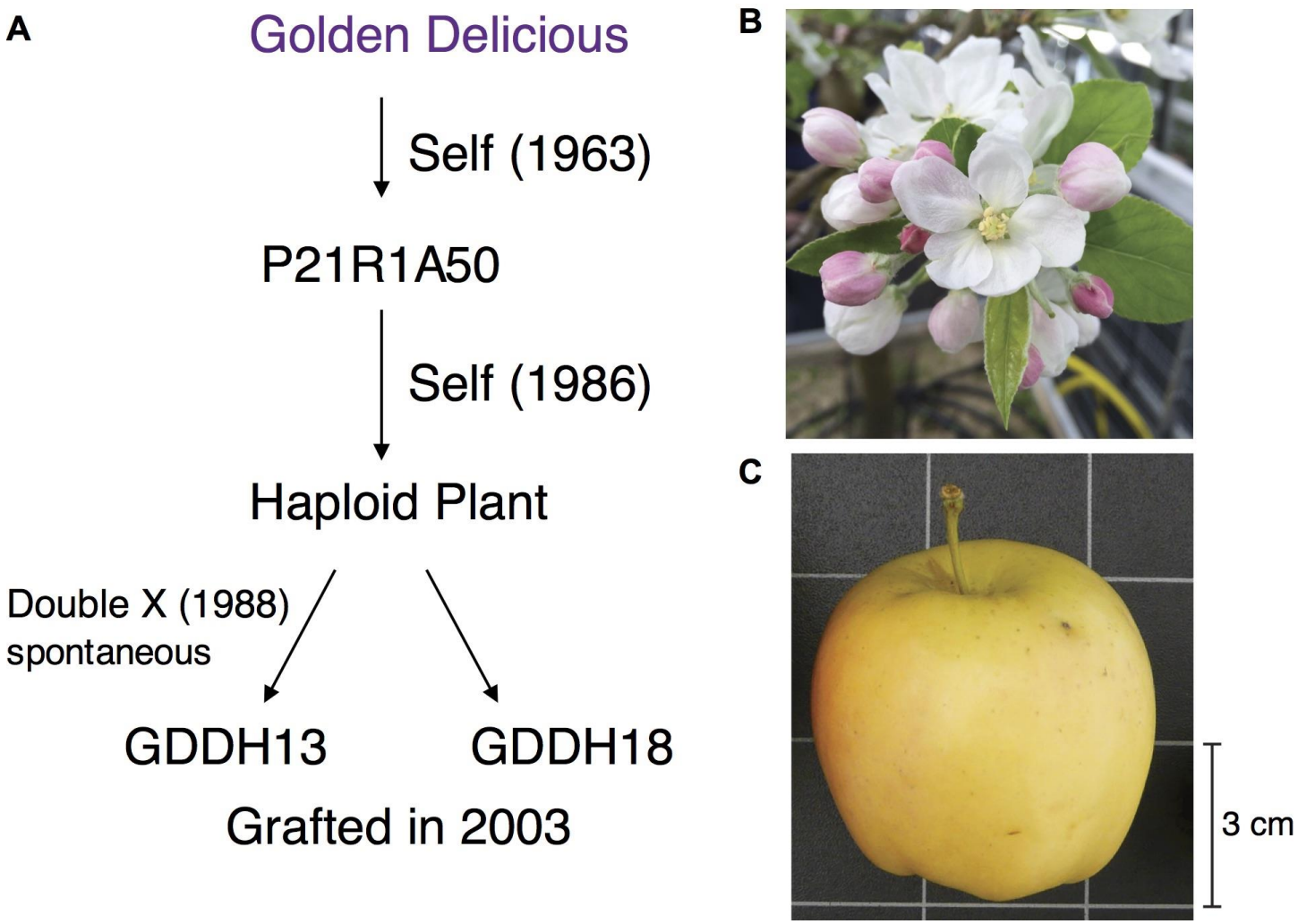
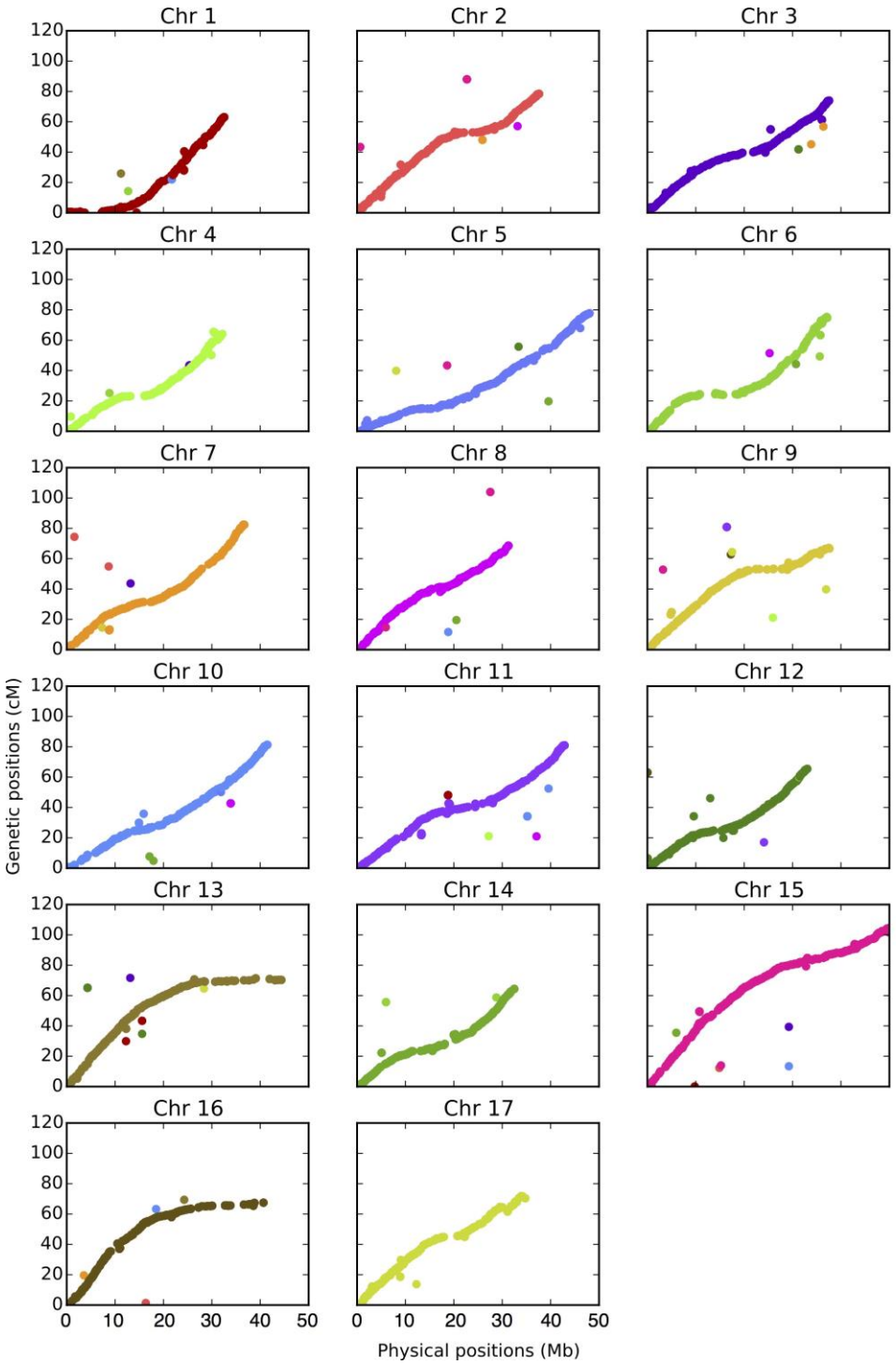


Supplementary Figure 1

Breeding the doubled-haploid GDDH13.

