genotypic mean below 5 (mean value = 3.41), while 91% in the South group had values above 5 (mean value = 7.49). The West group showed an intermediate distribution (mean value = 5.48).

### Population Structure and Linkage Disequilibrium

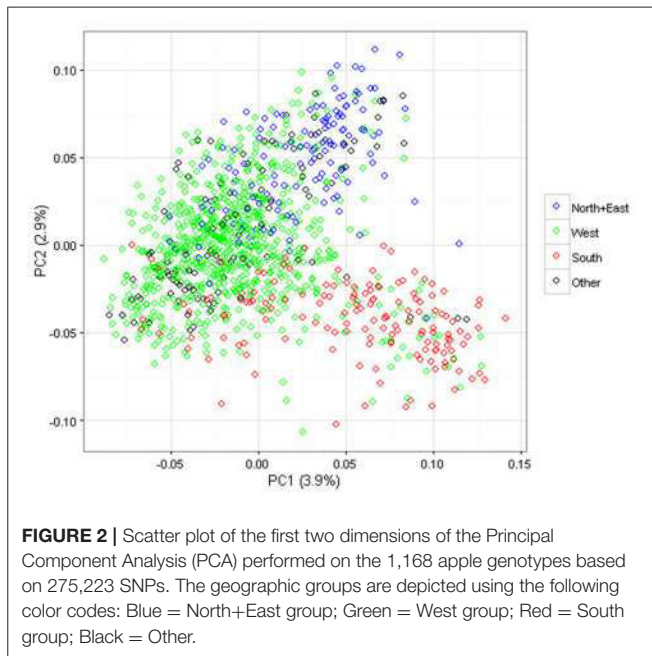
PCA was applied to summarize global genetic marker variation in the association panel: the ten largest Eigenvalues used to describe the whole population structure explained 17% of the overall variation. In Figure 2, the first two components of the PCA are represented. Genetic discrimination between the genotypes classified according to their geographic group of origin is visible in the bi-dimensional plot; genotypes from the North+East group are located in the upper part along the Y axis, while those from the West and South groups mostly occur on the left and right side along the X axis, respectively. The three groups North+East, West and South, explained 30 and 37% of the



**FIGURE 1 | (A)** Distribution of the genotypes according to ranges of genotypic means on flowering period at two different levels: (A1) Whole population; (A2) Geographic groups. The three geographical groups are depicted using the following color codes: Blue = North+East group; Green = West group; Red = South group. **(B)** Distribution of the genotypes according to ranges of genotypic means on ripening period at two different levels: (B1) Whole population; (B2) Geographic groups. The three geographical groups are depicted using the following color codes: Blue = North+East group; Green = West group; Red = South group.

variation for the first two dimensions of the PCA but less than 6.5% for the next eight dimensions.

LD ( $r^2$ ) was very variable in the SNP sets of 10 kb randomly sampled along the genome, spanning the entire range from absence to complete LD (Figure S2). The distribution of LD



was highly asymmetric, with half of the marker pairs showing a  $r^2$  value below 0.1 ( $r_{vs}^2$  value below 0.07 when corrected for relatedness and population structure). LD decay curves were very flat (**Figure 3**); mean  $r^2$  values were 0.24, 0.21, and 0.19 at 100 bp, 1 kb, and 5 kb, respectively, while mean  $r_{vs}^2$  values were 0.20, 0.17, and 0.13, respectively. Half of the adjacent marker pairs occurred within 587 bp, while 90% occurred within 4,975 bp. To estimate LD between a causal variant in the middle of a marker interval and its flanking markers, mean  $r^2$  values for marker pairs at half these distances (i.e., 293.5 and 2,487.5 bp) were computed: in the whole population, mean  $r^2$  values were 0.23 and 0.19 without correction, and 0.19 and 0.14 with correction.

## Genome-Wide Association Study (GWAS) Without Cofactor Inclusion

### Flowering Period

Using a single-locus mixed model with control for population structure and relatedness, 50 SNPs were significantly associated with flowering period for the whole population (Table S2). In a quantile-quantile (Q-Q) plot (not shown), close adherence was found between the observed and expected  $-\log_{10}(p)$  values till around 3, indicating that the significant SNPs are unlikely to be biased by population structure and relatedness. A strong association signal was found on chromosome 9 (49 SNPs) with a Bonferroni correction threshold of 5% ( $-\log_{10}(p) > 6.74$ ). The remaining SNP was located on the fictive chromosome 0 containing all unassigned scaffolds. The SNPs located on chromosome 9 spanned a distance of 3.24 Mb (265,164–3,509,888 bp).

Analyses of the individual collections revealed significant associations for flowering period only at INRA (29 SNPs), NFC (2 SNPs) and RBIPH (1 SNP) (Table S2). Twenty-one of these SNPs, all on chromosome 9, were also significant

for the whole population. Of the 11 SNPs identified only in individual collections, nine were located on chromosome 9, one on chromosome 4 (RBIPH), and one on chromosome 11 (NFC).

### Ripening Period

For ripening period, 82 SNPs exhibited a significant association for the whole population with adjustment for population structure and relatedness (Table S3). The Q-Q plot (not shown) was similar to the previous one. Most SNPs (70) were located on chromosome 3, spanning a distance of 2.05 Mb (29,196,200–31,243,065 bp). Nine SNPs were located on chromosome 16 and spanned a distance of 274.3 kb (9,032,064–9,306,332 bp), while three SNPs could not be mapped.

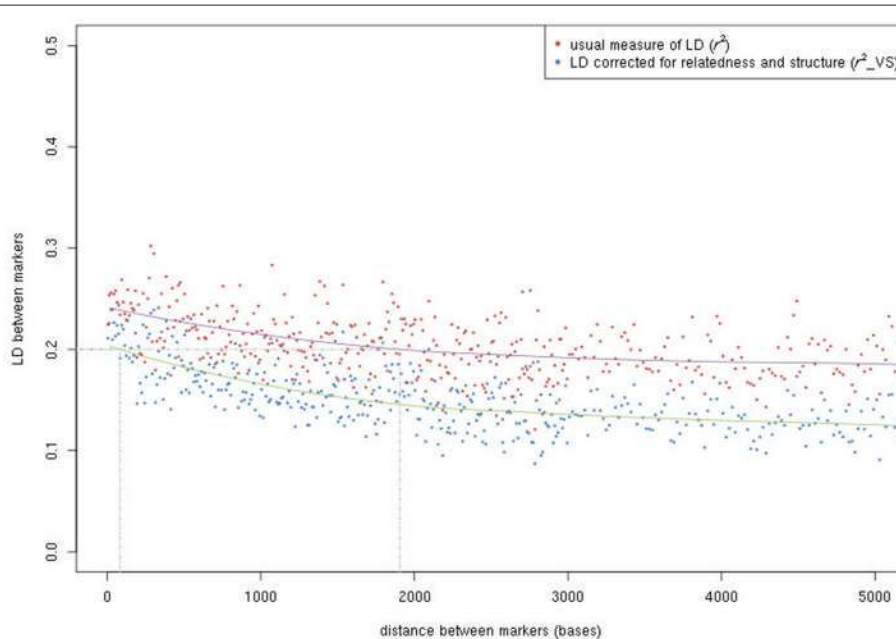
When GWAS was performed for each collection separately, numbers of significant SNPs were 38, 12, and 8 for NFC, INRA and SLU, respectively, two for both RBIPH and UNIBO, and only one for CRA-W (Table S3). Thirty-one of the 43 SNPs identified in the analyses of individual collections, showed a significant association also in the whole population. When analyses were carried out at collection-scale, all the identified SNPs were located on chromosome 3, except for three that were unmapped; none of the SNPs located on chromosome 16 with a significant association in the whole population, were identified in the GWAS of the individual collections.

## Genome-Wide Association Study (GWAS) Using SNPs as Cofactors

To further dissect the signal from the chromosomal regions containing the sets of significant SNPs for each trait, we performed a GWAS with MLM using SNPs as cofactors (Segura et al., 2012). MLM handles the putatively confounding effect of significant SNPs elsewhere on the genome, which considerably outperforms the existing single-locus mixed models by reducing the number of significantly associated SNPs rather than the number of peaks (Sauvage et al., 2014). The part of variance explained by structure and kinship estimated at step 0 of MLM (i.e., when no SNP were included as cofactors in the model) was 0.78 for flowering period and 0.84 for ripening period (Table 2).

### Flowering Period

The optimal MLM according to the EBIC criterion for the whole population retained two SNPs for flowering (FB\_AFFY\_0496090 “SNP.9-1” and FB\_AFFY\_0495650 “SNP.9-2”), both located on chromosome 9, only 27 kb apart (Table 3). These two SNPs were significantly associated with flowering period also in the initial analysis based on a single-locus mixed model (Table S2). With this optimal model, 8.9% of the whole phenotypic variance was explained by the pair of SNPs retained, 27.3% corresponded to the underlying population structure of the association panel, and 38.6% was associated with kinship (**Figure 4A**). In the re-sampling analysis conducted for estimating an approximate 95% confidence interval, the number of cofactors retained in the model was one in 76 subsets, two in 23 subsets and three in one subset (Table S4). SNP.9-1 and SNP.9-2 were selected as cofactors in 35 and 38 subsets, respectively, while other SNPs from the same chromosome were selected as cofactors in 45 subsets, among which FB\_AFFY\_4941692



**FIGURE 3** | LD decay according to the physical distance between SNPs. Both the usual  $r^2$  and the  $r^2$  after correcting for relatedness and population structure ( $r^2_{VS}$ ) are given.

**TABLE 2** | Summary of trait associations at the optimal models according to the EBIC criterion.

| Population              | No. cultivars | Model 0 PVE by<br>(structure + kinship) | Optimum model                               |                                       |                              |                     |
|-------------------------|---------------|---|---|---------------------------------------|------------------------------|---------------------|
|                         |               |   | PVE by (structure +<br>cofactors + kinship) | No. associations<br>without cofactors | No. significant<br>cofactors | PVE by<br>cofactors |
| <b>FLOWERING PERIOD</b> |               |   |   |                                       |                              |                     |
| Whole population        | 1,126         | 0.78                                    | 0.75  | 50                                    | 2                            | 0.09                |
| INRA                    | 251           | 0.93                                    | 0.90  | 29                                    | 1                            | 0.13                |
| UNIBO                   | 166           | 0.74                                    | 0.74  | 0                                     | 0                            | 0.00                |
| CRA-W                   | 221           | 0.72                                    | 0.72  | 0                                     | 0                            | 0.00                |
| RBIPH                   | 177           | 0.79                                    | 0.58  | 1                                     | 2                            | 0.27                |
| NFC                     | 288           | 0.78                                    | 0.77  | 2                                     | 4                            | 0.33                |
| SLU                     | 159           | 0.80                                    | 0.80  | 0                                     | 0                            | 0.00                |
| <b>RIPENING PERIOD</b>  |               |   |   |                                       |                              |                     |
| Whole population        | 1,149         | 0.84                                    | 0.85  | 82                                    | 6                            | 0.17                |
| INRA                    | 260           | 0.84                                    | 0.84  | 12                                    | 1                            | 0.13                |
| UNIBO                   | 178           | 0.88                                    | 0.86  | 2                                     | 1                            | 0.16                |
| CRA-W                   | 217           | 0.70                                    | 0.65  | 1                                     | 1                            | 0.12                |
| RBIPH                   | 176           | 0.80                                    | 0.78  | 2                                     | 2                            | 0.18                |
| NFC                     | 293           | 0.97                                    | 0.92  | 38                                    | 1                            | 0.22                |
| SLU                     | 160           | 0.94                                    | 0.89  | 8                                     | 4                            | 0.28                |

Part of Variance Explained (PVE) by population structure, cofactors and kinship, the number of associations without cofactors, the number of significant cofactors, the PVE by cofactors, and the ratio PVE by kinship/PVE by cofactors and kinship are showed. Data obtained for individual collections and the whole populations are provided.

