Title:

Chromosome-scale *de novo* diploid assembly of the apple cultivar 'Gala Galaxy'

Authors:

Giovanni A.L. Broggini[a,b,*], Ina Schlathölter[a,b], Giancarlo Russo[c], Dario Copetti[a], Steven A. Yates[a], Bruno Studer[a], Andrea Patocchi[b]

a. Molecular Plant Breeding, Institute of Agricultural Sciences, ETH Zurich, Universitaetstrasse 2, 8092 Zurich, Switzerland

b. Breeding Research, Research Division Plant Breeding, Agroscope, Mueller-Thurgau-Strasse 29, 8820 Waedenswil, Switzerland

c. Functional Genomics Center Zurich, University of Zurich and ETH Zurich, Winterthurerstrasse 190, 8057, Zurich, Switzerland

[*]Author for Correspondence: Giovanni A.L. Broggini, Molecular Plant Breeding, Institute of Agricultural Sciences, ETH Zurich, Universitaetstrasse 2, 8092 Zurich, Switzerland, +41584606308, giovanni.broggini@usys.ethz.ch

Abstract:

Apple (*Malus* × *domestica*) is one of the most important fruit crops in terms of worldwide production. Due to its self-incompatibility system and the long juvenile period, breeding of new apple cultivars combining traits desired by growers (e.g. yield, pest and disease resistance) and consumers (e.g. fruit size, color, and flavor) is a long and complex process. Genomics-assisted breeding strategies can facilitate the selection of germplasm leading to new cultivars. While the most complete apple genome assemblies available to date are from anther-derived homozygous lines, *de novo* assembly of apple genomes encompassing the natural heterozygosity remains challenging. Using long- and short-read sequencing technologies in combination with optical mapping, we *de novo* assembled a diploid and heterozygous genome of the apple cultivar 'Gala Galaxy'. This approach resulted in 154 hybrid scaffolds (N50 = 34.3 Mb) spanning 999.9 Mb and in 414.7 Mb of unscaffolded sequences. Anchoring 31 scaffolds with a genetic map was sufficient to represent an entire haploid genome of 17 pseudomolecules (719.4 Mb). The remaining sequences were assembled in a second set of 17 pseudomolecules, which spanned 601 Mb, leaving 80.6 Mb of unplaced sequences. A total of 41,264 genes were annotated using 74,900 transcripts derived from RNA sequencing of pooled leaf tissue samples. This study provides a high-quality diploid reference genome sequence encompassing the natural heterozygosity of the widely popular cultivar 'Gala Galaxy'. The DNA sequence resources and the assembly described here will serve as a solid foundation for fundamental and applied apple breeding research.

Keywords: Malus, heterozygous, genome, optical mapping, hybrid scaffolds

35 **Introduction:**

36 Apple (*Malus* × *domestica*) is the third most important fruit crop worldwide, with an annual production reaching 83
37 million metric tons in 2017 (FAOSTAT 2004). Most of the commercially successful cultivars are susceptible to a
38 number of pests and diseases (Turechek 2004), and the production of high-quality fruits requires many
39 applications of plant protection products. Given the increasing public demand to reduce fungicide input, disease
40 resistant cultivars are essential for future sustainable apple production. Genotypic information has the potential to
41 speed up the development of resistant cultivars of superior quality (Baumgartner, et al. 2016; Laurens, et al. 2018;
42 Peace, et al. 2019). However, the number of available molecular markers linked to fruit quality traits is low.
43 Several projects using genome-wide association studies or establishing genomic selection for fruit quality traits in
44 apple were initiated (Kumar, et al. 2012; Muranty, et al. 2015; Roth, et al. 2019), often relying on SNP chip-based
45 genotypic data (Bianco, et al. 2016; Bianco, et al. 2014). Once the association of genetic markers to a trait is
46 observed, a complete reference genome sequence is a requisite for linking these markers to the underlying
47 genomic features. With decreasing per base sequencing costs, sequencing-based genotyping (e.g. genotyping-
48 by-sequencing or skim sequencing) requiring a reference genome for sequence alignment may partly supplant
49 SNP chip-based genotyping. To meet the challenges described above, several projects already sequenced apple
50 genomes (Daccord, et al. 2017; Li, et al. 2016; Velasco, et al. 2010; Zhang, et al. 2019). The *de novo* assembly of
51 heterozygous genomes is recognized as a challenging task, as they generally result more fragmented than
52 homozygous genomes (Pryszcz and Gabaldon 2016). So far, Velasco (2010) and Li (2016) assembled the
53 genome of the heterozygous genotype 'Golden Delicious', with both assemblies resulting fragmented (N50 = 16
54 kb and 111 kb, respectively). Daccord *et al.* (2017) and Zhang *et al.* (2019) overcame the assembly challenges by
55 sequencing the genome of the homozygous anther-derived lines GDDH13 and HFTH1 obtaining more contiguous
56 assemblies (N50= 5.5 Mb and 7.0 Mb, respectively).