(A) 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. (B) and (C) show photographs of flowers and a fruit of GDDH13.

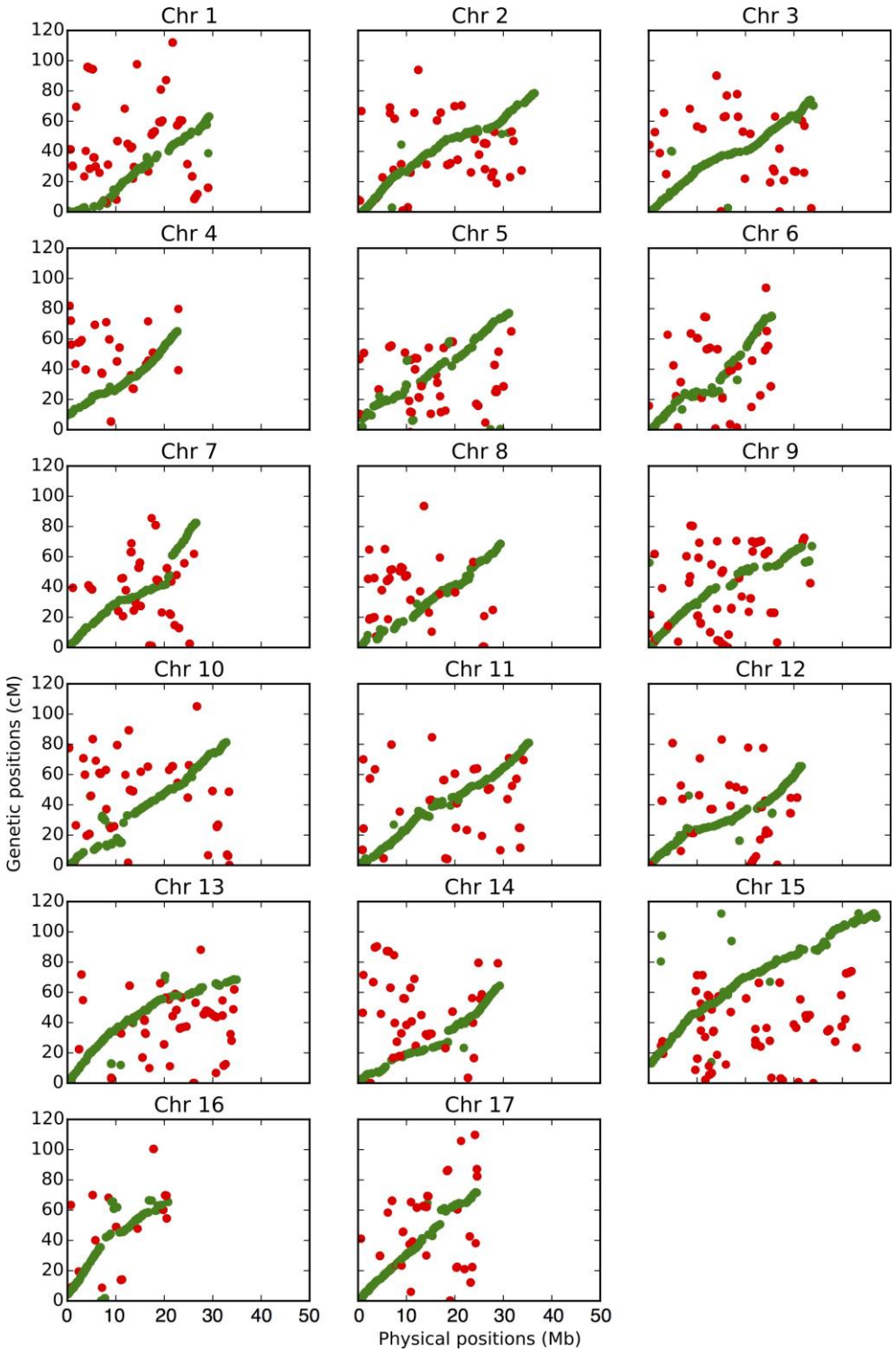
Nature Genetics: doi:10.1038/ng.3886



Supplementary Figure 2

Comparison of the genetic and physical GDDH13 position of the SNP markers of the integrated genetic map.

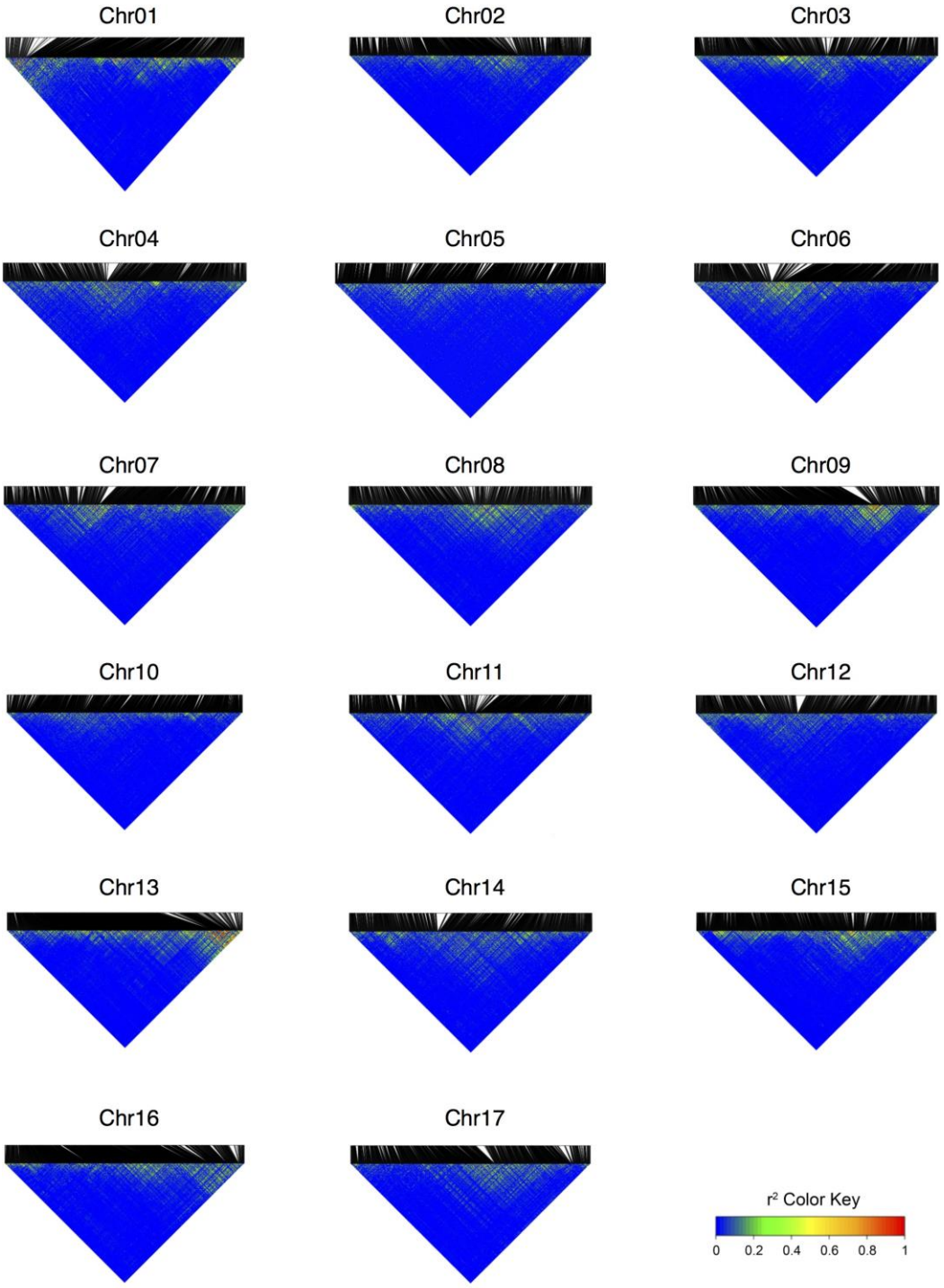
Graphical representation of the location of Single Nucleotide Polymorphism (SNP) markers on the physical map (x axis) compared to their position on the integrated genetic map (y axis) for all chromosomes. Each marker is plotted in the color of the chromosome to which it has been genetically mapped to. The colors correspond to Fig. 2 where syntenic chromosomes are depicted in similar color tones.