(“SNP.9-5”) was selected in 25 subsets (Table S4). For this region, the length of a 95% confidence interval was estimated at 157 kb.

GWAS of each single collection retained one, two and four SNPs (Tables 2, 3) for the collections of INRA, RBIPH, and NFC,

respectively, but none for the collections of CRA-W, SLU, and UNIBO. The part of variance explained by the markers selected in the optimal models for each collection was 13% (INRA), 27% (RBIPH), and 33% (NFC) (Table 2; Figure 4A). One of the two SNPs identified in the MLM analysis of the whole population,

**TABLE 3** | Summary of associations identified by Multi-Locus Mixed Model (MLMM) at the optimal models according to the EBIC criterion for flowering and ripening periods in the whole population and in the six individual collections.

| Population              | SNP code        | SNP short name | No. Cofactor <sup>a</sup> | Location of the SNPs |            | Alleles                 | p-value  | MAF               |
|-------------------------|-----------------|----------------|---------------------------|----------------------|------------|-------------------------|----------|-------------------|
|                         |                 |                |                           | Chromosome           | Position   |                         |          |                   |
| <b>FLOWERING PERIOD</b> |                 |                |                           |                      |            |                         |          |                   |
| Whole population        | FB_AFFY_0496090 | SNP.9-1        | 1                         | 9                    | 530,386    | <b>G/T</b> <sup>b</sup> | 1.33E-08 | 0.11              |
| Whole population        | FB_AFFY_0495650 | SNP.9-2        | 2                         | 9                    | 557,419    | <b>A/G</b>              | 6.81E-08 | 0.13              |
| INRA                    | FB_AFFY_0495650 | SNP.9-2        | 1                         | 9                    | 557,419    | <b>A/G</b>              | 1.06E-12 | 0.18              |
| RBIPH                   | FB_AFFY_6830175 | SNP.4-1        | 1                         | 4                    | 968,334    | <b>C/T</b>              | 3.16E-09 | 0.01              |
| RBIPH                   | FB_AFFY_1629518 | SNP.9-3        | 2                         | 9                    | 925,476    | <b>A/G</b>              | 8.01E-08 | 0.14              |
| NFC                     | FB_AFFY_6873601 | SNP.4-2        | 4                         | 4                    | 7,719,622  | <b>A/G</b>              | 3.94E-07 | 0.24              |
| NFC                     | FB_AFFY_7355751 | SNP.9-4        | 2                         | 9                    | 1,938,744  | <b>C/T</b>              | 3.15E-08 | 0.11              |
| NFC                     | FB_AFFY_2782466 | SNP.11-1       | 1                         | 11                   | 12,422,656 | <b>A/C</b>              | 8.30E-09 | 0.02              |
| NFC                     | FB_AFFY_9818101 | SNP.12-1       | 3                         | 12                   | 14,536,815 | <b>C/T</b>              | 5.35E-08 | 0.15              |
| <b>RIPENING PERIOD</b>  |                 |                |                           |                      |            |                         |          |                   |
| Whole population        | FB_AFFY_6730867 | SNP.3-3        | 4                         | 3                    | 30,430,113 | <b>A/G</b> <sup>c</sup> | 6.76E-15 | 0.10              |
| Whole population        | FB_AFFY_7541229 | SNP.3-4        | 5                         | 3                    | 30,465,002 | <b>C/T</b>              | 8.51E-10 | 0.09              |
| Whole population        | FB_AFFY_4981462 | SNP.3-6        | 2                         | 3                    | 30,700,183 | <b>C/T</b>              | 4.39E-19 | 0.18              |
| Whole population        | FB_AFFY_1209620 | SNP.3-7        | 1                         | 3                    | 30,726,252 | <b>A/G</b>              | 1.28E-13 | 0.41              |
| Whole population        | FB_AFFY_3795860 | SNP.10-1       | 6                         | 10                   | 38,390,484 | <b>A/G</b>              | 1.76E-08 | 0.23              |
| Whole population        | FB_AFFY_6370928 | SNP.16-1       | 3                         | 16                   | 9,146,297  | <b>C/T</b>              | 5.16E-12 | 0.14              |
| INRA                    | FB_AFFY_1253936 | SNP.3-5        | 1                         | 3                    | 30,590,166 | <b>A/C</b>              | 3.03E-14 | 0.08              |
| UNIBO                   | FB_AFFY_1253936 | SNP.3-5        | 1                         | 3                    | 30,590,166 | <b>A/C</b>              | 6.75E-09 | 0.06              |
| CRA-W                   | FB_AFFY_4741632 | SNP.3-2        | 1                         | 3                    | 30,318,639 | <b>A/G</b>              | 6.71E-08 | 0.11              |
| RBIPH                   | FB_AFFY_4981462 | SNP.3-6        | 1                         | 3                    | 30,700,183 | <b>C/T</b>              | 3.60E-10 | 0.20              |
| RBIPH                   | FB_AFFY_4836781 | SNP.15-1       | 2                         | 15                   | 10,377,731 | <b>C/T</b>              | 3.18E-08 | 0.37              |
| NFC                     | FB_AFFY_4981462 | SNP.3-6        | 1                         | 3                    | 30,700,183 | <b>C/T</b>              | 1.37E-18 | 0.14              |
| SLU                     | FB_AFFY_0899559 | SNP.3-1        | 4                         | 3                    | 24,220,838 | <b>A/G</b>              | 7.21E-07 | 0.10              |
| SLU                     | FB_AFFY_1209620 | SNP.3-7        | 1                         | 3                    | 30,726,252 | <b>A/G</b>              | 7.51E-15 | 0.33 <sup>d</sup> |
| SLU                     | FB_AFFY_6239519 | SNP.13-1       | 3                         | 13                   | 1,889,560  | <b>G/T</b>              | 1.41E-07 | 0.24              |
| SLU                     | FB_AFFY_3879540 | SNP.16-2       | 2                         | 16                   | 10,298,660 | <b>G/T</b>              | 2.34E-07 | 0.27              |

<sup>a</sup>Order of inclusion of the SNPs at the optimal model in MLMM according to the EBIC criterion.

<sup>b</sup>The allele associated with an early flowering period is highlighted in bold. The alternative allele is thus associated with a late flowering period.

<sup>c</sup>The allele associated with an early ripening period is highlighted in bold. The alternative allele is thus associated with a late ripening period.

<sup>d</sup>The allele found in SLU with the lowest frequency was the opposite to the one that appeared in the lowest frequency in the whole population and the other five individual collections.

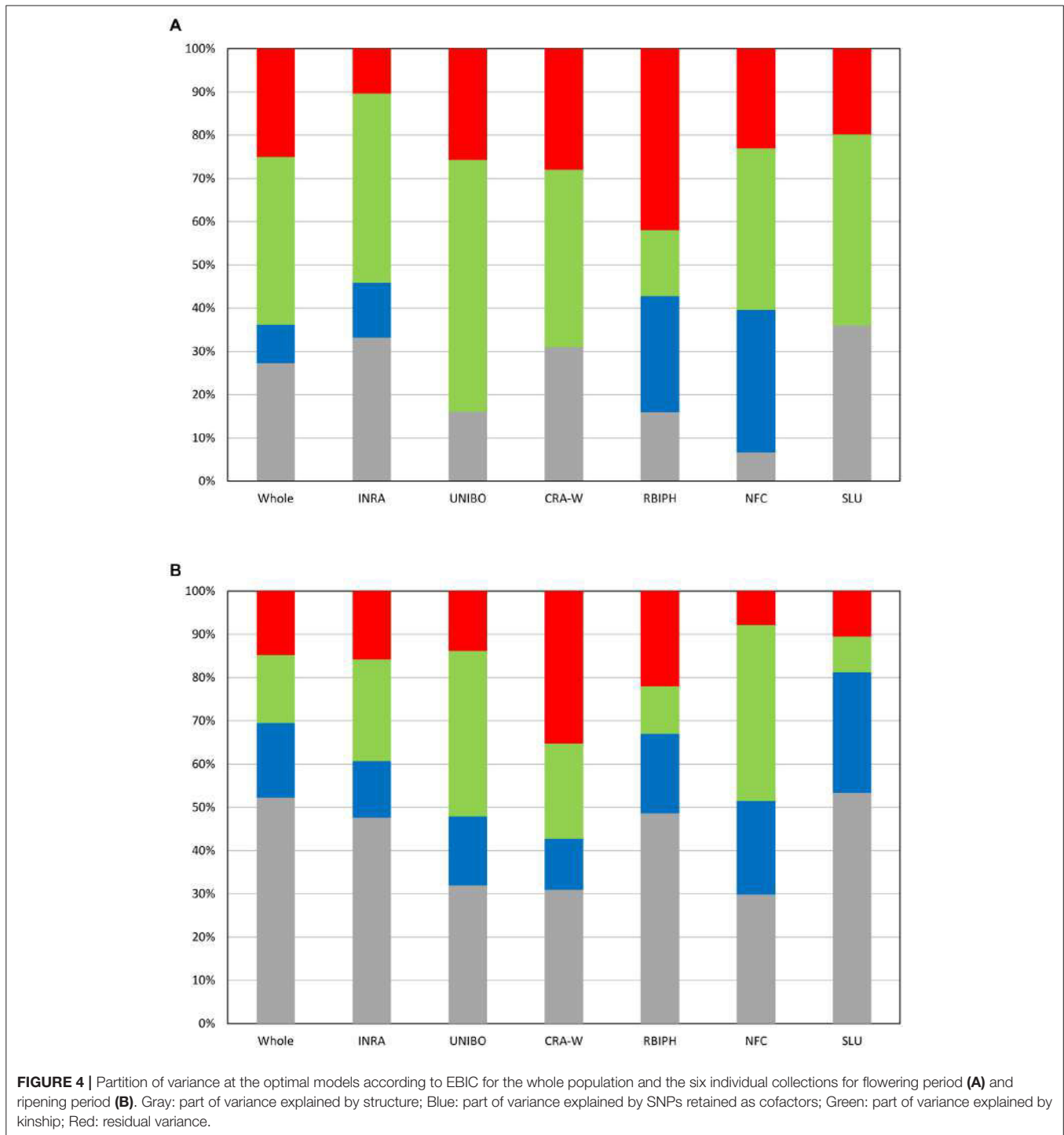
SNP.9-2, was also found for the INRA collection. Neither the two SNPs selected for the RBIPH collection (chromosomes 4 and 9) nor the four identified in the NFC collection (chromosomes 4, 9, 11, and 12) were retained in the analysis of the whole population.

### Ripening Period

The optimal model according to the EBIC criterion retained six SNPs for the whole population, four at the bottom of chromosome 3 (FB\_AFFY\_6730867, FB\_AFFY\_7541229, FB\_AFFY\_4981462, and FB\_AFFY\_1209620, denoted as “SNP.3-3,” “SNP.3-4,” “SNP.3-6,” and “SNP.3-7”), one at the bottom of chromosome 10 (FB\_AFFY\_3795860 “SNP.10-1”), and another on the top of chromosome 16 (FB\_AFFY\_6370928 “SNP.16-1”) (Table 3). The four SNPs identified on chromosome 3 were clustered two by two with a distance of only 35 and 26 kb within each cluster and a distance of about 296 kb between clusters. The SNP retained on chromosome 10 (SNP.10-1) and one of the four retained on chromosome 3 (SNP.3-3) did not show a significant

association with ripening period in the analysis based on a single-locus mixed model (Table S3). Altogether, the six SNPs explained 17.2% of the phenotypic variation, whereas population structure and kinship explained 52.2 and 15.7%, respectively (Figure 4B). When estimating the approximate 95% confidence interval with a re-sampling analysis, the number of cofactors retained varied between one and five, and between two and four in 94 of the 100 subsets (Table S5). SNP.3-6 and SNP.3-7, selected as the two first cofactors in the whole collection, were selected in 55 and 85 subsets, respectively. Other SNPs from chromosome 3 were selected in 59 subsets (Table S5). The length of a 95% confidence interval was estimated at only 152 kb for chromosome 3, but 1.39 Mb for chromosome 10, and 426 kb for chromosome 16.