57 A complete heterozygous genome assembly allows investigating haplotype divergence within a single cultivar, but
58 also to study the variation in the genome of sport mutants from a specific apple cultivar. Sport mutants are
59 variations identified as single branches of the original cultivar producing fruits with improved characteristics
60 (Foster and Aranzana 2018). For instance for the popular cultivar 'Gala', more than 30 sport mutants have been
61 commercialised (Dickinson and White 1986; Iglesias, et al. 2008) and represent one of the most successful
62 cultivar group worldwide. The exact reason for the emergence of sport mutants is unclear but likely due to
63 imperfect repair following errors in DNA replication, active transposable elements (Foster and Aranzana 2018;
64 Lee, et al. 2016) and/or epimutations (inheritable changes of DNA methylation, histone acetylation or chromatin
65 remodeling) that influence genes transcription (El-Sharkawy, et al. 2015). Therefore, the availability of a complete
66 heterozygous (diploid) genome assembly, besides supporting conventional breeding research, can help
67 understanding the events underlying the generation of novel sport mutants and, more generally, the accumulation
68 of mutations in this vegetatively propagated crop.

Material and Methods:

Assembly strategy

Genomic DNA was isolated from field-grown apple leaves of 'Gala Galaxy' and processed for library construction for long- and short-reads sequencing as described in the supplementary methods. Additional long-reads libraries were also generated from four genotypes of the 'Gala' sport mutants group; ‘Gala’ original, ‘Gala Royal’, ‘Gala Schnico Red’ and the cisgenic apple line C44.4.146 (Kost, et al. 2015). The long-reads libraries from 'Gala Galaxy' and the additional genotypes were sequenced using PacBio RSII and Sequel instruments, respectively. The complete PacBio sequence dataset was *de novo* assembled with FALCON Unzip, v.0.4.0 (Chin, et al. 2016). Contigs consisting in more than 50% organellar (chloroplast or mitochondrial) genome sequences were identified by BLAST search (Altschul, et al. 1990) and removed, together with contigs smaller than 1 kb. Short-reads libraries ('Gala Galaxy') were sequenced on an Illumina Hiseq4000 using the paired-end 150 bp module. These were used for polishing of the assembly and for genome size estimation as described in the supplementary methods.

Contiguity of the assembly was then increased combining two optical maps to generate dual enzyme hybrid scaffolds. DNA extraction and labelling, and assembly strategy of the optical maps are described in the supplementary methods. The scaffolds were then oriented and ordered by ALLMAPS (Tang, et al. 2015) to achieve a chromosome-scale assembly, as described in the supplementary methods. Several ALLMAPS runs integrated the two evidence sets producing chromosome-scale pseudomolecules for both sets of homologous chromosomes. The diploid assembly was investigated for completeness using the Benchmarking Universal Single-Copy Orthologs (Simao, et al. 2015) pipeline v3.0 with 1,440 conserved plant genes (embryophyta_odb9) available in the discovery environment of www.cyverse.org. To evaluate the correctness of the genome assembly, diagnostic genome features were visualized in RStudio (Version 1.2.5001) using the karyoploteR package (Gel and Serra 2017) as described in the supplementary methods.

Gene Annotation

For evidence-based gene annotation, RNA was extracted separately from 15 pools of three leaves collected from 1-year old field-grown ‘Gala Galaxy’ trees. Leaves were frozen in liquid nitrogen and ground to fine powder. RNA extraction protocol and libraries preparation are described in the supplementary methods. Libraries were sequenced on an Illumina HiSeq4000 instrument, generating paired-end reads (length = 150bp). Reads were checked for quality using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and trimmed using fastp (Chen, et al. 2018) retaining only sequences with Q>30. Libraries were mapped against the reference genome using bowtie2 v2.2.3 (Langmead 2010) and transcripts were identified with cufflinks v2.1 and tophat 2.0.13 (Trapnell, et al. 2012). Transcripts were functionally annotated as described in Knorst et al. (2019). Gene

annotations were used to investigate collinearity between haploid assemblies by generating synthenic dotplots with Synmap2 (Haug-Baltzell, et al. 2017).