Supplementary Figure 3

Comparison of the genetic and physical genome v.1 position of the markers of the integrated genetic map.

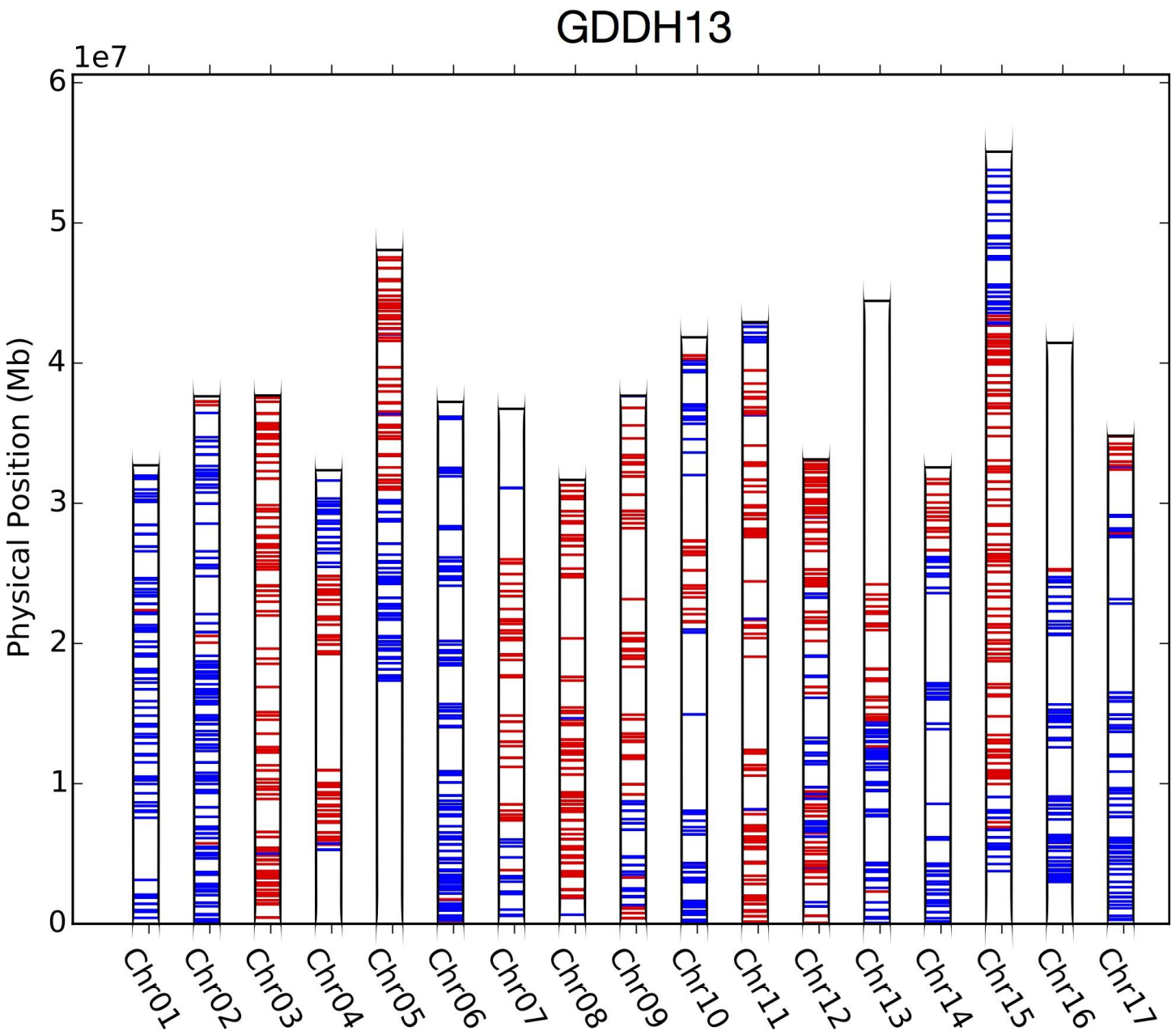
Like Suppl. Fig. 2 but using the previous genome release (v.1). All markers mapping to the correct chromosome are plotted in green, the ones to the wrong chromosome in red. This graph is shown to visualize the overall improvement that has been achieved by the new sequencing effort.



Supplementary Figure 4

Linkage disequilibrium analysis for all GDDH13 chromosomes.

Heatmap of genotypic linkage disequilibrium (LD, r^2) in all GDDH13 chromosomes in the 'Old Dessert' INRA apple core collection. The top part of the figure is a graphical representation of the location of single nucleotide polymorphism markers (SNPs) on the physical map (top) with correspondence to their order in a regular distribution (bottom). The color coding represents high LD (red) and low LD (blue).



Supplementary Figure 5

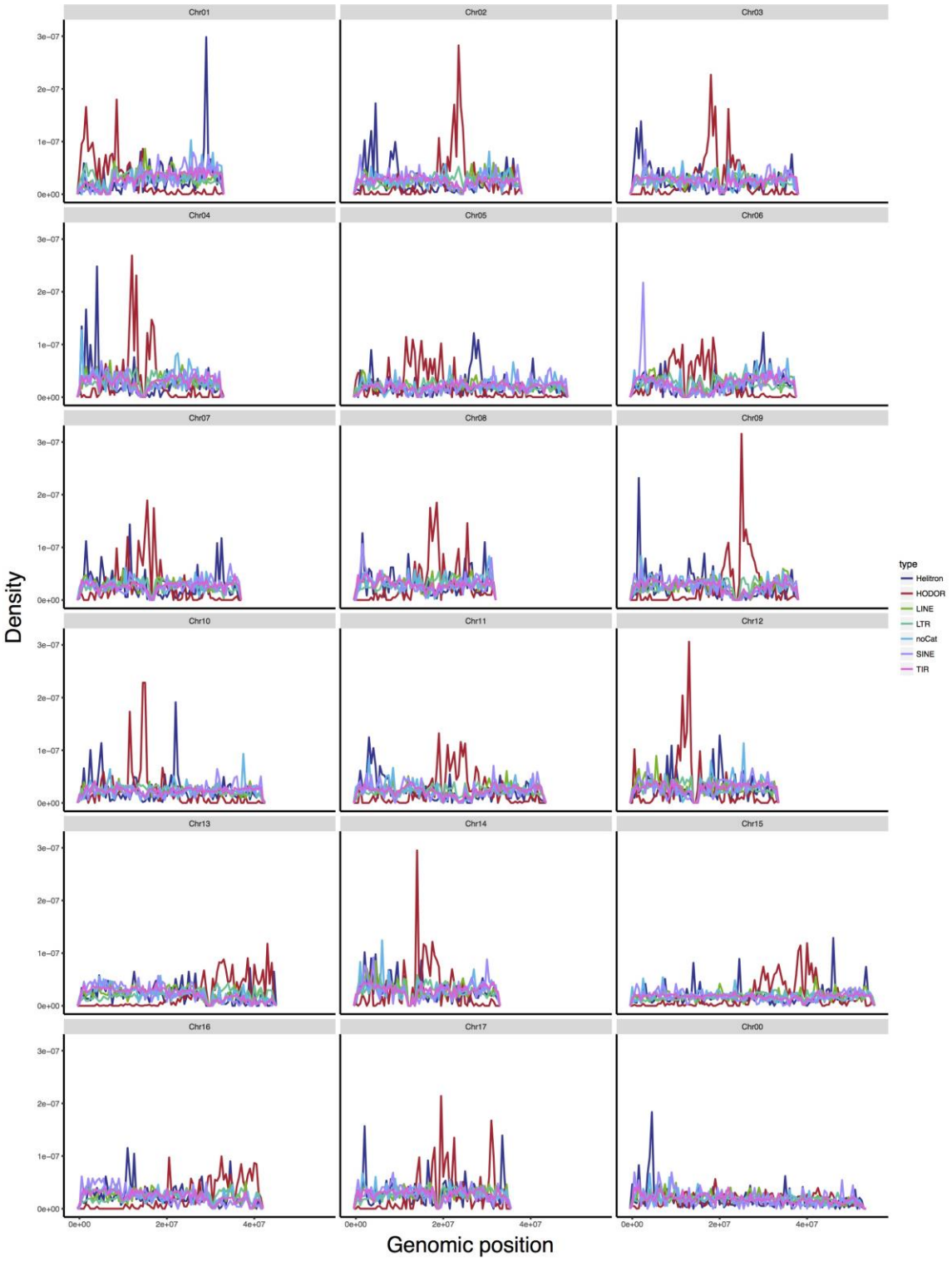
Haplotype map of GDDH13.

