

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

Nicolas Daccord, Jean-Marc Celton, G. Linsmith, C. Becker, Nathalie Choisne, E. Schijlen, H. van de Geest, L. Bianco, D. Micheletti, R. Velasco, et al.

▶ **To cite this version:**

Nicolas Daccord, Jean-Marc Celton, G. Linsmith, C. Becker, Nathalie Choisne, et al.. High-quality de novo assembly of the apple genome and methylome dynamics of early fruit development. Nature Genetics, 2017, 49 (7), pp.1099-1108. 10.1038/ng.3886 . hal-01602554

HAL Id: hal-01602554
https://hal.science/hal-01602554

Submitted on 26 May 2020

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.



Distributed under a Creative Commons Attribution 4.0 International License

nature genetics
OPEN

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

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

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.

© 2017 Nature America, Inc., part of Springer Nature. All rights reserved.

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[1]. In recent years, however, evidence supporting the importance of TEs in genome evolution, genome structure, regulation of gene expression and epigenetics has been mounting[2–5]. 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[6] but remained incomplete with inaccurate contig positions[7]; 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[8], 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[9–14]. 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[15].

In addition to DNA sequence modifications, it has been shown that epigenetic variations contribute to genome accessibility, functionality and structure[16,17]. 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[18–20].

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[21,22]. 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

[1]Institut de Recherche en Horticulture et Semences (IRHS), Université d'Angers, INRA, AGROCAMPUS-Ouest, SFR4207 QUASAV, Université Bretagne Loire, Angers, France. [2]Research and Innovation Center, Department of Genomics and Biology of Fruit Crops, Fondazione E Mach di San Michele all'Adige, Italy. [3]Department of Molecular Biology, Max Planck Institute for Developmental Biology, Tübingen, Germany. [4]UR1164 URGI, Research Unit in Genomics-Info, INRA, Université Paris-Saclay, Versailles, France. [5]Wageningen UR–Bioscience, Wageningen, the Netherlands. [6]Department of Biosciences, University of Milan, Milan, Italy. [7]LIPM, Université de Toulouse, INRA, CNRS, Castanet-Tolosan, France. [8]Agricultural Research Council, Biotechnology Platform, Onderstepoort, Pretoria, South Africa. [9]Wageningen UR–Plant Breeding, Wageningen, the Netherlands. [10]Present address: Gregor Mendel Institute of Molecular Plant Biology, Austrian Academy of Sciences, Vienna Biocenter (VBC), Vienna, Austria. [11]These authors contributed equally to this work. Correspondence should be addressed to E.B. (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[23]. 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[25], and the resulting contigs were combined with the PacBio reads using the DBG2OLC assembler[26]. 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 ≥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[23], 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