The optimal model for the analysis of each individual collection retained at least one SNP per collection (Tables 2, 3): four in the SLU collection, two in the RBIPH collection, and one each in the remaining collections. Part of variance explained by the markers identified for each collection, ranged



from 12% (CRA-W) to 28% (SLU) (Table 2; Figure 4B), with an average of 18%. Two out of the six SNPs retained in the whole population were identified also in some of the individual collections, i.e., SNP.3-6 in NFC and RBIPH, and SNP.3-7 in SLU, both of them belonging to the same lower cluster previously defined on chromosome 3. For INRA and UNIBO the same single SNP (FB\_AFFY\_1253936 “SNP.3-5”) on chromosome 3 was selected by MLM and was located in between the two

previously identified SNP clusters. In brief, the analyses of the individual collections identified six SNPs additional to the six ones identified in the whole population, three on chromosome 3, and one on each of chromosomes 13, 15, and 16.

For each trait, Manhattan plots obtained with the single-locus mixed model and the multi-locus mixed model for the whole population and for the individual collections, are shown in Figures S3, S4.



## Linkage Disequilibrium among SNPs Identified as Cofactors

Pairwise LD was assessed to test the independence of SNPs identified as cofactors for each trait in the whole population as well as in each geographic group. For flowering period, low LD ( $r^2 = 0.27$ ;  $r_{vs}^2 = 0.12$ ) was detected in the whole population between the two SNPs associated with the trait despite being located only 27 kb apart (Table 4A). Analysis of the results at the geographic-group level found almost complete equilibrium between these two SNPs in the North+East and South groups ( $r^2 = 4E^{-04}$  and 0.06, respectively), while LD was much higher ( $r^2 = 0.41$ ;  $r_{vs}^2 = 0.22$ ) in the West group. For ripening period, variable LD values were found among the four SNPs identified as cofactors located on chromosome 3 (Table 4B). Intermediate to high  $r^2$  values were found between SNP.3-3, SNP.3-4, and SNP.3-6, whereas low values were observed between these three SNPs and SNP.3-7. In the North+East and West groups,  $r^2$

values between these four SNPs were quite similar to those found in the whole population. By contrast, very low  $r^2$  were found for the South group, except for the pair SNP.3-3/SNP.3-6 ( $r^2 = 0.43$ ).

LD interconnections between the eight SNPs retained as cofactors and other SNPs residing within their surrounding regions of 1 Mb (Figure S6) showed that SNP.16-1 exhibited the highest number of connections with other SNPs at  $r^2 > 0.70$ , i.e., 53 SNPs delineating a region of 763 kb (Table S6). None or only a few (maximum 16) SNPs in the neighborhood of the remaining seven retained SNPs were linked with them at  $r^2 > 0.70$ . Accordingly, none of the triangular LD heat maps for the above mentioned regions showed a LD spatial pattern suggesting that they are organized in blocks of moderate/high LD (Figure S5). Conversely, networks of moderate LD ( $r^2 \sim 0.5-0.6$ ) were more frequently observed for the above mentioned regions (i.e., five regions with more than 40 connected SNPs).

**TABLE 4 | (A)** Pairwise LD between the two SNPs associated with flowering period in the whole population (A1) and in the three geographic groups: North+East (A2), West (A3), and South (A4). **(B)** Pairwise LD between the four SNPs associated with ripening period on chromosome 3 in the whole population (B1) and the three geographic groups: North+East (B2), West (B3), and South (B4).

| A                            |         |         |      | B                            |         |         |         |                      |                          |
|------------------------------|---------|---------|------|------------------------------|---------|---------|---------|----------------------|--------------------------|
| <b>(A1) Whole population</b> |         |         |      | <b>(B1) Whole population</b> |         |         |         |                      |                          |
| SNPs as cofactors            | SNP.9-1 | SNP.9-2 | MAF  | SNPs as cofactors            | SNP.3-3 | SNP.3-4 | SNP.3-6 | SNP.3-7              | MAF                      |
| SNP.9-1                      | 1.00    | 0.12    | 0.11 | SNP.3-3                      | 1.00    | 0.55    | 0.31    | 0.02                 | 0.10                     |
| SNP.9-2                      | 0.27    | 1.00    | 0.13 | SNP.3-4                      | 0.71    | 1.00    | 0.27    | 0.06                 | 0.09                     |
|                              |         |         |      | SNP.3-6                      | 0.56    | 0.54    | 1.00    | 0.22                 | 0.18                     |
|                              |         |         |      | SNP.3-7                      | 0.11    | 0.06    | 0.22    | 1.00                 | 0.41                     |
| <b>(A2) North+East group</b> |         |         |      | <b>(B2) North+East group</b> |         |         |         |                      |                          |
| SNPs as cofactors            | SNP.9-1 | SNP.9-2 | MAF  | SNPs as cofactors            | SNP.3-3 | SNP.3-4 | SNP.3-6 | SNP.3-7 <sup>a</sup> | MAF                      |
| SNP.9-1                      | 1.00    | 2.0E-03 | 0.09 | SNP.3-3                      | 1.00    | 0.79    | 0.36    | 0.09                 | 0.28                     |
| SNP.9-2                      | 4.2E-04 | 1.00    | 0.11 | SNP.3-4                      | 0.83    | 1.00    | 0.32    | 0.11                 | 0.27                     |
|                              |         |         |      | SNP.3-6                      | 0.48    | 0.45    | 1.00    | 0.33                 | 0.45                     |
|                              |         |         |      | SNP.3-7 <sup>a</sup>         | 0.16    | 0.21    | 0.28    | 1.00                 | 0.33 (0.67) <sup>a</sup> |
| <b>(A3) West group</b>       |         |         |      | <b>(B3) West group</b>       |         |         |         |                      |                          |
| SNPs as cofactors            | SNP.9-1 | SNP.9-2 | MAF  | SNPs as cofactors            | SNP.3-3 | SNP.3-4 | SNP.3-6 | SNP.3-7              | MAF                      |
| SNP.9-1                      | 1.00    | 0.22    | 0.13 | SNP.3-3                      | 1.00    | 0.69    | 0.34    | 0.04                 | 0.06                     |
| SNP.9-2                      | 0.41    | 1.00    | 0.14 | SNP.3-4                      | 0.77    | 1.00    | 0.37    | 0.07                 | 0.06                     |
|                              |         |         |      | SNP.3-6                      | 0.49    | 0.55    | 1.00    | 0.21                 | 0.12                     |
|                              |         |         |      | SNP.3-7                      | 0.09    | 0.12    | 0.21    | 1.00                 | 0.39                     |
| <b>(A4) South group</b>      |         |         |      | <b>(B4) South group</b>      |         |         |         |                      |                          |
| SNPs as cofactors            | SNP.9-1 | SNP.9-2 | MAF  | SNPs as cofactors            | SNP.3-3 | SNP.3-4 | SNP.3-6 | SNP.3-7              | MAF                      |
| SNP.9-1                      | 1.00    | 0.07    | 0.09 | SNP.3-3                      | 1.00    | 0.02    | 0.41    | 0.02                 | 0.11                     |
| SNP.9-2                      | 0.06    | 1.00    | 0.05 | SNP.3-4                      | 0.01    | 1.00    | 0.04    | 4.7E-03              | 0.01                     |
|                              |         |         |      | SNP.3-6                      | 0.43    | 0.07    | 1.00    | 0.03                 | 0.13                     |
|                              |         |         |      | SNP.3-7                      | 0.03    | 0.02    | 0.04    | 1.00                 | 0.17                     |

Values below the diagonal line refer to the usual  $r^2$  and above the diagonal line refer to  $r_{vs}^2$  (i.e., with correction for relatedness and population structure) MAF of the SNPs are given for the whole population and the three geographic groups.

<sup>a</sup> The allele found in the North-East group with the lowest frequency at the SNP.3-7 was the opposite to the one that appeared in the lowest frequency in the whole population and the other two geographic groups.

Pairwise LD between SNPs are highlighted using the following color scale: red ( $r^2 = 1$ ) and yellow ( $r^2 = 0$ ).

## Allele Frequencies, Effects and Genetic Variants for the SNPs Identified as Cofactors

For each of the eight SNPs identified as cofactors in the analysis of the whole population, the minor (i.e., less frequent) allele remained the same across the individual collections, the three geographic groups and the whole population, except for two cases: SLU collection and North+East group for SNP.3-7 associated with ripening period (Tables 3, 4A,B). In the two latter cases, the allele of SNP.3-7 with the lowest frequency was the alternate one to that in the other five collections and two geographic groups thus exhibiting a strong shift in the frequency of the G allele associated to early ripening period. Large differences between geographic groups were also observed for the minor allele frequencies (MAF) of the four SNPs associated with ripening period located on chromosome 3 (Table 4B), again indicating a North-South gradient.

The phenotypic effects of the two SNPs identified for flowering period indicated a strong mean difference (>1.9) between genotypic means for genotypes homozygous for the alternative alleles (Table S7; Figure S7). Dominance effects and epistatic interaction effects were significant, despite explaining a very small part of the variance (Table 5; Table S7). For ripening period, an even higher mean difference (frequently >2.5) was observed between the genotypic means of alternative homozygous genotypes for the four SNPs on chromosome 3 (Table S8; Figure S8), while less variation was found for SNP.10-1 and SNP.16-1. Variation in genotypic frequencies at each SNP was again very pronounced between geographic groups (Figure S8). Globally, dominance effects and epistatic interaction effects between the six SNPs were not significant, except some partial dominance occasionally observed for SNP.3-3, SNP.3-4, and SNP.10-1 (Table 5; Table S8).

The joint effect associated with the two SNPs identified for flowering period was assessed by comparing the average values for genotypes with different combinations of alleles in the whole population. Among the five genetic variants with a frequency above 1% (Table 6), variants 1 and 5 combining, respectively, the two alleles associated with early (GG/AA)

and late (TT/GG) flowering at a homozygous stage differed on average by 3.73 corresponding to 3.24  $\sigma$  (in standard-deviation units). The double heterozygous variant 3 (GT/GA) exhibited an intermediate value. For the four SNPs identified on chromosome 3 for ripening period, only 26 combinations out of the 81 potential variants were observed, 10 of which accounted for ~95% of the association panel (Table 6). The genetic variants accumulating homozygous alleles associated to early ripening period (variant 10: AA/TT/TT/GG) or late ripening period (variant 1: GG/CC/CC/AA) differed by 4.63 on average corresponding to 2.25  $\sigma$ . Out of the 397 genotypes belonging to variant 1, only 4.3% belonged to the North+East group, while 69.7 and 19.7% belonged to the West and South groups, respectively, representing 12, 35.7, and 52.7% of the total genotypes from North+East, West, and South groups, respectively. By contrast, the infrequent variant 10 (~2%) was common in the North+East group (52.2%) but more scarce (17.4%) and totally absent in the West and South groups, respectively. Multiple comparisons indicated no significant differences between variants 1 and 8, between variants 4, 5, 6, and 7, or between variants 9 and 10 (Table 6).

## Candidate Gene Identification

For flowering period, we considered the interval 451,830–635,974 bp (i.e., 184 kb) on chromosome 9, corresponding to the fusion of the 95% confidence intervals of the two SNPs selected as cofactors. In this interval, we found 28 gene models (Table S9) including putative transcription factors containing e.g., a NAM/NAC (MD09G1006400), a WRKY (MD09G1008800), a SBP (MD09G1008900) domain, and a putative glutaredoxin (MD09G1007400). In a second run, we also considered the 95% confidence interval covering SNP.9-5 which was selected in 25 subsets of the re-sampling analysis (Table S4) despite not being detected in the initial analysis. The corresponding interval 654,780–811,891 bp (i.e., 157 kb) was almost contiguous to the previous one, thus defining a wider region of ~360 kb (451,830–811,891 bp). Thirty-eight additional gene models were found in this enlarged interval (Table S9) including a putative SRF transcription factor containing a MADS- and a K-box (MD09G1009100), another putative SRF transcription factor (not detected by automatic annotation pipeline), and a gene model containing a SWIB/MDM2 domain (MD09G1011600).