The genetic phase obtained through the construction of a genetic map from a cross between Golden Delicious and Renetta Grigia di Torriana was used to depict the two haplotypes of Golden Delicious in the GDDH13 assembly. The blue and red color represents the two different alleles of the heterozygous markers of Golden Delicious.

Supplementary Figure 6

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

Insert figure caption here by deleting or overwriting this text; captions may run to a second page if necessary. To ensure accurate appearance in the published version, please use the Symbol font for all symbols and Greek letters.

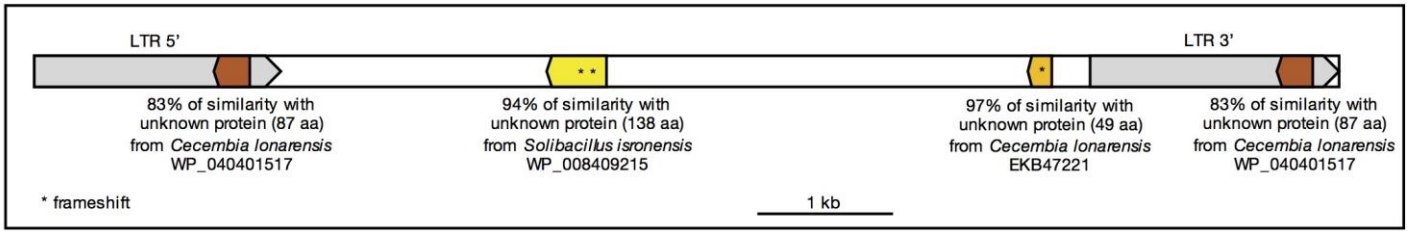


Supplementary Figure 7

Global TE and *HODOR* distribution shown for all chromosomes.

Chromosomal density plots of all TE families (see color code) and *HODOR* (red) shown on all chromosomes (including Chr00).

A



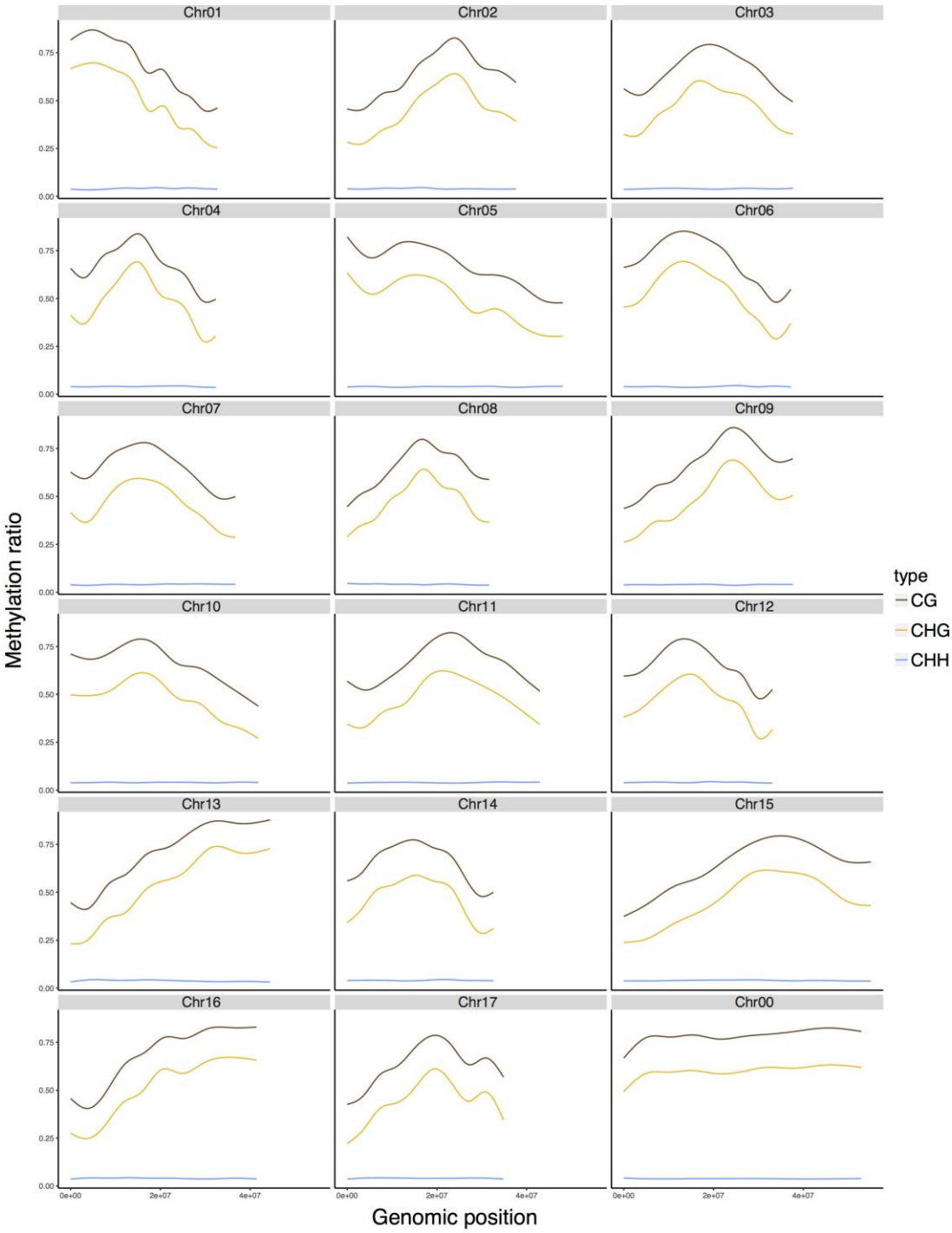
B

	seed	seedling	root	stem	leaf	flower	fruit	
probe_id	4442x2596	X4102	GD	GD	Leaf_M14	M74	M20	
AryANE_v1_00107370	1.89	-0.08	0.02	-0.01	0.23	-0.3	-0.03	sense
AryANE_v1_00044359	2.78	0.68	0.61	0.24	1.13	1.06	0.54	antisense

Supplementary Figure 8

HODOR structure and expression.