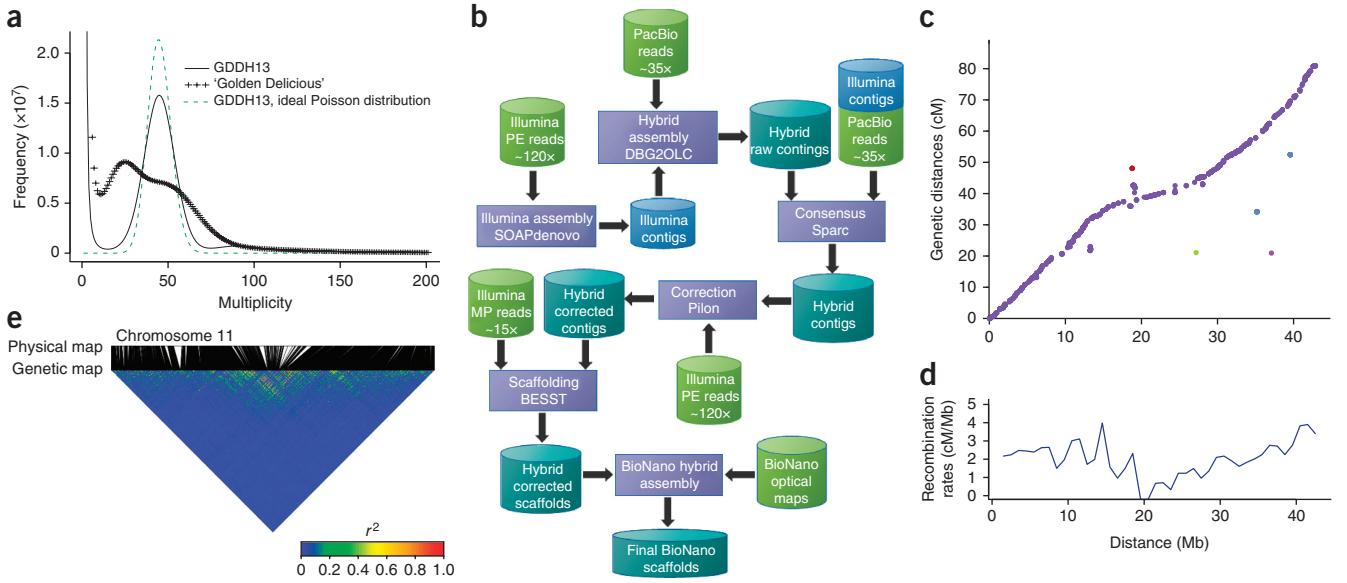


Figure 1 Assembly and validation of the GDDH13 doubled-haploid apple genome. (**a**) k-mer (23 bp) spectra of the doubled-haploid GDDH13 and the heterozygous Golden Delicious[33] genomes. The *x* axis represents k-mer multiplicity, and the *y* axis represents the number of k-mers with a given multiplicity 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).

© 2017 Nature America, Inc., part of Springer Nature. All rights reserved.

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

	GDDH[13]	Li *et al.*[33]	Velasco *et al.*[6]
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[27,28]. 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[29], we generated a haplotype map for GDDH13, which allowed the identification of recombination breakpoints (**Supplementary Fig. 5**).

Finally, the completeness of the assembly was tested by searching for 248 core eukaryotic genes[30] (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[31].

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[32]. 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[6,33] was performed using the BUSCO v2 method, which is based on a benchmark of 1,440 conserved plant genes[34]. The results indicate that our apple genome annotation is the most complete, despite having the lowest number of predicted genes (**Table 1**).

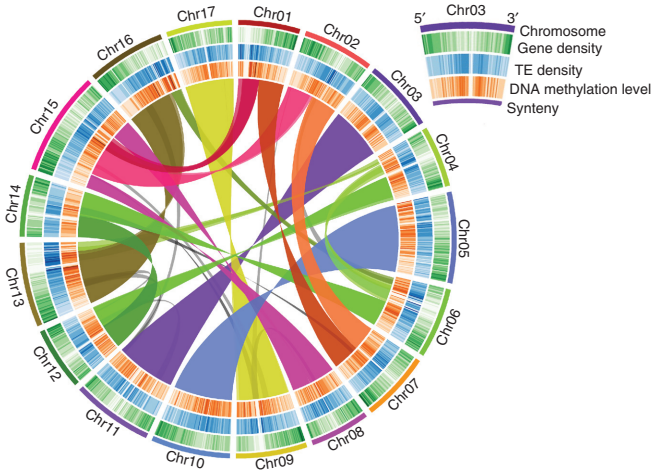


Figure 2 Synteny and distribution of genomic and epigenomic features of the apple genome. The rings indicate (from outside to inside, as indicated in the inset) chromosomes (Chr), heat maps representing gene density (green), TE density (blue) and DNA methylation levels (orange). A map connecting homologous 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[35] 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.genomevolution.org) and visualized with Circos[36]. Results of this analysis (**Fig. 2**) showed an even clearer genome duplication pattern than has previously been reported[6]. 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[37]) 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**.