For ripening period, we considered two intervals on chromosome 3, one corresponding to the fusion of the 95% confidence intervals of SNP.3-6 and SNP.3-7 which overlapped (30,624,429–30,802,006 bp, i.e., 178 kb), and the second for the confidence interval of SNP.3-3 and SNP.3-4 (30,354,359–30,540,756 bp, i.e., 186 kb). Only eleven gene models were found in the first interval and 24 in the second, with 6 additional gene models in between (Table S10). Two successive genes encoding a putative transcription factor containing a NAM/NAC domain (MD03G1222600 and MD03G1222700) were found in the very close vicinity of SNP.3-6 and SNP.3-7, both SNPs located in between the two genes. An Ultrapetala transcription factor (MD03G1220200) was found close to SNP.3-3, and a protein tyrosine kinase

**TABLE 5 |** Test of dominance and epistatic effects among the SNPs selected as cofactors in the GWAS of the whole population for flowering and ripening periods.

| Trait            | Effects              | d.f.            | F-test | p-value  | PVE (%) |
|------------------|----------------------|-----------------|--------|----------|---------|
| Flowering period | Additive             | 2               | 79.8   | 4.3E-33  | 9.1     |
|                  | Dominance            | 2               | 5.9    | 2.7E-03  | 0.7     |
|                  | Epistatic            | 4               | 9.0    | 3.7E-07  | 2.0     |
| Ripening period  | Additive             | 6               | 106.6  | 3.4E-106 | 17.4    |
|                  | Dominance            | 6               | 1.1    | 3.6E-01  | 0.2     |
|                  | Dominance of SNP.3-4 | 1               | 4.8    | 2.8E-02  | 0.1     |
|                  | Dominance of SNP.3-6 | 1               | 4.5    | 3.3E-02  | 0.1     |
|                  | Epistatic            | 41 <sup>a</sup> | 1.2    | 2.3E-01  | 1.3     |

<sup>a</sup>Some combinations of SNP genotypes did not exist in the whole population, which reduced the df for all interactions between 6 SNPs.



**TABLE 6** | Joint effect of the two SNPs associated with flowering period in the whole population and of the ten most frequent genetic variants defined by the four SNPs on chromosome 3 associated with ripening period in the whole population.

| Genetic variant         | Genotypes at SNPs <sup>a,b</sup> | N° cultivars | Frequency | Mean | Median | SD   | Min  | Max  | Tukey groups |
|-------------------------|----------------------------------|--------------|-----------|------|--------|------|------|------|--------------|
| <b>FLOWERING PERIOD</b> |                                  |              |           |      |        |      |      |      |              |
| Variant 1               | <b>GG/AA</b>                     | 760          | 0.67      | 4.65 | 4.73   | 0.95 | 1.73 | 8.87 | a            |
| Variant 2               | <b>GT/GA</b>                     | 126          | 0.11      | 5.83 | 5.73   | 1.32 | 2.73 | 8.88 | b            |
| Variant 3               | <b>GG/GA</b>                     | 121          | 0.11      | 4.43 | 4.47   | 1.13 | 1.73 | 8.24 | a            |
| Variant 4               | <b>GT/AA</b>                     | 89           | 0.08      | 4.49 | 4.67   | 1.00 | 2.34 | 7.37 | a            |
| Variant 5               | TT/GG                            | 11           | 0.01      | 8.38 | 7.82   | 1.08 | 5.73 | 9.24 | c            |
| <b>RIPENING PERIOD</b>  |                                  |              |           |      |        |      |      |      |              |
| Variant 1               | GG/CC/CC/AA                      | 397          | 0.35      | 6.75 | 6.80   | 1.43 | 2.21 | 9.84 | a            |
| Variant 2               | GG/CC/CC/ <b>AG</b>              | 336          | 0.29      | 5.69 | 5.63   | 1.35 | 0.88 | 9.50 | b            |
| Variant 3               | GG/CC/CC/ <b>GG</b>              | 73           | 0.06      | 4.72 | 4.81   | 1.61 | 1.21 | 8.48 | c            |
| Variant 4               | GG/CC/ <b>CT/AG</b>              | 61           | 0.05      | 3.75 | 3.74   | 1.73 | 0.88 | 8.77 | d            |
| Variant 5               | <b>AG/CT/CT/AG</b>               | 59           | 0.05      | 3.75 | 3.85   | 1.35 | 1.21 | 7.15 | d            |
| Variant 6               | <b>AG/CT/CT/GG</b>               | 44           | 0.04      | 2.89 | 2.54   | 1.32 | 0.54 | 6.85 | d            |
| Variant 7               | GG/CC/ <b>CT/GG</b>              | 39           | 0.03      | 3.49 | 3.21   | 1.61 | 0.55 | 6.66 | d            |
| Variant 8               | <b>AG/CC/CT/AA</b>               | 29           | 0.03      | 7.44 | 7.67   | 1.51 | 4.82 | 9.80 | a            |
| Variant 9               | <b>AG/CT/TT/GG</b>               | 28           | 0.02      | 2.01 | 2.09   | 0.62 | 1.16 | 4.14 | e            |
| Variant 10              | <b>AA/TT/TT/GG</b>               | 23           | 0.02      | 2.11 | 2.21   | 1.10 | 0.54 | 4.42 | e            |

<sup>a</sup>The allele associated with an early flowering period is highlighted in bold; order of SNPs is as follows: SNP:9-1/SNP:9-2.

<sup>b</sup>The allele associated with an early ripening period is highlighted in bold; order of SNPs is as follows: SNP:3-3/SNP:3-4/SNP:3-6/SNP:3-7.

(MD03G1221300) close to SNP.3-4. On chromosome 10, we considered the 95% confidence interval 37,695,471–39,085,497 bp for SNP.10-1 and found 153 gene models (Table S11). Among them were four putative transcription factors, two of which contained a NAM/NAC domain (MD10G1288300 and MD10G1299900) while another two contained an Apetala-2 domain (MD10G1290400 and MD10G1290900). A carbohydrate phosphorylase putatively involved in starch metabolism was also identified (MD10G1289300). On chromosome 16, we considered a 95% confidence interval 8,933,453–9,359,141 bp for SNP.16-1 and found 38 gene models (Table S12). Together with two putative transcription factors encoding either a NAM/NAC domain or a TIFY domain (MD16G1125800 and MD16G1127400, respectively), we especially identified a gene model encoding an auxin responsive protein (MD16G1124300) and another gene model encoding a sugar bidirectional transporter (MD16G1125300).

A nearly perfect microsynteny (with some minor rearrangement) was revealed between apple and peach in all the four confidence interval genomic regions estimated in our study, as shown by the numerous conserved homologs between the two species in those regions (Tables S9–S12).

## DISCUSSION

### Genomic Regions Controlling Variation in Phenological Traits

The SNPs retained as cofactors in the GWAS on the whole population defined one genomic region controlling flowering period and three controlling ripening period. Additional regions were identified when conducting GWAS for individual

collections. The associations found accounted for varying levels of trait variation (0–33% for flowering period; 12–28% for ripening period) across the whole population and individual collections. We applied a conservative approach in identifying SNPs as cofactors for *p*-values below a defined threshold of Bonferroni correction at 5%. Implementation of those stringent parameters was essential to eliminate false positives, but have probably sacrificed some true associations with small effects.

The top of chromosome 9 was recently indicated as being involved in the genetic control of flowering or bud burst period (Celton et al., 2011; Allard et al., 2016). The regions pointed out in these contributions overlap with the confidence interval found in our study although the region indicated by Allard et al. (2016) is shifted slightly downstream since the very top of the chromosome was not mapped in their experiment. The regions indicated in these studies were, however, much larger than the confidence interval we report: Celton et al. (2011) examined a region of almost 16 cM corresponding to 4.04 Mb and comprising 983 gene models in the apple genome v1.0 of the Genome Database for *Rosaceae* (GDR, <https://www.rosaceae.org/>), whereas Allard et al. (2016) indicated a region of 10 cM corresponding to 1.8 Mb and comprising 622 gene models. The numerous recombination events accumulated in our association panel reduced the associated region to 184 kb with only 28 gene models in the GDDH13 genome. An extended interval of ~360 kb was nevertheless proposed to take into account the results of the re-sampling analysis, thus generating a final set of 66 candidate gene models. Interestingly, Trainin et al. (2016) identified a common haplotype on the top of chromosome 9 shared by a small subset of mostly Israeli apple cultivars adapted to low-chill conditions such as the well-known “Anna.” They

defined an interval of about 1.7 Mb but suggested that the genetic factor/s responsible for early bud-break could be located in a region of about only 190 kb (between SNP-A6-2 and SNP-A4). Mapping these SNPs on the GDDH13 genome, we found the corresponding interval to be 730,978–923,844 bp, which overlaps the extended interval accounting for SNP.9-5 (451,830–811,891 bp). This co-localization raises the question of the allelic control of flowering period in that particular genomic region as described below.

Chromosomes 3, 10, and 16 have shown associations with ripening period in previous linkage mapping studies (Liebhard et al., 2003; Kenis et al., 2008; Chagné et al., 2014; Kunihisa et al., 2014). None of these studies attempted to define a confidence interval for the physical position of the reported QTLs, thus preventing an accurate comparison of the precision in QTL location between studies. Recently, Migicovsky et al. (2016) did not find any associations with ripening period on chromosomes 10 and 16 in a GWAS based on single-locus tests, but identified associations with two SNPs on chromosome 3 located within the coding region of NAC18.1 (GenBank ID: NM\_001294055.1) which corresponds to a gene model (MD03G1222600) at position ~30,697,000 bp of GDDH13 genome. Interestingly, this position fits perfectly within the 95% confidence interval of SNP.3-6/SNP.3-7 (30,624,429–30,802,006 bp). Since this genomic region has been identified in various environments and genetic backgrounds, it therefore appears to potentially be a major factor in the genetic control of ripening period.

## GWAS on Phenological Traits Suggests Presence of Allelic Heterogeneity

For each trait, MLM analysis for the whole population retained SNPs in weak LD despite being in close vicinity. Two SNPs retained as cofactors for flowering period on chromosome 9 were only 27 kb apart. Four SNPs retained for ripening period on chromosome 3 spanned a region of 296 kb, with two sub-regions spanning only 35 and 26 kb, respectively. Identification of multiple significant SNPs within or near a single gene may suggest either allelic heterogeneity or the presence of an untyped causal variant that requires multiple SNPs to be adequately tagged, or both (Atwell et al., 2010; Dickson et al., 2010; Segura et al., 2012). Allelic heterogeneity refers to the presence of more than two functional alleles of a given gene affecting a phenotypic trait (Wood et al., 2011). Indeed, the biallelic nature of SNPs reduces their ability to tag multiple alleles and explains the need for several SNPs to tag them. Also, maximizing the genetic variance in the association panel by including geographically distant accessions with both different and complex evolutionary histories is expected to improve resolution, but has the potential to introduce genetic heterogeneity (i.e., multiple causal variants with various dates of appearance and frequencies) which can generate false “synthetic” associations when only single-locus tests are used (Korte and Farlow, 2013). Fortunately, the MLM approach is able to disentangle the contribution of genetic heterogeneity by including “competing” variants as cofactors within the mixed model setting and thus helps to discard false

“synthetic” associations (Segura et al., 2012; Korte and Farlow, 2013). For flowering period, the two detected SNPs can either fit with allelic heterogeneity or untyped causal variant requesting more than one SNP. But more interestingly, the co-localization of our confidence interval with the small genomic region identified by Trainin et al. (2016) for the extreme phenotype of low-chilling requirement, opens the question of the local genomic architecture of this trait. Since bud-break and consequently flowering period of Israeli cultivars occur much earlier than in traditional European cultivars (Trainin et al., 2016), either two different polymorphic genes or a single gene with at least three alleles may be responsible for the co-location of detectable genotypic variation for flowering period and early bud-break. In the latter case, at least two alleles would control the genotypic difference we observed here for flowering period, and another more “extreme” allele would confer the early bud-break of Israeli cultivars. Alternatively, this extreme allele could be proposed as an epi-allele when considering epigenetic control (Ríos et al., 2014). For ripening period, a model including the two nearby genomic regions detected on chromosome 3 can also be proposed with the presence of two closely positioned genes (~300 kb apart), each with possible allelic heterogeneity. Such a complex pattern of association has never been highlighted before for flowering and ripening periods in apple. Nevertheless, additional genetic studies would be required to be certain about the multi-allelic and multi-genic architecture of the detected regions by using e.g., local haplotype sharing methods (Xu and Guan, 2014) provided that sufficient SNPs are available.