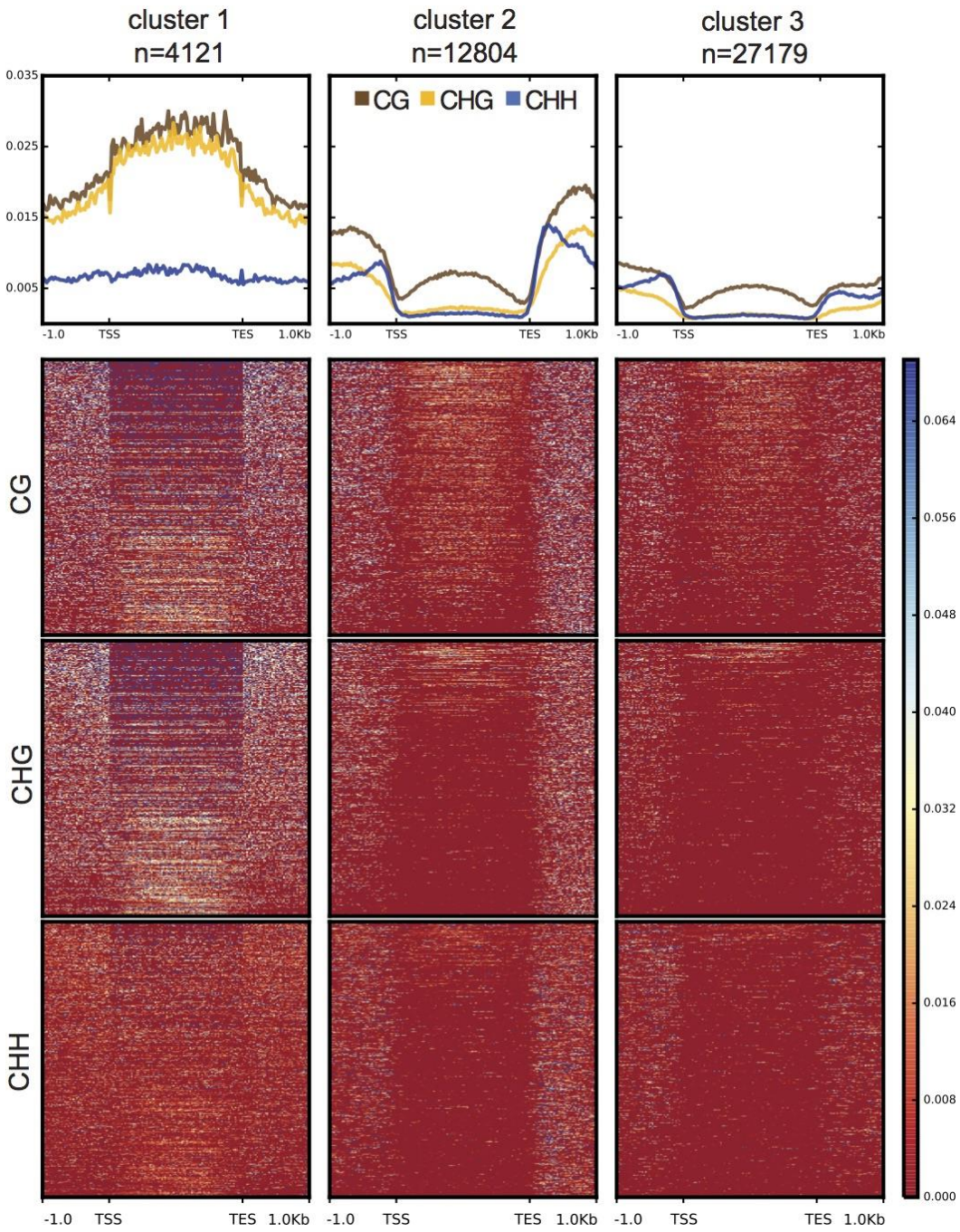
(A) Annotation of the *HODOR* consensus sequence. (B) log2 expression values of *HODOR* in different tissues.



Supplementary Figure 9

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.



Supplementary Figure 10

DNA methylation distribution on genes in apple.

Three main clusters of genes presenting different DNA methylation distributions over their body and surrounding sequences were detected. Cluster 1 includes genes rich in DNA methylation in their body, upstream and downstream of their transcription start site (TSS) and transcription end site (TES) respectively. Cluster 2 genes are located in genomic regions enriched in DNA methylation and cluster 3 genes are located in regions with less overall DNA methylation. The images at the bottom represent the level of DNA methylation for each gene of a cluster along its coding sequence, with a color coding, indicated at the right of the images, from low (red) to high (blue).

Supplementary Notes

Results

Genome sequencing, assembly and scaffolding

Homozygosity of the doubled-haploid:

Homozygosity of the GDDH13 line was confirmed by observation of the k-mer spectrum of Illumina reads. In Fig. 1A k-mer spectra of GDDH13 and of the heterozygous 'Golden Delicious' [1] 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.

Assembly details:

Details on the assembly pipeline (Fig. 1b): Illumina paired-end reads (150bp, 120X) were assembled with SOAPdenovo. The resulting Illumina contigs and PacBio reads (8.5kb mean length, 35X) were assembled with DBG2OLC into raw hybrid contigs. A consensus step was performed using the raw hybrid contigs, the Illumina contigs, and the PacBio reads with Sparc. A correction using Illumina paired-end reads (150bp, 120X) was performed with Pilon. A first scaffolding was performed using Illumina mate pair reads (15X) with BESST. The final scaffolding was performed using BioNano optical maps to obtain the final assembly.

All 162 scaffolds and 912 contigs resulting from the hybrid assembly of the BioNano optical map and Illumina/PacBio first assembly were mined for SNP (Single Nucleotide Polymorphism) markers that were genetically mapped on a high density multi-family genetic linkage map [2], by aligning their probe sequences [3]. Based on this information all 162 scaffolds plus 111 contigs were anchored, oriented and assembled into 17 pseudomolecules corresponding to the 17 chromosomes (with 10kb of undefined bases (N) added between each scaffold/contig, Supplemental Table 2), plus one pseudomolecule containing 801 contigs that could not be assigned to a chromosome (Chr00, 801 contigs, 45 Mb, sorted by descending order of size, with 10 kb of undefined bases (N) added between each contig). Based on the SNP markers data and on a visual analysis of the BioNano optical maps, we identified seven scaffolds with major inconsistencies between the genetic and physical maps. These seven scaffolds were split based on the latest correct SNP position and the previously conflicting sequence was re-assigned to its correct location on the chromosomes. This brought the total number of assembled sequences to 1,081.

Genome quality assessment

Discrepancy between genetic linkage map and pseudo-molecule assembly:

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).

Comparison of GDDH13 genome assembly with previous assemblies of the heterozygous 'Golden Delicious':

In the previous apple 'Golden Delicious' heterozygous genome sequence (Velasco et al. 2010), several groups of contiguous markers showed high LD within each group and with other distantly located groups of markers and at the same time no LD to neighboring groups of markers [4]. In the present version of the GDDH13 genome, no abrupt jumps in LD were identified.

Genome annotation

Description of Gene Ontology annotation:

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 [5], 32,109 genes (76%) were distributed among 3,853 gene families.