© 2017 Nature America, Inc., part of Springer Nature. All rights reserved.

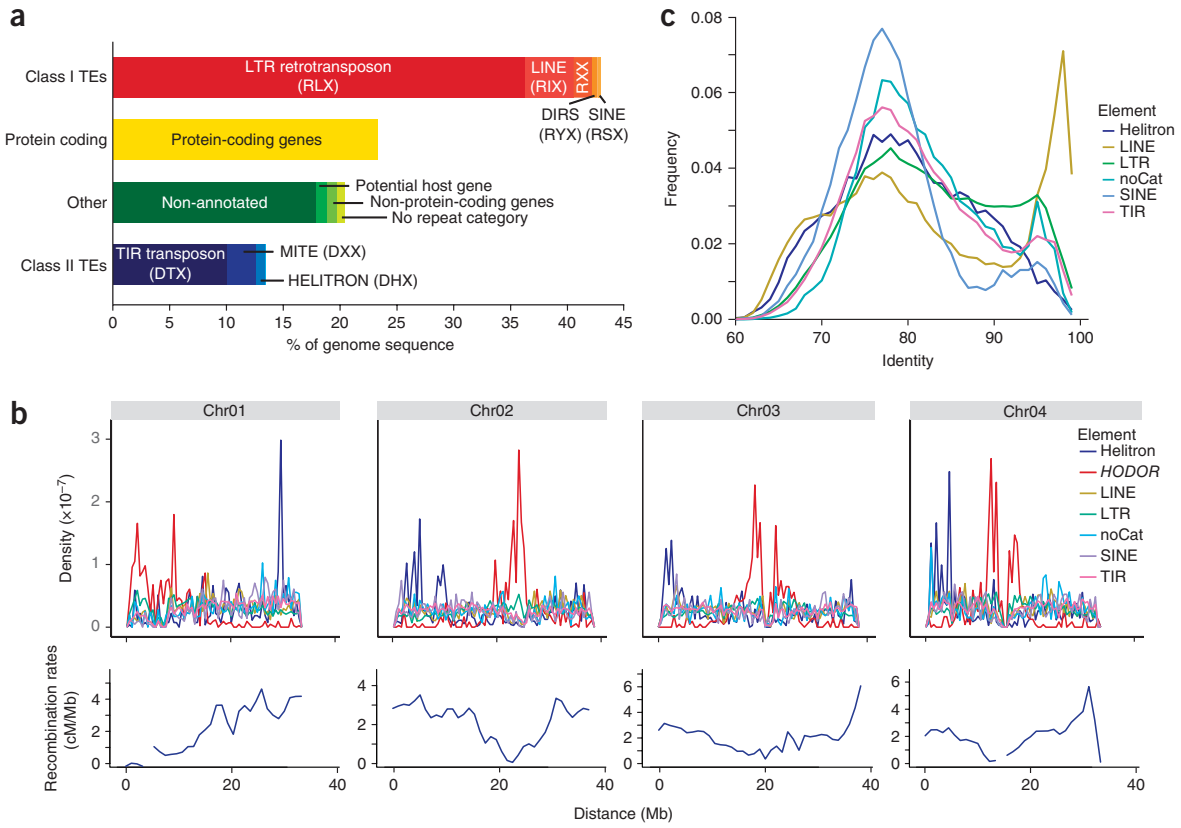


Figure 3 Distribution and evolution of transposable elements in the apple genome. (**a**) Percentage of base pairs of the assembled GDDH13 genome that represent genes, pseudo-genes, TEs and non-annotated regions. Retrotransposons (class I) are shown in shades of red, and DNA transposons (class II) are 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[38] 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[39], 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 ≤ 1×10^{-29}) similarities with sequences encoding three short bacterial proteins of unknown function (**Supplementary Fig. 8a**), and mining of transcriptome data[35] 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 (1.3×10^{-8} base substitutions per site per year[40,41]), we estimated the age of those insertions as described by the International Human Genome Sequencing Consortium[42]. 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 ~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[43,44].