## Unexplained Genetic Variation May Be Accounted for by Multiple Factors

The limited number of detected genomic regions associated with the traits and the low/moderate amount of phenotypic variance accounted for by the retained SNPs suggests that several, if not many additional genomic regions are involved in the genetic control of these traits. Here, as with other GWAS, we were challenged by the so-called “missing heritability” syndrome (i.e., traits exhibiting both high heritability and tiny effect variants; Maher, 2008; Manolio et al., 2009; Visscher et al., 2010; Yang et al., 2010; Zuk et al., 2012). In our experiment, a significant proportion of the phenotypic variance not captured by the SNP cofactors could be explained by relatedness accounting for polygenic effects (15–58% for flowering period, 8–41% for ripening period) and population structure mostly accounting for genetic differentiation over geographic groups (7–36% for flowering period, 30–53% for ripening period). The large proportion of phenotypic variance under genetic control clearly indicates that additional genomic regions are still to be discovered. Interestingly, at the whole population level, the part of variance explained by relatedness for flowering period (39%) was more than twice the estimate for ripening period (16%), while the inverse was observed for the part of variance explained by structure (27% for flowering period, 52% for ripening period), thus indicating differential impact of relatedness and geographic structure on these two phenological traits.

Several factors may have hampered the detection of additional genomic regions. Genetic architecture consisting of many common variants with small effects and/or rare variants with large effects can reduce the statistical power of GWAS (Brachi et al., 2011; Gibson, 2011; Stranger et al., 2011; Korte and Farlow, 2013). The wide diversity in our association panel may have favored the inclusion of several rare variants with strong effects that could not be detected in the present study. The rapid LD decay and the LD pattern between causal variants and genotyped SNPs are two other limiting factors (Manolio et al., 2009; Visscher et al., 2010; Stranger et al., 2011). Despite the use of a high-density SNP array, it is possible that some genomic regions with causal variants were insufficiently covered by SNPs (i.e., null or incomplete LD), thus preventing detection of the corresponding variance. Denser genotyping may be required to find new associations given that both their effect and frequency are large enough to be detected by GWAS. Also, other factors may account for the unexplained genetic variation: (i) quality and precision of the phenotypic (historical) data (Myles et al., 2009; Migicovsky et al., 2016), (ii) genotype  $\times$  environment (G $\times$ E) interactions, (iii) epistatic effects that were not systematically investigated in our experiment, or even, (iv) epigenetic variation.

## Population Structure and Geographic Adaptation

Our association panel consisted mostly of local and/or old dessert apple cultivars selected as representative subsets by each institute. The phenotypic differences observed in the geographic-scale analyses (North+East, South and West groups) are probably explained by adaptive selection to different environments. Adaptive traits are frequently filtered by environmental gradients that coincide with patterns of population structure due to the differential fixation of alleles among groups of cultivars, following diversifying selection and/or genetic drift (Atwell et al., 2010; Brachi et al., 2011; Lasky et al., 2015; Nicolas et al., 2016). Despite genetic structure being weak at the whole population scale in our study (only 17% of the genotypic variation was explained by the ten largest Eigenvalues of the PCA), this structure explained a moderate (flowering period: 27%) or even high (ripening period: 52%) proportion of the phenotypic variance in GWAS. These results are in line with the phenotypic differences observed at a geographic scale, since the first two principal components were highly associated with geographic grouping (30 and 37%). A similar observation was made by Migicovsky et al. (2016). Differential selection together with genetic drift where the germplasm originated (North+East, West and South) may have favored or selected specific alleles or combinations of alleles in different geographic regions/environments. A good example is given by SNP.3-7, which was associated to ripening period with a frequency of 67% for its G allele in the North+East group but only 17% in the South group. Similarly, when considering the genetic variants combining the four SNPs retained on chromosome 3, variant 10 combining all earliness-associated SNP alleles at a homozygous state, was very common in accessions of the North+East group while totally absent in the South. In apple, harvest period is

probably the trait with the strongest impact of geographical adaptation, since local weather conditions define the length of harvesting season.

## Putative Functions of Genes Controlling Phenotypic Variation in Apple Flowering Period

Gene models of particular interest were identified in the interval defined on chromosome 9, including a putative NAC gene (MD09G1006400). NAC-domain proteins are transcription factors involved in the genetic control of flowering time in *Arabidopsis* (Yoo et al., 2007), where two NAC proteins in association with a JM14 gene (a histone demethylase) apparently take part in flowering time regulation (Ning et al., 2015). In addition, a putative WRKY transcription factor was identified. This gene model (MD09G1008800, corresponding to MDP0000154734 in GDR) was cited by Trainin et al. (2016) as a putative candidate for early bud-break of Israeli apple cultivars. The WRKY gene family was recently proposed to play a role in dormancy regulation in peach (Chen et al., 2016). Based on RNAseq, MD09G1008800 transcription was detected mainly in apple roots and only slightly in fruits, thus limiting its potential role in flowering.

Three other candidate gene models are of special interest for the genetic control of flowering period. MD09G1009100 and another non-predicted gene model are similar to SRF transcription factors containing a MADS domain, putatively homologous to the *FLC* (*FLOWERING LOCUS C*) gene involved with the *FRIGIDA* gene in vernalization response of *Arabidopsis* (reviewed by Amasino and Michaels, 2010). MADS-box genes, such as the DAM (dormancy associated MADS-box) family members, were previously shown to be the master regulators of dormancy establishment and maintenance in *Prunus* and *Pyrus* species (Bielenberg et al., 2008; Ubi et al., 2010; Yamane et al., 2011; Saito et al., 2013; Sánchez-Pérez et al., 2014; Zhebentyayeva et al., 2014). Related DAM-like genes with dormancy-dependent expression have been identified in other perennial species such as leafy spurge (Horvath et al., 2008, 2010), raspberry (Mazzitelli et al., 2007), blackcurrant (Hedley et al., 2010), and kiwifruit (Wu et al., 2012). Also, MD09G0010600 is predicted as a SWIB/MDM2-domain containing gene, a member of a family of chromatin remodeling complexes that modify DNA accessibility by restructuring nucleosomes (Jerzmanowski, 2007). These three genes (MDP0000167381/MDP0000126259, MDP0000296123, and MDP0000315892/MDP0000317368 in GDR v1.0, respectively) were also mentioned by Trainin et al. (2016). Conversely, the other candidate genes highlighted by these authors were located outside of our largest confidence interval, as were all the candidate genes cited by Celton et al. (2011). Finally, special attention should be given to the MADS-domain containing gene (MD09G1009100 = MDP0000167381 = MDP0000126259 in its shorter version) since it was upregulated in several differential expression situations when comparing the low chilling requirement sport “Castel Gala” with “Royal Gala” (Porto et al., 2015). By contrast, the other two genes (MADS-box: MDP0000207984, and *PRE1*-like:

MDP0000320691) highlighted by Porto et al. (2015), were located either on another chromosome or considerably downstream on chromosome 9.

Whilst the candidate genes we identified did not encompass all of those that have previously been proposed to have a role in flowering time, it is clear that a number of them have putative roles in the regulation of flowering time in apple or other plants.

## Putative Functions of Genes Controlling Phenotypic Variation in Apple Ripening Period

Concerning ripening period, three main genomic regions were identified (on chromosomes 3, 10, and 16) with candidate genes belonging to the NAC family, surrounded by other genes putatively involved in apple ripening. In the genomic region of chromosome 3, two NAC transcription factors (MD03G1222600 and MD03G1222700) are strongly indicated as candidate genes for the control of this trait. A member of this gene family (i.e., ppa008301m, according to the *P. persica* genome version v1.0) was identified in a major locus on chromosome 4 controlling maturity date in peach, and a 9 bp DNA insertion in its last exon was described as a variant putatively linked to early ripening (Pirona et al., 2013). Most interestingly, one of the two NAC genes of apple (i.e., MD03G1222700) cited above showed to be the best homolog of this particular peach NAC gene which was renamed Prupe.4G186800 in the *P. persica* genome version v2.1 (Verde et al., 2017). The second apple NAC gene was also showed to be homolog to the second peach NAC gene cited by Pirona et al. (2013) (i.e., ppa007577m.v1.0 equivalent to Prupe.4G187100.v2.1), and a strong microsynteny was observed between *Malus* and *Prunus* all along the analyzed confidence interval (Table S10). The importance of NAC transcription factors in controlling fruit ripening traits has been described also in tomato (Zhu et al., 2014) and kiwifruit (Nieuwenhuizen et al., 2015). Very recently, two NAC members (called SINAC4/9) were indicated as regulators of ethylene biosynthesis and ethylene-related genes in tomato (Kou et al., 2016).

The genes identified on chromosome 10 appeared to be involved in the same metabolic pathways: two NAM/NAC (MD10G1288300 and MD10G1299900) and two Apetala2 (MD10G1290400 and MD10G1290900) transcription factors. Members of the plant-specific APETALA2/ethylene response factor (AP2/ERF) superfamily of transcription factors act downstream of the ethylene signaling pathway and are strongly conserved throughout the plant kingdom (Xie et al., 2016). They are apparently associated with several plant developmental and growth processes, including fruit ripening (Licausi et al., 2010; Karlova et al., 2014; Xie et al., 2016).

On chromosome 16, two additional putative transcription factors encoding either a NAM/NAC domain (MD16G1125800) or a TIFY domain (MD16G1127400) were identified as well as one gene encoding an auxin responsive protein (MD16G1124300) and one gene for a sugar bidirectional

transporter (MD16G1125300) with high homology with a senescence associated protein (SAG 29) of *Arabidopsis*. TIFY transcription factors comprise a plant-specific family involved in the regulation of various developmental processes and responses to phytohormones. Among the 30 members of this family characterized in apple (Li et al., 2015), are the jasmonate zim-domain (JAZ) proteins, known to be repressors of JA signaling and, consequently, actors of the cross-talk among multiple hormone signaling pathways including ethylene and gibberellins (An et al., 2016). The expression patterns of genes in the JA biosynthesis pathway was found to be correlated with genes in the ethylene biosynthesis pathway, emphasizing the role of JA biosynthesis and its signaling on apple fruit maturation (Lv et al., 2015).

Altogether, candidate genes identified after GWAS highlight the probable role of transcription factors, controlling the ethylene biosynthesis or regulatory pathway, for ripening in apple. Other candidates such as the gene encoding for an auxin responsive protein may also be considered since an ethylene–auxin interplay at a late ripening stage has been proposed in apple (Tadiello et al., 2016).