Transposable elements and repeats annotation

Description of TE content in the GDDH13 genome:

Of the TE identified in the GDDH13 genome, 2.2% (0.9% of genome assembly) were labeled as unclassified. In addition, potential host genes represent 0.97% of the genome assembly

Analysis of *HODOR*:

Within the GDDH13 genome, we identified 3 autonomous LTR-retrotransposons consensuses with all typical domains of copia elements (in red in Supplemental Table 3). These 3 consensuses showed sequence homologies with a mean identity of 82% (85% on each LTR) and allowed the identification of common features such as 'TG' and 'CA' dinucleotides in LTRs, specific primer binding sites (PBS) and polypurine tract (PPT) sites. These results suggest a putative mobilization of *HODOR* by some of these elements.

Discussion

GDDH13 assembly comparison with previous version of the heterozygous 'Golden Delicious' genome:

Following Illumina-only sequence assembly, Li and colleagues [1] report a N50 of 0.5 kb, while at the same assembly step we obtained a N50 of 7.3 kb.

GDDH13 annotation:

Our gene prediction reduced the estimated number of annotated genes in apple from 63,541 (www.rosaceae.org [6]) to 42,140. 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 [7]. 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 [8]. 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, [9]) and diploid strawberry (34,809 genes, 240 Mb, [10]). 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).

GDDH13 linkage groups orientation:

In apple and pear, the denotation and orientation of linkage groups of all recent genetic linkage maps as of the pseudo-chromosomes of the first 'Golden Delicious' heterozygous genome assembly [6] is based on the genetic linkage map of Maliepaard *et al.* [11].

Since this first published genetic map, large stretches of sequence homology between pairs of linkage groups has been detected by multi-locus targeting RFLP and microsatellite markers, but this information has not been used to orient the linkage groups. These observations have been since confirmed and duplication patterns definitions have been considerably improved ([6,12], this study). In particular, the homology of chromosomes 5 and 10 but in opposite orientation *sensu* Maliepaard *et al.* [11] has been confirmed. As mirrored orientations hamper a straight forward comparison, in this new genome we follow a newer convention [2] where the orientation of this pair of chromosomes becomes aligned by the inversion of Chr05, being the least frequently reported of the two in previous genetic studies on QTL, gene-discovery and characterization.

GDDH13 genome duplication and synteny:

Interestingly, we observed that in most instances, chromosome rearrangements following genome duplication occurred primarily around interstitial regions (eg. Chr01, Chr02, Chr04, Chr06, Chr07, Chr12, Chr14), with some exceptions (top of Chr04 syntenous to bottom of Chr13, top of Chr06 syntenous to bottom of Chr16).

GDDH13 genome methylation analysis:

The global DNA methylation analysis revealed a higher DNA methylation level than what was observed in Arabidopsis, more similar to soybean, which has a 1.1 Gb genome [13] but a lower level than in maize which has a very high TE content (86% for CG, 74% for CHG and 5.4% for CHH methylation [72]).

GDDH13 young fruit versus leaf methylome analysis:

Among the genes having a DMR within their putative promoter region between fruit and leaf tissues, we identified an apple gene orthologous to *CMT2* from Arabidopsis. This plant-specific DNA methyltransferase methylates mainly cytosines in CHH contexts, predominantly at TE elements [14]. It has previously been shown that natural variations in *CMT2* can be associated with genome-wide differences in CHH methylation levels [15,16].

GDDH13 fruit methylome versus GDDH18 fruit methylome:

Among the genes having a DMR in their putative promoter region, we identified three that could potentially contribute to fruit size difference:

An *SPL13* orthologue was differentially methylated between GDDH13 and GDDH18. The miR156/SPL module has been implicated in developmental timing in several species such as Arabidopsis and rice [17], and was found to be involved in tomato fruit development by controlling the rate of cell differentiation [18]. The different methylation profiles observed in the promoter of this gene between fruits of GDDH13 and GDDH18 may contribute to the observed phenotypic difference. On the other hand, plant hormones such as auxin, gibberellic acid and ethylene have been associated with control of plant size and architecture [19]. In previous studies, ethylene treatment has been linked to both stimulation [20] and inhibition of cellular division [21]. In Arabidopsis, increased ethylene signaling through over-expression of *ACS6* resulted in reduced stature and smaller leaf size [22]. More recently, increased expression of *ACS8*, an enzyme involved in the first steps in the biosynthesis of ethylene in poplar, was shown to negatively affect leaf size through a reduction in cell number [19]. The differential methylation profile observed in the promoter region of *ACS8* between fruits of GDDH13 and GDDH18 may thus contribute to the observed decrease in number of cell layers in the parenchyma of GDDH18 fruits. Finally, we identified a DMR in the promoter of *CYP71A25* orthologue. In Arabidopsis, expression of this gene was found to be up-regulated in the M phase of mitosis, indicating its potential involvement in cell-cycle [23].

Supplemental Material and Methods

Plant material

Origin of the two doubled-haploid 'Golden Delicious' apple trees was described by [24]. 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 which leaves spontaneously produced two independent and simultaneous chromosome doubling events *in vitro*, named GDDH13 and GDDH18. These plants were then rooted and grown in the orchard (1989, first fruits in 1995).

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

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 (clonaly 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.

Histological staining and microscopy

Our aim here was to characterize fruitlets development at a cellular level and to identify the stage at which fruits from GDDH13 and GDDH18 differentiated. For this, two fruitlets per tree were harvested from -3 DAP to 28 DAP, fixed in 4% (v/v) glutaraldehyde in phosphate buffer (0.1M; pH 7.2) for 3 hours at 4°C under vacuum, and rinsed in two changes of buffer. Fixed fruit samples were then dehydrated in a graded ethanol series and embedded in technovit 7100 resin (Kulzer Histo-technique kit, Labonord, Templemars, France) according to Kroes et al. (European Journal of Plant Pathology 104, 725–736, 1998). Specimens were then stored at 37°C. Median sections of fruitlets were cut at 3 µm with a Leica RM2165 microtome, mounted on glass slides, stained with toluidine blue, mounted after dehydration in a synthetic resin and examined under a Leica DM1000 microscope equipped with a Qimaging Micropublisher 3.3 RTV camera. The number of cell layers was determined by counting the number of cells in the parenchyme, traversed by a straight line starting from the fruit centre to the epidermis.

DNA purification

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.

Nature Genetics: doi:10.1038/ng.3886

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

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 [6] using BWA mem with default settings. All Illumina reads were then subjected to kmer spectrum based error correction using SoapEC [26] 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 [27] to remove short fragment read pairs in forward reverse orientation leaving only true mate pair fragments in forward reverse orientation. The FastUniq software [28] 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

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.

Nature Genetics: doi:10.1038/ng.3886

BioNano genomics genome mapping

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.Bsp*Q1 (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.

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 [26] in multi kmer mode with all kmer values from 51 to 127 and filtering out kmers with frequency lower than 3 prior to assembly. The resulting assembly had an N50 of 7,289 bp (Table 1). 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 and Illumina contigs were combined to perform a hybrid assembly using the DBG2OLC pipeline [29] with the following parameters: kmer size 17 as advised by [30], removeChimera parameter 1. A broad range of the three critical parameters (AdaptativeTh, 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 [31] before calling a consensus sequence with Sparc [32]. Parameter sweeps were performed for the critical DBG2OLC parameters in order to optimize the N50 and the Assembly size. A genome assembly was

also performed using only PacBio reads. This assembly was performed using the PbcR pipeline and run with the parameters advised in [33].

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 [34]. This alignment was then used with Pilon [35] which computed a consensus base for each position.

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 [36] using the default parameter.

BioNano scaffolding

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.