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.

© 2017 Nature America, Inc., part of Springer Nature. All rights reserved.

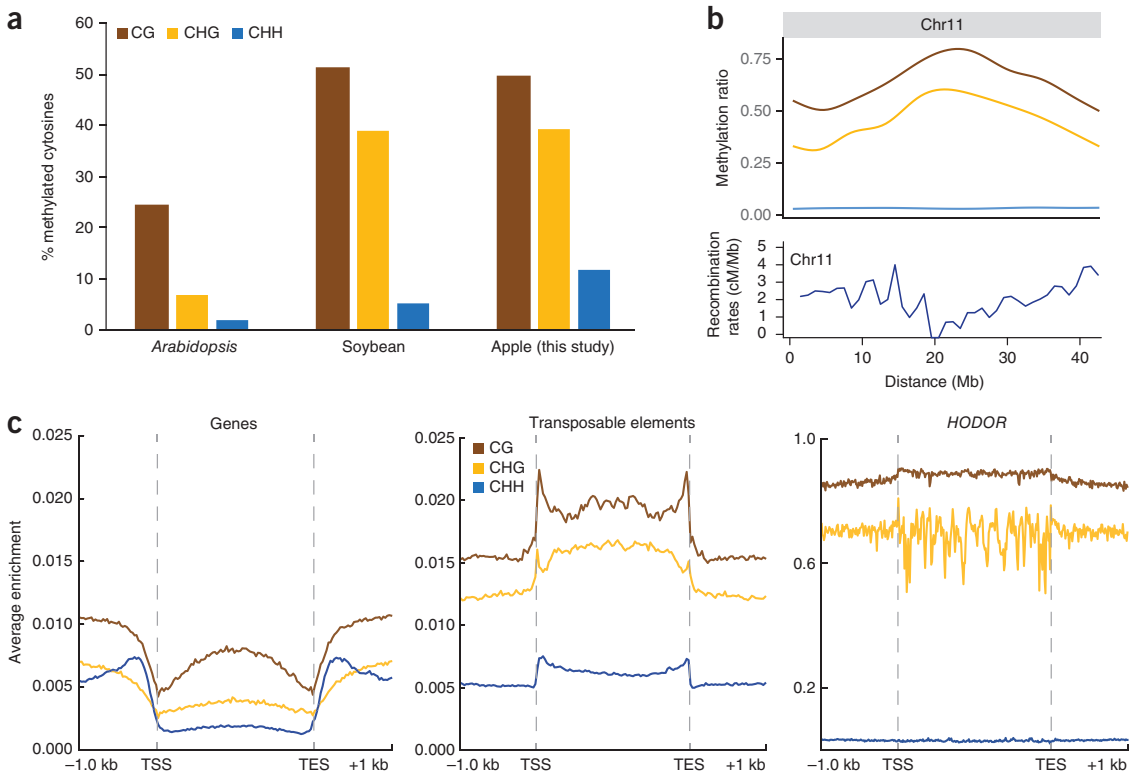


Figure 4 DNA methylation landscape of the GDDH13 genome. (**a**) Percentage of DNA methylation distributions of the three methylation contexts (CG, CHG or CHH) in *Arabidopsis*[44], soybean[60] and apple. For apple, the percentages were estimated based on the number of cytosines that had a methylation ratio ≥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[45,46], 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[35], 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**).