## FUTURE PERSPECTIVES

GWAS mapping is a powerful tool for the identification of genomic regions associated with important traits, but results can be restricted by too much genetic heterogeneity, insufficient marker density and an overly strong impact of population structure. In the present study, narrow genomic regions controlling two phenological traits and a rather low number of candidate genes were identified, while other regions remained unidentified because of the relationship between traits and geographic structure. Enlarging the diversity panel with more genotypes, especially from Southern and Northern+Eastern groups, might improve detection of loci associated with those traits in each geographic group. Also, a combination of linkage and association analyses may achieve higher statistical power and resolution (Jansen et al., 2003; Flint-Garcia et al., 2005; Pascual et al., 2016) and reduce the confidence interval of the detected genomic regions which would call for validating the function of certain candidate genes by genetic transformation, especially gene editing (Busov et al., 2005; Nishitani et al., 2016). Still, the current set of phenotypic and genotypic data is already useful to establish genome-wide predictions (Meuwissen et al., 2001; Muranty et al., 2015) of the breeding values of the studied genotypes for these two traits.

## AUTHOR CONTRIBUTIONS

JU, HM, and DL carried out the statistical analyses under the supervision of C-ED and ST. CD and ER managed the leaf sample collection and part of DNA extraction. AT performed the remaining DNA extraction and the whole runs of the Axiom<sup>®</sup> plates on the GeneTitan<sup>®</sup> under the supervision of CP. CD performed the whole SNP genotyping analysis and validation



process with the help of HM and C-ED. Selection of germplasm and acquiring phenotypic data were performed by C-ED, AG, RG, LF, ML, PH, MO, FP, JS, HN, RGr, LD, and ST. MB and SM brought expertise about GWAS methodology and results interpretation. SA performed the candidate gene analysis thanks to the new apple genome sequence developed by ND and J-MC. C-ED conceived and coordinated the study. FL coordinated the EU FruitBreedomics project including this study. JU, HM, and C-ED wrote the manuscript with decisive contributions of SA and LD. HN, MO, MT, LB, RV, MB, LG-G, and SM critically reviewed the manuscript. All authors read and approved the final manuscript.

## FUNDING

This work has been funded under the EU seventh Framework Programme by the FruitBreedomics project No. 265582: Integrated Approach for increasing breeding efficiency in fruit tree crops (<http://www.fruitbreedomics.com/>).

## REFERENCES

- Abbott, A. G., Zhebentyayeva, T., Barakat, A., and Liu, Z. (2015). "The genetic control of bud-break in trees," in *Advances in Botanical Research*, Vol. 74, eds P. Christophe and A. B. Anne-Françoise (San Diego, CA: Academic Press), 201–228.
- Allard, A., Legave, J. M., Martinez, S., Kelner, J. J., Bink, M. C. A. M., Di Guardo, M., et al. (2016). Detecting QTLs and putative candidate genes involved in budbreak and flowering time in an apple multiparental population. *J. Exp. Bot.* 67, 2875–2888. doi: 10.1093/jxb/erw130
- Amasino, R. M. (2005). Vernalization and flowering time. *Curr. Opin. Biotechnol.* 16, 154–158. doi: 10.1016/j.copbio.2005.02.004
- Amasino, R. M., and Michaels, S. D. (2010). The timing of flowering. *Plant Physiol.* 154, 516–520. doi: 10.1104/pp.110.161653
- An, X. H., Hao, Y. J., Li, E. M., Xu, K., and Cheng, C. G. (2016). Functional identification of apple MdJAZ2 in *Arabidopsis* with reduced JA-sensitivity and increased stress tolerance. *Plant Cell Rep.* 36, 255–265. doi: 10.1007/s00299-016-2077-9
- Aranzana, M. J., Kim, S., Zhao, K., Bakker, E., Horton, M., Jakob, K., et al. (2005). Genome-wide association mapping in *Arabidopsis* identifies previously known flowering time and pathogen resistance genes. *PLoS Genet.* 1:e60. doi: 10.1371/journal.pgen.0010060
- Atwell, S., Huang, Y. S., Vilhjálmsson, B. J., Willems, G., Horton, M., Li, Y., et al. (2010). Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. *Nature* 465, 627–631. doi: 10.1038/nature08800
- Balding, D. J. (2006). A tutorial on statistical methods for population association studies. *Nat. Rev. Genet.* 7, 781–791. doi: 10.1038/nrg1916
- Bates, D. M., and Sarkar, D. (2007). *lme4: Linear Mixed-effects Models using Eigen and Classes*. R package version 0.99875–99876.
- Bianco, L., Cestaro, A., Linsmith, G., Muranty, H., Denancé, C., Théron, A., et al. (2016). Development and validation of the Axiom®Apple 480K SNP genotyping array. *Plant J.* 86, 62–74. doi: 10.1111/tj.13145
- Bianco, L., Cestaro, A., Sargent, D. J., Banchi, E., Derdak, S., Di Guardo, M., et al. (2014). Development and validation of a 20K single nucleotide polymorphism (SNP) whole genome genotyping array for apple (*Malus × domestica* Borkh.). *PLoS ONE* 9:e110377. doi: 10.1371/journal.pone.0110377
- Bielenberg, D. G., Wang, Y., Li, Z., Zhebentyayeva, T., Fan, S., Reighard, G. L., et al. (2008). Sequencing and annotation of the evergrowing locus in peach (*Prunus Persica* [L.] Batsch) reveals a cluster of six MADS-box transcription factors as

## ACKNOWLEDGMENTS

The authors thank Vincent Segura for his help in defining the process for testing dominance and epistatic effects with MLM, the Migale (<http://migale.jouy.inra.fr/>) and GenoToul (<http://bioinfo.genotoul.fr/>) bioinformatics platforms for giving access to computing facilities, the past and present curatorial staff and field technicians for maintaining the apple collections at INRA Experimental Unit (UE Horti), CRA-W, RBIPH, AUB-UNIBO, NFC-Brogdale (UK), and SLU, and acknowledge Defra (UK) for supporting the characterization of the collections. JU has been partially supported by an Early Stage Research Fellowship of the Institute of Advanced Studies (University of Bologna).

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fpls.2017.01923/full#supplementary-material>

- candidate genes for regulation of terminal bud formation. *Tree Genet. Genomes* 4, 495–507. doi: 10.1007/s11295-007-0126-9
- Boss, P. K., Bastow, R. M., Mylne, J. S., and Dean, C. (2004). Multiple pathways in the decision to flower: enabling, promoting, and resetting. *Plant Cell* 16, S18–S31. doi: 10.1105/tpc.015958
- Brachi, B., Morris, G. P., and Borevitz, J. O. (2011). Genome-wide association studies in plants: the missing heritability is in the field. *Genome Biol.* 12:232. doi: 10.1186/gb-2011-12-10-232
- Busov, V. B., Brunner, A. M., Meilan, R., Filichkin, S., Ganio, L., Gandhi, S., et al. (2005). Genetic transformation: a powerful tool for dissection of adaptive traits in trees. *New Phytol.* 167, 9–74. doi: 10.1111/j.1469-8137.2005.01412.x
- Campoy, J. A., Ruiz, D., and Egea, J. (2011). Dormancy in temperate fruit trees in a global warming context: a review. *Sci. Hortic.* 130, 357–372. doi: 10.1016/j.scienta.2011.07.011
- Cannell, M. G. R., and Smith, R. I. (1986). Climate warming, spring budburst and frost damage on trees. *J. Appl. Ecol.* 23, 177–191. doi: 10.2307/2403090
- Castède, S., Campoy, J. A., Quero-García, J., Le Dantec, L., Lafargue, M., Barreneche, T., et al. (2014). Genetic determinism of phenological traits highly affected by climate change in *Prunus avium*: flowering date dissected into chilling and heat requirements. *New Phytol.* 202, 703–715. doi: 10.1111/nph.12658
- Celton, J. M., Martinez, S., Jammes, M. J., Bechti, A., Salvi, S., Legave, J. M., et al. (2011). Deciphering the genetic determinism of bud phenology in apple progenies: a new insight into chilling and heat requirement effects on flowering dates and positional candidate genes. *New Phytol.* 192, 378–392. doi: 10.1111/j.1469-8137.2011.03823.x
- Chagné, D., Crowhurst, R. N., Troggio, M., Davey, M. W., Gilmore, B., Lawley, C., et al. (2012). Genome-wide SNP detection, validation, and development of an 8K SNP array for apple. *PLoS ONE* 7:e31745. doi: 10.1371/journal.pone.0031745
- Chagné, D., Dayatilake, D., Diack, R., Oliver, M., Ireland, H., Watson, A., et al. (2014). Genetic and environmental control of fruit maturation, dry matter and firmness in apple (*Malus × domestica* Borkh.). *Hortic. Res.* 1:14046. doi: 10.1038/hortres.2014.46
- Chen, J., and Chen, Z. (2008). Extended Bayesian information criteria for model selection with large model space. *Biometrika* 95, 759–771. doi: 10.1093/biomet/asn034
- Chen, M., Tan, Q., Sun, M., Li, D., Fu, X., Chen, X., et al. (2016). Genome-wide identification of WRKY family genes in peach and analysis of WRKY expression during bud dormancy. *Mol. Genet. Genomics* 291, 1319–1332. doi: 10.1007/s00438-016-1171-6