Scaffold validation and anchoring to genetic map

An integrated multi-parental genetic linkage map of apple [2] 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® 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 [2]. The probe sequence of the 15,417 markers [3] 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 [37]. 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

The "Old Dessert" INRA core collection, comprising 278 accessions [38], was genotyped with the Axiom® Apple-480K SNP genotyping array [39] 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* [40] (R package version 1.16.0). Heatmaps of pairwise LD between markers were plotted using the R package *LDheatmap* [41]. For each chromosome, one marker every ten was used to illustrate LD at a whole genome scale.

mRNA-Seq

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 [42]), 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[42]). 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 [43]. 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.

DNA extraction from leaf and developing fruits and bisulfite sequencing

Young leaves and developing fruits 9 days after pollination (DAP) were collected from two biological replicates of a GDDH13 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 Epitect 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. DMRs were computed according to [44]. After mapping of the reads the average coverages ranged from 7 to 10-fold.

DNA methylation distribution plots and gene clustering by methylation patterns were performed with deepTools [45].

smallRNA alignment

Apple sRNA derived from mature fruit parenchyme [46] 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.

Genome 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 [47] and SOAPdenovo-trans [48]. 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 [49], Red [50] and BLASTx comparisons against Repbase [51]. The structural annotation of coding genes was performed using EuGene [52] by combining Gmap transcript mapping [53], 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 has been completed by tRNAscan-SE [54], RNAmmer [55] and RfamScan [56] in order to annotate non-protein coding genes.

Functional annotation of proteins was performed using InterProScan [57]. The functional annotation was then completed by the prediction of targeted signals using the TargetP software [58].

Genome synteny

SynMap (CoGe, www.genomevolution.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.

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

Malus domestica predicted genes (MDP) sequences obtained from the heterozygous genome annotation [6] were mapped to the GDDH13 genome assembly using the best BLAT [59] 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++ [60].

Repeat annotation

The TE*denovo* pipeline [61,62] from the REPET package v2.5 (https://urgi.versailles.inra.fr/Tools/REPET) was used to detect transposable elements (TEs) in genomic sequences and to provide a consensus sequence for each TE family. First, the 18 pseudo-molecules of this genome assembly were deconstructed into « virtual » contigs by removing stretches > 11 undefined bases (Ns) to exclude gaps in these sequences. This step generated 2,170 « virtual » contigs with a N50 of 589 Kb, for a total length of 625 Mb. A subset of the « virtual » contigs (with size > 400 Kb), representing approximately 412 Mb, was selected to run the TE*denovo* repeats detection pipeline. A minimum of 5 sequences per group was used as parameter and a library of 15,850 consensus sequences classified according to structural and functional features (similarities with characterized TEs from RepBase21.01 database [51] and domains from Pfam27.0, specially formatted for REPET) was generated. After removing the redundancy and filtering out consensus sequences classified as SSR and unclassified consensus sequences constructed with < 10 copies in the genome, a library of 6,219 consensus sequences was used to annotate the TE copies in the whole genome using the TEannot pipeline [63] from the REPET package v2.5 (with default parameters). To refine the TE annotation, consensus sequences showing no full-length fragments (i.e. fragment covering more than 95% of the consensus sequence) in the genome were filtered out and a subset of 2,517 consensus sequences were used to run a second TEannot iteration on the whole genome. Consensus sequences 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* (*High Copy Golden Delicious Repeat*) consensus sequence on the PacBio assembly with REPET package v2.5 pipelines. TE insertion ages were calculated using the adapted T=K/r formula for non-duplicated LTR sequences [64] where T is the time to most common ancestry, K is the sequence divergence, and r is the substitution rate. r was set to the TE-specific rate of 1.3×10^{-8} [65]. The observed sequence divergence was corrected with the Jukes and Cantor model [66].

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 2014). A SNP and small indels identification was performed using freebayes (Garrison 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 Suppl. Table 4.

Supplementary References

1. Li, X. *et al.* Improved hybrid de novo genome assembly of domesticated apple (Malus x domestica). *Gigascience* **5,** 1–5 (2016).
2. 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,** 16057–13 (2016).
3. Bianco, L. *et al.* Development and Validation of a 20K Single Nucleotide Polymorphism (SNP) Whole Genome Genotyping Array for Apple (Malus x domestica Borkh). *PLOS ONE* **9,** (2014).
4. Leforestier, D. Localization of genomic regions controlling the variation of fruit quality and disease resistance traits in apple: selection signatures and association genetics. (www.theses.fr, 2015).
5. Finn, R. D. *et al.* The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res* **44,** D279–85 (2016).
6. Velasco, R. *et al.* The genome of the domesticated apple (Malus × domestica Borkh.). *Nat Rev Genet* **42,** 833–839 (2010).
7. Veeckman, E., Ruttink, T. & Vandepoele, K. Are We There Yet? Reliably Estimating the Completeness of Plant Genome Sequences. *Plant Cell* **28,** tpc.00349.2016–25 (2016).
8. Denton, J. F. *et al.* Extensive Error in the Number of Genes Inferred from Draft Genome Assemblies. *PLoS Comp Biol* **10,** e1003998 (2014).
9. Verde, I. *et al.* The high-quality draft genome of peach (Prunus persica) identifies unique patterns of genetic diversity, domestication and genome evolution. *Nat Rev Genet* **45,** 487–494 (2013).
10. Shulaev, V. *et al.* The genome of woodland strawberry (Fragaria vesca). *Nat Genet* **43,** 109–116 (2010).
11. Maliepaard, C. *et al.* Aligning male and female linkage maps of apple (Malus pumila Mill.) using multi-allelic markers. *Theor Appl Genet* **97,** 60–73 (1998).
12. Celton, J. M., Tustin, D. S., Chagne, D. & Gardiner, S. E. Construction of a dense genetic linkage map for apple rootstocks using SSRs developed from Malus ESTs and Pyrus genomic sequences. *Tree Genetics & Genomes* **5,** 93–107 (2009).
13. Schmutz, J. *et al.* Genome sequence of the palaeopolyploid soybean. *Nature* **463,** 178–183 (2010).
14. Zemach, A. *et al.* The Arabidopsis nucleosome remodeler DDM1 allows DNA methyltransferases to access H1-containing heterochromatin. *Cell* **153,** 193–205 (2013).
15. Dubin, M. J. *et al.* DNA methylation in Arabidopsis has a genetic basis and shows evidence of local adaptation. *eLife* **4,** e05255 (2015).
16. Kawakatsu, T. *et al.* Epigenomic Diversity in a Global Collection of Arabidopsis thaliana Accessions. *Cell* **166,** 492–505 (2016).
17. Preston, J. C. & Hileman, L. C. Functional Evolution in the Plant SQUAMOSA-PROMOTER