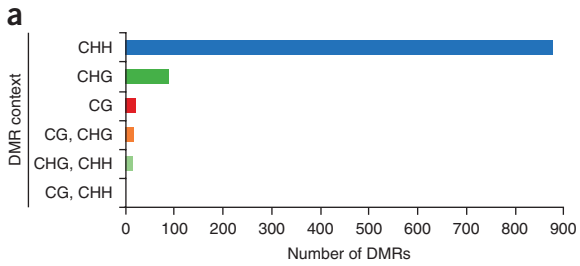
DNA methylation and fruit development

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[47]. 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

© 2017 Nature America, Inc., part of Springer Nature. All rights reserved.



b

GeneID	DMR context	*Arabidopsis* annotation	Putative function
MD01G0162300	CHH	Sugar isomerase (SIS) domain-containing protein	Carbohydrate metabolism
MD13G0052100	CHH	AP1 (APETALA1); DNA binding / transcription factor	Development
MD06G0169300	CHH	AGL8 (AGAMOUS-LIKE 8), FUL (FRUITFUL)	Development
MD02G0223400	CHH	SUVH1 (SU(VAR)3-9 HOMOLOG 1)	Epigenetic regulation
MD06G0019200	CHH	CHROMOMETHYLASE 2 (CMT2)	Epigenetic regulation
MD01G0157100	CHH	SEP1, AGL2 (AGAMOUS LIKE-2); DNA binding / transcription factor	Flower/ovule development
MD15G0009100	CHH	CLV1 (CLAVATA 1); ATP binding / kinase/ protein serine/threonine kinase	Fruit development
MD03G0074700	CHH	ATEXPA4 (ARABIDOPSIS THALIANA EXPANSIN A4)	Fruit development and ripening
MD14G0002100	CHH	Short-chain dehydrogenase/reductase (SDR) family protein	Fruit ripening
MD03G0017900	CHH	EIL3 (ETHYLENE-INSENSITIVE3-LIKE3); transcription factor	Fruit ripening
MD03G0223600	CHH	ATGID1B/GID1B (GA INSENSITIVE DWARF1B)	Fruit set, fruit growth



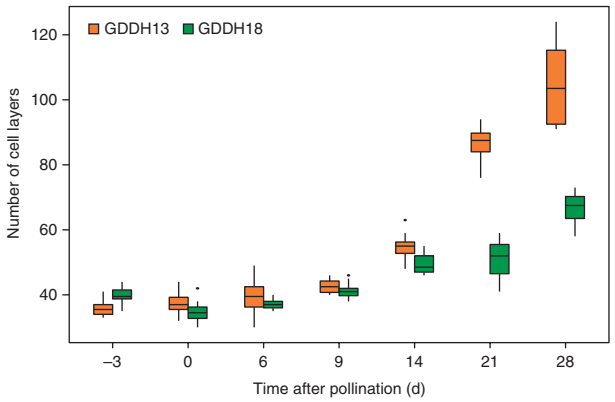


Figure 5 Differentially methylated regions between apple tree leaves and young fruits. (**a**) DMR content in samples of GDDH13 leaves and young fruits (CHH, $n = 875$ DMRs; CHG, $n = 88$ DMRs; CG, $n = 21$ DMRs; CG and CHG, $n = 17$ DMRs; CHG and CHH, $n = 14$ DMRs; CG and CHH, $n = 2$ DMRs). Most of 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[33]. Following hybrid assembly of PacBio and Illumina reads, Li and colleagues[33] 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[48], 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[23] 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

© 2017 Nature America, Inc., part of Springer Nature. All rights reserved.

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[49] and the 45,293 genes that were identified after filtering out overlapping genes from the original apple genome annotation[49] (**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[50]. 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[48] and an uplift event occurring at the Tian Shan mountains[51], which cover the region where the ancestor of the apple originates from[52]. 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[53] 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*[43], 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[54]. Consistent with this observation, it has been shown for *Arabidopsis* that cell-type-specific DNA methylation differences mainly occur at CHH sites[55]. 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[56–58].

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[59]. 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/nuccore/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. Giraud 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;

© 2017 Nature America, Inc., part of Springer Nature. All rights reserved.

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/.

© 2017 Nature America, Inc., part of Springer Nature. All rights reserved.