- Cook, B. I., Wolkovich, E. M., and Parmesan, C. (2012). Divergent responses to spring and winter warming drive community level flowering trends. *Proc. Natl. Acad. Sci. U.S.A.* 109, 9000–9005. doi: 10.1073/pnas.1118364109
- Daccord, N., Celton, J. M., Linsmith, G., Becker, C., Choisine, N., Schijlen, E., et al. (2017). High-quality *de novo* assembly of the apple genome and methylome dynamics of early fruit development. *Nat. Genet.* 49, 1099–1106. doi: 10.1038/ng.3886
- Dickson, S. P., Wang, K., Krantz, I., Hakonarson, H., and Goldstein, D. B. (2010). Rare variants create synthetic genome-wide associations. *PLoS Biol.* 8:e1000294. doi: 10.1371/journal.pbio.1000294
- Di Guardo, M., Bink, M. C. A. M., Guerra, W., Letschka, T., Lozano, L., Busatto, N., et al. (2017). Deciphering the genetic control of fruit texture in apple by multiple family-based analysis and genome-wide association. *J. Exp. Bot.* 68, 1451–1466. doi: 10.1093/jxb/erx017
- Dirlwanger, E., Quero-Garcia, J., Le Dantec, L., Lambert, P., Ruiz, D., Dondini, L., et al. (2012). Comparison of the genetic determinism of two key phenological traits, flowering and maturity dates, in three *Prunus* species: peach, apricot and sweet cherry. *Heredity* 109, 280–292. doi: 10.1038/hdy.2012.38
- Ensminger, I., Schmidt, L., and Lloyd, J. (2008). Soil temperature and intermittent frost modulate the rate of recovery of photosynthesis in Scots pine under simulated spring conditions. *New Phytol.* 177, 428–442. doi: 10.1111/j.1469-8137.2007.02273.x
- Erez, A. (2000). “Bud dormancy; phenomenon, problems and solutions in the tropics and subtropics” in *Temperate Fruit Crops in Warm Climates*, ed A. Erez (Dordrecht: Kluwer Academic Publishers), 17–48. doi: 10.1007/978-94-017-3215-4\_2
- Farneti, B., Di Guardo, M., Khomenko, I., Cappellin, L., Biasioli, F., Velasco, R., et al. (2017). Genome-wide association study unravels the genetic control of the apple volatolome and its interplay with fruit texture. *J. Exp. Bot.* 68, 1467–1478. doi: 10.1093/jxb/erx018
- Fleckinger, J. (1964). “Phénologie et arboriculture fruitière,” in *Le bon jardinier, (Tome I, 2ème partie)*, eds P. Grisvard and V. C. Chaudun (La Maison Rustique), 362–372.
- Flint-Garcia, S. A., Thuillet, A., Yu, J., Pressoir, G., Romero, S. M., Mitchell, S. E., et al. (2005). Maize association population: a high-resolution platform for quantitative trait locus dissection. *Plant J.* 44, 1054–1064. doi: 10.1111/j.1365-313X.2005.02591.x
- Fox, J. (2003). Effect displays in R for generalised linear models. *J. Stat. Soft.* 8, 1–27. doi: 10.18637/jss.v008.i15
- Gardner, K. M., Brown, P., Cooke, T. F., Cann, S., Costa, F., Bustamante, C., et al. (2014). Fast and cost-effective genetic mapping in apple using next-generation sequencing. *G3* 4, 1681–1687. doi: 10.1534/g3.114.011023
- Gibson, G. (2011). Rare and common variants: twenty arguments. *Nat. Rev. Genet.* 13, 135–145. doi: 10.1038/nrg3118
- Hänninen, H., and Tanino, K. (2011). Tree seasonality in a warming climate. *Trends Plant Sci.* 16, 412–416. doi: 10.1016/j.tplants.2011.05.001
- Hayes, B. (2013). Overview of statistical methods for Genome-Wide Association Studies (GWAS). *Methods Mol. Biol.* 1019, 149–169. doi: 10.1007/978-1-62703-447-0\_6
- Hedley, P. E., Russell, J. R., Jorgensen, L., Gordon, S., Morris, J. A., Hackett, C. A., et al. (2010). Candidate genes associated with bud dormancy release in blackcurrant (*Ribes nigrum* L.). *BMC Plant Biol.* 10:202. doi: 10.1186/1471-2229-10-202
- Horvath, D. P., Chao, W. S., Suttle, J. C., Thimmapuram, J., and Anderson, J. V. (2008). Transcriptome analysis identifies novel responses and potential regulatory genes involved in seasonal dormancy transitions of leafy spurge (*Euphorbia esula* L.). *BMC Genomics* 9:536. doi: 10.1186/1471-2164-9-536
- Horvath, D. P., Sung, S., Kim, D., Chao, W., and Anderson, J. (2010). Characterization, expression and function of DORMANCY ASSOCIATED MADS-BOX genes from leafy spurge. *Plant Mol. Biol.* 73, 169–179. doi: 10.1007/s11103-009-9596-5
- Ingvarsson, P. K., and Street, N. R. (2011). Association genetics of complex traits in plants. *New Phytol.* 189, 909–922. doi: 10.1111/j.1469-8137.2010.03593.x
- Ionescu, I. A., Moller, B. L., and Sánchez-Pérez, R. (2017). Chemical control of flowering time. *J. Exp. Bot.* 68, 369–382. doi: 10.1093/jxb/erw427
- Jansen, R. C., Jannink, J. L., and Beavis, W. D. (2003). Mapping quantitative trait loci in plant breeding populations: use of parental haplotype sharing. *Crop Sci.* 43, 829–834. doi: 10.2135/cropsci2003.8290
- Jerzmanowski, A. (2007). SWI/SNF chromatin remodeling and linker histones in plants. *Biochim. Biophys. Acta* 1769, 330–345. doi: 10.1016/j.bbaexp.2006.12.003
- Johnston, J. W., Gunaseelan, K., Pidakala, P., Wang, M., and Schaffer, R. J. (2009). Co-ordination of early and late ripening events in apples is regulated through differential sensitivities to ethylene. *J. Exp. Bot.* 60, 2689–2699. doi: 10.1093/jxb/erp122
- Jung, C., Pillen, K., Staiger, D., Coupland, G., and von Korff, M. (2017). Editorial: recent advances in flowering time control. *Front. Plant Sci.* 7:2011. doi: 10.3389/fpls.2016.02011
- Karlova, R., Chapman, N., David, K., Angenent, G. C., Seymour, G. B., and de Maagd, R. A. (2014). Transcriptional control of fleshy fruit development and ripening. *J. Exp. Bot.* 65, 4527–4541. doi: 10.1093/jxb/eru316
- Kenis, K., Keulemans, J., and Davey, M. W. (2008). Identification and stability of QTLs for fruit quality traits in apple. *Tree Genet. Genomes* 4, 647–661. doi: 10.1007/s11295-008-0140-6
- Korte, A., and Farlow, A. (2013). The advantages and limitations of trait analysis with GWAS: a review. *Plant Methods* 9:29. doi: 10.1186/1746-4811-9-29
- Kou, X., Liu, C., Han, L., Wang, S., and Xue, Z. (2016). NAC transcription factors play an important role in ethylene biosynthesis, reception and signaling of tomato fruit ripening. *Mol. Genet. Genomics* 291, 1205–1217. doi: 10.1007/s00438-016-1177-0
- Kunihisa, M., Moriya, S., Abe, K., Okada, K., Haji, T., Hayashi, T., et al. (2014). Identification of QTLs for fruit quality traits in Japanese apples: QTLs for early ripening are tightly related to preharvest fruit drop. *Breed. Sci.* 64, 240–251. doi: 10.1270/jsbbs.64.240
- Lasky, J. R., Upadhyaya, H. D., Ramu, P., Deshpande, S., Hash, C. T., Bonnette, J., et al. (2015). Genome-environment associations in sorghum landraces predict adaptive traits. *Sci. Adv.* 1, 1–13. doi: 10.1126/sciadv.1400218
- Laurens, F., Aranzana, M. J., Arús, P., Bassi, D., Bonany, J., Corelli, L., et al. (2012). Review of fruit genetics and breeding programmes and a new European initiative to increase fruit breeding efficiency. *Acta Hortic.* 929, 95–102. doi: 10.17660/ActaHortic.2012.929.12
- Leforestier, D., Ravon, E., Muranty, H., Cornille, A., Lemaire, C., Giraud, T., et al. (2015). Genomic basis of the differences between cider and dessert apple varieties. *Evol. Appl.* 8, 650–661. doi: 10.1111/eva.12270
- Lê, S., Josse, J., and Husson, F. (2008). FactoMineR: an R package for multivariate analysis. *J. Stat. Soft.* 25, 1–18. doi: 10.18637/jss.v025.i01
- Li, X., Yin, X., Wang, H., Li, J., Guo, C., Gao, H., et al. (2015). Genome-wide identification and analysis of the apple (*Malus × domestica* Borkh.) TIFY gene family. *Tree Genet. Genomes* 11, 1–13. doi: 10.1007/s11295-014-0808-z
- Licausi, F., Giorgi, F. M., Zenoni, S., Ost, F., Pezzotti, M., and Perata, P. (2010). Genomic and transcriptomic analysis of the AP2/ERF superfamily in *Vitis vinifera*. *BMC Genomics* 11:719. doi: 10.1186/1471-2164-11-719
- Liebhart, R., Kellerhals, M., Pfammatter, W., Jertmini, M., and Gessler, C. (2003). Mapping quantitative physiological traits in apple (*Malus × domestica* Borkh.). *Plant Mol. Biol.* 52, 511–526. doi: 10.1023/A:1024886500979
- Lv, J., Rao, J., Johnson, F., Shin, S., and Zhu, Y. (2015). Genome-wide identification of jasmonate biosynthetic genes and characterization of their expression profiles during apple (*Malus × domestica*) fruit maturation. *Plant Growth Reg.* 75, 355–364. doi: 10.1007/s10725-014-9958-0
- Maher, B. (2008). Personal genomes: the case of the missing heritability. *Nature* 456, 18–21. doi: 10.1038/456018a
- Mangin, B., Siberchicot, A., Nicolas, S., Doligez, A., This, P., and Cierco-Ayrolles, C. (2012). Novel measures of linkage disequilibrium that correct the bias due to population structure and relatedness. *Heredity* 108, 285–291. doi: 10.1038/hdy.2011.73
- Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorf, L. A., Hunter, D. J., et al. (2009). Finding the missing heritability of complex diseases. *Nature* 461, 747–753. doi: 10.1038/nature08494
- Mariette, S., Tai, F. W. J., Roch, G., Barre, A., Chague, A., Decroocq, S., et al. (2016). Genome-wide association links candidate genes to resistance to *Plum pox virus* in apricot (*Prunus armeniaca*). *New Phytol.* 209, 773–784. doi: 10.1111/nph.13627
- Mazzitelli, L., Hancock, R. D., Haupt, S., Walker, P. G., Pont, S. D., McNicol, J., et al. (2007). Co-ordinated gene expression during phases of dormancy release in raspberry (*Rubus idaeus* L.) buds. *J. Exp. Bot.* 58, 1035–1045. doi: 10.1093/jxb/erl266