BINDING PROTEIN-LIKE (SPL) Gene Family. *Frontiers in Plant Science* **4,** (2013).
18. Wang, H. & Wang, H. The miR156/SPL Module, a Regulatory Hub and Versatile Toolbox, Gears up Crops for Enhanced Agronomic Traits. *Molecular Plant* **8,** 677–688 (2015).
19. Plett, J. M., Williams, M., LeClair, G., Regan, S. & Beardmore, T. Heterologous over-expression of ACC SYNTHASE8 (ACS8) in Populus tremula x P. alba clone 717-1B4 results in elevated levels of ethylene and induces stem dwarfism and reduced leaf size through separate genetic pathways. *Frontiers in Plant Science* **5,** (2014).
20. Love, J. *et al.* Ethylene is an endogenous stimulator of cell division in the cambial meristem of Populus. *Proceedings of the National Academy of Sciences* **106,** 5984–5989 (2009).
21. Dubois, M. *et al.* Ethylene Response Factor6 acts as a central regulator of leaf growth under water-limiting conditions in Arabidopsis. *Plant Physiol* **162,** 319–332 (2013).
22. Liu, Y. & Zhang, S. Phosphorylation of 1-aminocyclopropane-1-carboxylic acid synthase by MPK6, a stress-responsive mitogen-activated protein kinase, induces ethylene biosynthesis in Arabidopsis. *Plant Cell* **16,** 3386–3399 (2004).
23. Menges, M. & Murray, J. A. H. Synchronous Arabidopsis suspension cultures for analysis of cell-cycle gene activity. *Plant J* **30,** 203–212 (2002).
24. Lespinasse, Y., Bouvier, L., Djulbic, M. & Chevreau, E. Haploidy in apple and pear. *Acta Hortic.* 49–54 (1999). doi:10.17660/ActaHortic.1998.484.4
25. Jaskiewicz, M., Peterhänsel, C. & Conrath, U. Detection of histone modifications in plant leaves. *J Vis Exp* e3096–e3096 (2011). doi:10.3791/3096
26. Luo, R. *et al.* SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience* **1,** (2012).
27. O'Connell, J. *et al.* NxTrim: optimized trimming of Illumina mate pair reads. *Bioinformatics* **31,** btv057–2037 (2015).
28. Xu, H. *et al.* FastUniq: A Fast De Novo Duplicates Removal Tool for Paired Short Reads. *PLOS ONE* **7,** e52249 (2012).
29. 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).
30. Ye, C., Hill, C. M., Wu, S., Ruan, J. & Ma, Z. S. DBG2OLC: Efficient Assembly of Large Genomes Using Long of the Third Generation Sequencing Technologies. *arxiv* 1–41 (2016).
31. Chaisson, M. J. & Tesler, G. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics* **13,** 238 (2012).
32. Ye, C. & Ma, Z. S. Sparc: a sparsity-based consensus algorithm for long erroneous sequencing reads. *PeerJ* **4,** e2016 (2016).
33. Berlin, K. *et al.* Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nature Biotechnology* (2015). doi:10.1038/nbt.3238
34. Li, H. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics* **30,** 2843–2851 (2014).
35. Walker, B. J. *et al.* Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLOS ONE* **9,** e112963 (2014).
36. Sahlin, K., Vezzi, F., Nystedt, B., Lundeberg, J. & Arvestad, L. BESST--efficient scaffolding of large fragmented assemblies. *BMC Bioinformatics* **15,** 281 (2014).
37. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* (2011). doi:10.1093/bioinformatics/btr011
38. 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 Molecular Biology Reporter* **34,** 1–18 (2016).
39. Bianco, L. *et al.* Development and validation of the Axiom(®) Apple480K SNP genotyping array. *Plant J* **86,** 62–74 (2016).
40. Clayton, D. snpStats: SnpMatrix and XSnpMatrix classes and methods. *R package version 1.16.0* (2014).
41. Shin, J. H., Blay, S., McNeney, B. & Graham, J. LDheatmap: an R function for graphical display of pairwise linkage disequilibria between single nucleotide polymorphisms. *JSS*

(2006).
42. Segonne, S. M. *et al.* Multiscale investigation of mealiness in apple: an atypical role for a pectin methylesterase during fruit maturation. *BMC Plant Biol* **14,** 1995 (2014).
43. Nobile, P. M. *et al.* Identification of a novel α-L-arabinofuranosidase gene associated with mealiness in apple. *J Exp Bot* **62,** 4309–4321 (2011).
44. Hagmann, J. *et al.* Century-scale Methylome Stability in a Recently Diverged Arabidopsis thaliana Lineage. *PLoS Genet* **11,** e1004920 (2015).
45. Ramírez, F., Dündar, F., Diehl, S., Grüning, B. A. & Manke, T. deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res* **42,** W187–91 (2014).
46. Celton, J. M. *et al.* Widespread anti-sense transcription in apple is correlated with siRNA production and indicates a large potential for transcriptional and/or post-transcriptional control. *The New phytologist* **203,** 287–299 (2014).
47. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology* **29,** 644–652 (2011).
48. Xie, Y. *et al.* SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. *Bioinformatics* **30,** 1660–1666 (2014).
49. Ellinghaus, D., Kurtz, S. & Willhoeft, U. LTRharvest , an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics* **9,** 18 (2008).
50. Girgis, H. Z. Red: an intelligent, rapid, accurate tool for detecting repeats de-novo on the genomic scale. *BMC Bioinformatics* **16,** 860 (2015).
51. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA* **6,** 11 (2015).
52. Foissac, S. *et al.* Genome annotation in plants and fungi: EuGene as a model platform. *Current Bioinformatics* **3,** 87–97 (2008).
53. Wu, T. D. & Watanabe, C. K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21,** 1859–1875 (2005).
54. Lowe, T. M. & Chan, P. P. tRNAscan-SE On-line: integrating search and context for analysis of transfer RNA genes. *Nucleic Acids Res* **44,** W54–7 (2016).
55. Lagesen, K. *et al.* RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res* **35,** 3100–3108 (2007).
56. Nawrocki, E. P. *et al.* Rfam 12.0: updates to the RNA families database. *Nucleic Acids Res* **43,** D130–7 (2015).
57. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30,** 1236–1240 (2014).
58. Emanuelsson, O., Brunak, S., Heijne, von, G. & Nielsen, H. Locating proteins in the cell using TargetP, SignalP and related tools. *Nature Protocols* **2,** 953–971 (2007).
59. Kent, W. J. BLAT--the BLAST-like alignment tool. *Genome Res* **12,** 656–664 (2002).
60. Guéguen, L. *et al.* Bio++: efficient extensible libraries and tools for computational molecular evolution. *Molecular Biology and Evolution* **30,** 1745–1750 (2013).
61. Flutre, T., Duprat, E., Feuillet, C. & Quesneville, H. Considering transposable element diversification in de novo annotation approaches. *PLOS ONE* **6,** e16526 (2011).
62. Hoede, C. *et al.* PASTEC: an automatic transposable element classification tool. *PLOS ONE* **9,** e91929 (2014).
63. Quesneville, H. *et al.* Combined evidence annotation of transposable elements in genome sequences. *PLoS Comp Biol* **1,** 166–175 (2005).
64. la Chaux, de, 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. *Mobile DNA* **3,** 2 (2012).
65. Ma, J. & Bennetzen, J. L. Rapid recent growth and divergence of rice nuclear genomes. *Proc Natl Acad Sci USA* **101,** 12404–12410 (2004).
66. Jukes, T. H. & Cantor, C. R. in *Mammalian Protein Metabolism* 21–132 (Elsevier, 1969). doi:10.1016/B978-1-4832-3211-9.50009-7

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

Supplemental Table 1. Metrics of the different steps performed for the GDDH13 genome assembly.

			Number	Total size (Mb)	% BioNano map assembly (649.7 Mb)	% genome size (651 Mb)
Hybrid assembly	Contigs		912	57.1	8.8	8.8
	Scaffolds	with N	162 (+7 after chimera identification)	643.2	99.0	98.8
		without N	162 (+7 after chimera identification)	568.2	87.5	87.3
Pseudo-chromosomes	Scaffolds and contigs anchored without N		280	580.4	89.3	89.2
	Contigs in Chr00, without N		801	45.0	6.9	6.9
	Total without N		1081	625.4	96.3	96.1
Annotation	Genes	Protein coding genes	42,140	151.2	23.3	23.2
		Non protein coding genes	1,965	4.0	0.6	0.6
	Transposable Elements	Class I	393,464	278.8	42.9	42.8
		Class II	299,637	87.3	13.4	13.4
		No Category	23,581	6.1	0.9	0.9
		Total	716,682	372.2	57.3	57.2

Supplemental Table 2. Summary of the genome assembly features and annotations of the apple GDDH13 genome.