1. Veeckman, E., Ruttink, T. & Vandepoele, K. Are we there yet? Reliably estimating the completeness of plant genome sequences. *Plant Cell* **28**, 1759–1768 (2016).
2. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
3. Fedoroff, N.V. Transposable elements, epigenetics and genome evolution. *Science* **338**, 758–767 (2012).
4. 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).
5. Lisch, D. How important are transposons for plant evolution? *Nat. Rev. Genet.* **14**, 49–61 (2013).
6. Velasco, R. *et al.* The genome of the domesticated apple (*Malus* × *domestica* Borkh.). *Nat. Genet.* **42**, 833–839 (2010).
7. 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).
8. Ansorge, W.J. Next-generation DNA sequencing (II): techniques, applications. *Next Generat. Sequenc. & Applic.* **1**, 1–10 (2016).
9. Zhang, G. *et al.* Hybrid *de novo* genome assembly of the Chinese herbal plant danshen (*Salvia miltiorrhiza* Bunge). *Gigascience* **4**, 62 (2015).
10. VanBuren, R. *et al.* Single-molecule sequencing of the desiccation-tolerant grass *Oropetium thomaeum*. *Nature* **527**, 508–511 (2015).
11. 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).
12. 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).
13. Mahesh, H.B. *et al. Indica* rice genome assembly, annotation and mining of blast-disease-resistance genes. *BMC Genomics* **17**, 242 (2016).
14. 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).
15. Cui, L. *et al.* Widespread genome duplications throughout the history of flowering plants. *Genome Res.* **16**, 738–749 (2006).
16. Roudier, F., Teixeira, F.K. & Colot, V. Chromatin indexing in *Arabidopsis*: an epigenomic tale of tails and more. *Trends Genet.* **25**, 511–517 (2009).
17. He, G., Elling, A.A. & Deng, X.W. The epigenome and plant development. *Annu. Rev. Plant Biol.* **62**, 411–435 (2011).
18. Cubas, P., Vincent, C. & Coen, E. An epigenetic mutation responsible for natural variation in floral symmetry. *Nature* **401**, 157–161 (1999).
19. Becker, C. *et al.* Spontaneous epigenetic variation in the *Arabidopsis thaliana* methylome. *Nature* **480**, 245–249 (2011).
20. Ong-Abdullah, M. *et al.* Loss of *Karma* transposon methylation underlies the mantled somaclonal variant of oil palm. *Nature* **525**, 533–537 (2015).
21. 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).
22. Telias, A. *et al.* Apple skin patterning is associated with differential expression of *MYB10*. *BMC Plant Biol.* **11**, 93 (2011).
23. 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).
24. Lespinasse, Y., Bouvier, L., Djulbic, M. & Chevreau, E. Haploidy in apple and pear. *Acta Hortic.* **484**, 49–54 (1998).
25. Luo, R. *et al.* SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. *Gigascience* **1**, 18 (2012).
26. 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).
27. 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).
28. Bianco, L. *et al.* Development and validation of the Axiom Apple480K SNP genotyping array. *Plant J.* **86**, 62–74 (2016).
29. Falginella, L. *et al.* A major QTL controlling apple skin russeting maps on the linkage group 12 of 'Renetta Grigia di Torriana'. *BMC Plant Biol.* **15**, 150 (2015).
30. Parra, G., Bradnam, K. & Korf, I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23**, 1061–1067 (2007).
31. Bradnam, K.R. *et al.* Assemblathon 2: evaluating *de novo* methods of genome assembly in three vertebrate species. *Gigascience* **2**, 10 (2013).
32. Foissac, S. *et al.* Genome annotation in plants and fungi: EuGene as a model platform. *Curr. Bioinform.* **3**, 87–97 (2008).
33. Li, X. *et al.* Improved hybrid *de novo* genome assembly of domesticated apple (*Malus* × *domestica*). *Gigascience* **5**, 35 (2016).
34. 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).
35. 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).
36. Krzywinski, M. *et al.* Circos: an information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645 (2009).
37. Flutre, T., Duprat, E., Feuillet, C. & Quesneville, H. Considering transposable element diversification in *de novo* annotation approaches. *PLoS One* **6**, e16526 (2011).
38. Quesneville, H. *et al.* Combined evidence annotation of transposable elements in genome sequences. *PLOS Comput. Biol.* **1**, 166–175 (2005).
39. Jurka, J. *et al.* Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110**, 462–467 (2005).
40. 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).
41. 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).
42. Lander, E.S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
43. Lister, R. *et al.* Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. *Cell* **133**, 523–536 (2008).
44. Cokus, S.J. *et al.* Shotgun bisulphite sequencing of the *Arabidopsis* genome reveals DNA methylation patterning. *Nature* **452**, 215–219 (2008).
45. Matzke, M.A. & Mosher, R.A. RNA-directed DNA methylation: an epigenetic pathway of increasing complexity. *Nat. Rev. Genet.* **15**, 394–408 (2014).
46. Law, J.A. & Jacobsen, S.E. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nat. Rev. Genet.* **11**, 204–220 (2010).
47. Hagmann, J. *et al.* Century-scale methylome stability in a recently diverged *Arabidopsis thaliana* lineage. *PLoS Genet.* **11**, e1004920 (2015).
48. Zhang, H. *et al.* Dramatic improvement in genome assembly achieved using doubled-haploid genomes. *Sci. Rep.* **4**, 6780–6785 (2014).
49. Wu, J. *et al.* The genome of the pear (*Pyrus bretschneideri* Rehd.). *Genome Res.* **23**, 396–408 (2013).
50. Belyayev, A. Bursts of transposable elements as an evolutionary driving force. *J. Evol. Biol.* **27**, 2573–2584 (2014).
51. Balukhovsky, A.N. & Khain, V.E. *Historical Geotectonics—Mesozoic and Cenozoic* (CRC Press, 1997).
52. 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).
53. 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).
54. 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).
55. Kawakatsu, T. *et al.* Unique cell-type-specific patterns of DNA methylation in the root meristem. *Nat. Plants* **2**, 16058 (2016).
56. 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).
57. Liu, R. *et al.* A DEMETER-like DNA demethylase governs tomato fruit ripening. *Proc. Natl. Acad. Sci. USA* **112**, 10804–10809 (2015).
58. 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).
59. 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).
60. 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 SOAPdevo 2.223 (ref. 25). Next, the PacBio reads and Illumina contigs were combined to perform a hybrid assembly using the DBG2OLC pipeline[26].