- Meuwissen, T., Hayes, B., and Goddard, M. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157, 1819–1829.
- Micheletti, D., Dettori, M. T., Micali, S., Aramini, V., Pacheco, I., Linge, C. D. S., et al. (2015). Whole-genome analysis of diversity and SNP-major gene association in peach germplasm. *PLoS ONE* 10:e0136803. doi: 10.1371/journal.pone.0136803
- Migicovsky, Z., Gardner, K. M., Sawler, J., Money, D., Bloom, J. S., Zhong, G. Y., et al. (2016). Genome to phenome mapping in apple using historical data. *Plant Genome* 9, 1–15. doi: 10.3835/plantgenome2015.11.0113
- Muranty, H., Troggo, M., Sadok, I. B., Rifaï, M. A., Auwerkerken, A., Banchi, E., et al. (2015). Accuracy and responses of genomic selection on key traits in apple breeding. *Hortic. Res.* 2:15060. doi: 10.1038/hortres.2015.60
- Myles, S., Peiffer, J., Brown, P. J., Ersoz, E. S., Zhang, Z., Costich, D. E., et al. (2009). Association mapping: critical considerations shift from genotyping to experimental design. *Plant Cell* 21, 2194–2202. doi: 10.1105/tpc.109.068437
- Neale, D. B., and Savolainen, O. (2004). Association genetics of complex traits in conifers. *Trends Plant Sci.* 9, 325–330. doi: 10.1016/j.tplants.2004.05.006
- Nicolas, S. D., Péros, J. P., Lacombe, T., Launay, A., Le Paslier, M. C., Bérard, A., et al. (2016). Genetic diversity, linkage disequilibrium and power of a large grapevine (*Vitis vinifera* L.) diversity panel newly designed for association studies. *BMC Plant Biol.* 16:74. doi: 10.1186/s12870-016-0754-z
- Nieuwenhuizen, N. J., Chen, X., Wang, M. Y., Matich, A. J., Perez, R. L., Allan, A. C., et al. (2015). Natural variation in monoterpene synthesis in kiwifruit: transcriptional regulation of terpene synthases by NAC and ETHYLENE-INSENSITIVE3-like transcription factors. *Plant Physiol.* 167, 1243–1258. doi: 10.1104/pp.114.254367
- Ning, Y. Q., Ma, Z. Y., Huang, H. W., Mo, H., Zhao, T. T., Li, L., et al. (2015). Two novel NAC transcription factors regulate gene expression and flowering time by associating with the histone demethylase JM14. *Nucleic Acids Res.* 43, 1469–1484. doi: 10.1093/nar/gku1382
- Nishitani, C., Hirai, N., Komori, S., Wada, M., Okada, K., Osakabe, K., et al. (2016). Efficient genome editing in apple using a CRISPR/Cas9 system. *Sci. Rep.* 6:31481. doi: 10.1038/srep31481
- Pascual, L., Albert, E., Sauvage, C., Duangjit, J., Bouchet, J. P., Bitton, F., et al. (2016). Dissecting quantitative trait variation in the resequencing era: complementarity of bi-parental, multi-parental and association panels. *Plant Sci.* 242, 120–130. doi: 10.1016/j.plantsci.2015.06.017
- Pirana, R., Eduardo, I., Pacheco, I., Da Silva, L. C., Miculan, M., Verde, I., et al. (2013). Fine mapping and identification of a candidate gene for a major locus controlling maturity date in peach. *BMC Plant Biol.* 13:166. doi: 10.1186/1471-2229-13-166
- Porto, D. D., Bruneau, M., Perini, P., Anzanello, R., Renou, J. P., Pessoa dos Santos, H., et al. (2015). Transcription profiling of the chilling requirement for bud break in apples: a putative role for *FLC*-like genes. *J. Exp. Bot.* 66, 2659–2672. doi: 10.1093/jxb/erv061
- Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., et al. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81, 559–575. doi: 10.1086/519795
- R Core Team, A. (2014). *R: A Language and Environment for Statistical Computing*. [Internet]. Vienna: R Foundation for Statistical Computing. Available online at: <http://www.Rproject.org/>
- Ríos, G., Leida, C., Conejero, A., and Badenes, M. L. (2014). Epigenetic regulation of bud dormancy events in perennial plants. *Front. Plant Sci.* 5:247. doi: 10.3389/fpls.2014.00247
- Saito, T., Bai, S., Ito, A., Sakamoto, D., Saito, T., Ubi, B. E., et al. (2013). Expression and genomic structure of the dormancy-associated MADS box genes MADS13 in Japanese pears (*Pyrus pyrifolia* Nakai) that differ in their chilling requirement for endodormancy release. *Tree Physiol.* 33, 654–667. doi: 10.1093/treephys/tp037
- Sánchez-Pérez, R., Del Cuetto, J., Dicenta, F., and Martínez-Gómez, P. (2014). Recent advancements to study flowering time in almond and other *Prunus* species. *Front. Plant Sci.* 5:334. doi: 10.3389/fpls.2014.00334
- Sauvage, C., Segura, V., Bauchet, G., Stevens, R., Do, P. T., Nikoloski, Z., et al. (2014). Genome-Wide Association in tomato reveals 44 candidate loci for fruit metabolic traits. *Plant Physiol.* 165, 1120–1132. doi: 10.1104/pp.114.241521
- Segura, V., Vilhjálmsson, B. J., Platt, A., Korte, A., Seren, Ü., Long, Q., et al. (2012). An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nat. Genet.* 44, 825–830. doi: 10.1038/ng.2314
- Stranger, B. E., Stahl, E. A., and Raj, T. (2011). Progress and promise of genome-wide association studies for human complex trait genetics. *Genetics* 187, 367–383. doi: 10.1534/genetics.110.120907
- Tadiello, A., Longhi, S., Moretto, M., Ferrarini, A., Tononi, P., Farneti, B., et al. (2016). Interference with ethylene perception at receptor level sheds light on auxin and transcriptional circuits associated with the climacteric ripening of apple fruit (*Malus x domestica* Borkh.). *Plant J.* 88, 963–975. doi: 10.1111/tpj.13306
- Trainin, T., Zohar, M., Shimoni-Shor, E., Doron-Faigenboim, A., Bar-Ya'akov, I., Hatib, K., et al. (2016). A unique haplotype found in apple accessions exhibiting early bud-break could serve as a marker for breeding apples with low chilling requirements. *Mol. Breed.* 36:158. doi: 10.1007/s11032-016-0575-7
- Ubi, B. E., Sakamoto, D., Ban, Y., Shimada, T., Ito, A., Nakajima, I., et al. (2010). Molecular cloning of dormancy associated MADS-box gene homologs and their characterization during seasonal endodormancy transitional phases of Japanese pear. *J. Am. Soc. Hort. Sci.* 135, 174–182.
- Urrestarazu, J., Denancé, C., Ravon, E., Guyader, A., Guisnel, R., Feugey, L., et al. (2016). Analysis of the genetic diversity and structure across a wide range of germplasm reveals prominent gene flow in apple at the European level. *BMC Plant Biol.* 16:130. doi: 10.1186/s12870-016-0818-0
- Verde, I., Bassil, N., Scalabrin, S., Gilmore, B., Lawley, C. T., Gasic, K., et al. (2012). Development and evaluation of a 9K SNP array for peach by internationally coordinated SNP detection and validation in breeding germplasm. *PLoS ONE* 7:e35668. doi: 10.1371/journal.pone.0035668
- Verde, I., Jenkins, J., Dondini, L., Micali, S., Pagliarini, G., Vendramin, E., et al. (2017). The Peach v2.0 release: high-resolution linkage mapping and deep resequencing improve chromosome-scale assembly and contiguity. *BMC Genomics* 18:225. doi: 10.1186/s12864-017-3606-9
- Visscher, P. M., Yang, J., and Goddard, M. E. (2010). A commentary on ‘common SNPs explain a large proportion of the heritability for human height’ by Yang et al. (2010). *Twin Res. Hum. Genet.* 13, 517–524. doi: 10.1375/twin.13.6.517
- Vitasse, Y., Lenz, A., and Körner, C. (2014). The interaction between freezing tolerance and phenology in temperate deciduous trees. *Front. Plant Sci.* 5:541. doi: 10.3389/fpls.2014.00541
- Wegrzyn, J. L., Eckert, A. J., Choi, M., Lee, J. M., Stanton, B. J., Sykes, R., et al. (2010). Association genetics of traits controlling lignin and cellulose biosynthesis in black cottonwood (*Populus trichocarpa*, *Salicaceae*) secondary xylem. *New Phytol.* 188, 15–32. doi: 10.1111/j.1469-8137.2010.03415.x
- Wilczek, A. M., Roe, J. L., Knapp, M. C., Cooper, M. D., Lopez-Gallego, C., Martin, L. J., et al. (2009). Effects of genetic perturbation on seasonal life history plasticity. *Science* 323, 930–934. doi: 10.1126/science.1165826
- Wood, A. R., Hernandez, D. G., Nalls, M. A., Yaghootkar, H., Gibbs, J. R., Harries, L. W., et al. (2011). Allelic heterogeneity and more detailed analyses of known loci explain additional phenotypic variation and reveal complex patterns of association. *Hum. Mol. Genet.* 20, 4082–4092. doi: 10.1093/hmg/ddr328
- Wu, R. M., Walton, E. F., Richardson, A. C., Wood, M., Hellens, R. P., and Varkonyi-Gasic, E. (2012). Conservation and divergence of four kiwifruit SVP-like MADSbox genes suggest distinct roles in kiwifruit bud dormancy and flowering. *J. Exp. Bot.* 63, 797–807. doi: 10.1093/jxb/err304
- Xie, X. I., Yin, X. R., and Chen, K. S. (2016). Roles of APETALA2/Ethylene-Response factors in regulation of fruit quality. *Crit. Rev. Plant Sci.* 35, 120–130. doi: 10.1080/07352689.2016.1213119
- Xu, H., and Guan, Y. (2014). Detecting local haplotype sharing and haplotype association. *Genetics* 197, 823–838. doi: 10.1534/genetics.114.164814
- Yamane, H., Ooka, T., Jotatsu, H., Hosaka, Y., Sasaki, R., and Tao, R. (2011). Expressional regulation of PpDAM5 and PpDAM6, peach (*Prunus persica*) dormancy-associated MADS-box genes, by low temperature and dormancy breaking reagent treatment. *J. Exp. Bot.* 62, 3481–3488. doi: 10.1093/jxb/err028
- Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., et al. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* 42, 565–571. doi: 10.1038/ng.608
- Yano, K., Yamamoto, E., Aya, K., Takeuchi, H., Lo, P. C., Hu, L., et al. (2016). Genome-wide association study using whole-genome sequencing rapidly



- identifies new genes influencing agronomic traits in rice. *Nat. Genet.* 48, 927–934. doi: 10.1038/ng.3596
- Yoo, S. Y., Kim, Y., Kim, S. Y., Lee, J. S., and Ahn, J. H. (2007). Control of flowering time and cold response by a NAC-domain protein in *Arabidopsis*. *PLoS ONE* 2:e642. doi: 10.1371/journal.pone.0000642
- Zhebentyayeva, T. N., Fan, S., Chandra, A., Bielenberg, D. G., Reighard, G. L., Okie, W. R., et al. (2014). Dissection of chilling requirement and bloom date QTLs in peach using a whole genome sequencing of sibling trees from an F2 mapping population. *Tree Genet. Genomes* 10, 35–51. doi: 10.1007/s11295-013-0660-6
- Zhou, X., and Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* 44, 821–824. doi: 10.1038/ng.2310
- Zhu, C. S., Gore, M., Buckler, E. S., and Yu, J. M. (2008). Status and prospects of association mapping in plants. *Plant Genome* 2, 121–133. doi: 10.3835/plantgenome2008.02.0089
- Zhu, M., Chen, G., Zhou, S., Tu, Y., Wang, Y., Dong, T., et al. (2014). A new tomato NAC (NAM/ATAF1/2/CUC2) transcription factor, SINAC4, functions as a positive regulator of fruit ripening and carotenoid accumulation. *Plant Cell Physiol.* 55, 119–135. doi: 10.1093/pcp/pct162
- Zuk, O., Hechter, E., Sunyaev, S. R., and Lander, E. S. (2012). The mystery of missing heritability: genetic interactions create phantom heritability. *Proc. Natl. Acad. Sci. U.S.A.* 109, 1193–1198. doi: 10.1073/pnas.1119675109

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer AMA and handling editor declared their shared affiliation.

Copyright © 2017 Urrestarazu, Muranty, Denancé, Leforestier, Ravon, Guyader, Guisnel, Feugey, Aubourg, Celton, Daccord, Dondini, Gregori, Lateur, Houben, Ordidge, Paprstein, Sedlak, Nybom, Garkava-Gustavsson, Troglio, Bianco, Velasco, Poncet, Théron, Moriya, Bink, Laurens, Tartarini and Durel. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

**Titre :** Analyse bioinformatique du génome et de l'épigénome du pommier

**Mots clés :** assemblage de génome, annotation de genes, épigénétique, méthylation différentielle

**Résumé :** La pomme est l'un des fruits les plus consommés au monde. En utilisant les dernières technologies de séquençage (PacBio) et de cartes optiques (BioNano), nous avons généré un assemblage *de novo* de haute qualité du génome du pommier (*Malus domestica Borkh.*). Nous avons réalisé une annotation des gènes et des éléments transposables pour permettre à cet assemblage d'être utilisé en tant que génome de référence. La grande contiguité de l'assemblage a permis de détecter les éléments transposables de façon exhaustive, ce qui fournit une opportunité sans précédents d'étudier les régions non-caractérisées d'un génome d'arbre. Nous avons également trouvé que le génome du pommier est entièrement dupliqué, comme montré par les relations de syntenie entre les chromosomes.

En utilisant du Whole Genome Bisulfite Sequencing (WGBS) ainsi que l'assemblage précédemment généré, nous avons montré des cartes de méthylation de l'ADN pour tout le génome et montré une corrélation générale entre la méthylation de l'ADN près des gènes et l'expression des gènes. De plus, nous avons identifié plusieurs Régions Différentiellement Méthylées (RDMs) entre les méthylomes de fruits et de feuilles du pommier, associées à des gènes candidats qui pourraient être impliqués dans des traits agronomiques importants tel que le développement du fruit. Enfin, nous avons développé un pipeline rapide, simple et complet qui prend entièrement en charge l'analyse des données WGBS, de l'alignement des reads au calcul des RDMs.

**Title :** Bioinformatic analysis of the apple genome and epigenome

**Keywords :** genome assembly, gene annotation, epigenetics, differential methylation

**Abstract :** Apple is one of the most consumed fruits in the world. Using the latest sequencing (PacBio) and optical mapping (BioNano) technologies, we have generated a high-quality *de novo* assembly of the apple (*Malus domestica Borkh.*) genome. We performed a gene annotation as well as a transposable element annotation to allow this assembly to be used as a reference genome. The high-contiguity of the assembly allowed to exhaustively detect the transposable elements, which represented over half the assembly, thus providing an unprecedented opportunity to investigate the uncharacterized regions of a tree genome. We also found that the apple genome is entirely duplicated as showed by the syntenic links between chromosomes.

Using Whole Genome Bisulfite Sequencing (WGBS) and the previously generated assembly, we produced genome-wide DNA methylation maps and showed a general correlation between DNA methylation next to genes and gene expression. Moreover, we identified several Differentially Methylated Regions (DMRs) between apple fruits and leaf methylomes associated to candidate genes that could be involved in agronomically relevant traits such as apple fruit development. Finally, we developed a complete and easy-to-use pipeline which aim is to handle the complete treatment of WGBS data, from the reads mapping to the DMRs computing. It can handle datasets having a low number of biological replicates.