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[61]. 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[63] 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[23] 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[64] 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[27], was genotyped with the Axiom Apple-480K SNP genotyping array[28]. 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[65].

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.*[47]. DNA methylation distribution plots and gene clustering analyses by methylation patterns were performed with deepTools[66].

Small RNA alignment. Apple sRNA sequences derived from mature fruit parenchyma[35] were aligned to the Golden Delicious doubled-haploid pseudo-molecules 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[67] and SOAPdenovo-trans[68]. 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[32] by combining Gmap transcript mapping[69], 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[70], RNAmmer[71] and RfamScan[72] 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[73]. The functional annotation was then completed by the prediction of targeted signals using the TargetP software[74].

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[6] were mapped to the GDDH13 genome assembly using the best BLAT[75] hit. Comparison of the two genome annotations was done using Bio++[76].

Repeat annotation. The TE*denovo* pipeline[37,77] 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[38] 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[78]. The observed sequence divergence was corrected with the Jukes and Cantor model[79]. 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 MJAX00000000.1. The raw Illumina mRNA sequences were submitted to the NCBI under BioProject ID PRJNA191060, and the GDDH18 genome reads were deposited under BioProject ID PRJNA379390. DNA methylation data can be accessed on the Gene Expression Omnibus website under accession codes GSE87014 and 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).

© 2017 Nature America, Inc., part of Springer Nature. All rights reserved.

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).

© 2017 Nature America, Inc., part of Springer Nature. All rights reserved.

doi:10.1038/ng.3886