K-mer Analysis Toolkit

The use of a haploid reference for sequencing read mapping may result problematic if one or both haplotypes from a heterozygous genotypes show high divergence to the corresponding haplotype in the reference. Sequence reads from diverging haplotypes would not be mapped to the reference and result in information loss. The k-mer analysis toolkit (Mapleson, et al. 2016) was used to assess the presence and count of Illumina sequencing data (in k-mers) of 'Gala Galaxy' (heterozygous) in different assemblies. For this analysis, only forward (R1) Illumina reads ('Gala Galaxy') were used in the KAT analysis with the GDDH13 assembly (Daccord, et al. 2017) as well as the diploid assembly reported here as reference. In addition, we included in this analysis a graph-based phased genome generated as described in the supplementary methods with Whatshap (Patterson, et al. 2015) using the primary haploid assembly MDGGph_v1.0 as reference.

Results and Discussion

Sequencing: For the genotype 'Gala Galaxy', PacBio RSII generated a total of 32 Gb long-reads sequence data using 29 SMRT cells, while Illumina HiSeq4000 instrument generated 694 million short-reads (104 Gb, submitted to SRA, accession XXXXXX). Using 19-mers, this latter data allowed estimating a genome size of 710.5 Mb (data not shown). To increase coverage of long-reads sequences, PacBio Sequel generated additional long-reads sequence data for other four genotypes of the 'Gala' group; 'Gala' original (5 SMRT cells, 9.4 Gb), 'Gala Royal' (8 SMRT cells, 8.8 Gb), 'Gala Schniga® SchniCo red' (2 SMRT cells, 8.3 Gb) and C44.4.146 (3 SMRT cells, 11.6 Gb). Combining all generated long-reads data, FALCON unzip produced 2,061 primary contigs and 5,663 haplotigs (N50 = 0.59 Mb, total length = 1,308.97 Mb, Table 1). A total of 201 organellar contigs (total of 8.57 Mb) and 31 contigs smaller than 1 kb were removed from the assembly.

Scaffolding: Two optical maps were used for scaffolding of the assembled contigs. The alignment of 316,748 (out of 2,290,666) NLRS-labelled molecules produced 1,662 optical contigs (N50 = 0.65 Mb, total length = 948.63 Mb, coverage = 20×). The alignment of 329,921 (out of 2,117,673) DLE-1 labelled input DNA molecules produced 296 optical contigs (N50 = 15.53 Mb, total length = 1,806.45 Mb, coverage = 29×). Dual enzyme hybrid scaffolding was performed using Bionano Access v1.2.1 and Bionano Solve v3.2.1. Default settings were used to perform the hybrid scaffolding. Dual enzyme hybrid scaffolding (incorporating both DLS and NLRS maps) resulted in a hybrid assembly with N50 of 13.7 Mb (scaffold only N50 = 34.28 Mb) for a total length of 1,414.64 Mb and consisted of 154 scaffolds (999.96 Mb) and 6,336 unscaffolded contigs (414.68 Mb, Table 1).

Anchoring: The first round of analyses on the whole hybrid assembly with ALLMAPS identified 31 scaffolds that were sufficient to represent one haploid genome. Seven pseudomolecules (chr8, 9, 12, 13, 14, 16, and 17) were covered by one scaffold each. For each one of the remaining pseudomolecules (covered by two to three scaffolds), overlapping regions between scaffolds were searched and eight overlapping regions were removed manually. When re-anchored using ALLMAPS, a primary haploid genome (MDGGph_v1.0) assembly spanning 719 Mb in 17 chromosome-scale pseudomolecules (chr1-17) was generated. Its size is in agreement with the estimated genome size of 710 Mb. MDGGph_v1.0 served as haploid reference assembly in order to map long and short reads generating the graph-based genome for subsequent k-mer analysis. The remaining sequences were assembled again using ALLMAPS into a second set of 17 pseudomolecules (secondary haploid genome assembly MDGGsh_v1.0, chr_s1-17) spanning 601 Mb, with 80 Mb unanchored sequences. In four cases the secondary haploid assembly showed longer pseudomolecules when compared to the primary assembly (chromosomes chr_s5, chr_s 6, chr_s11 and chr_s 14, supplementary Table 1). This is possibly due to differences in chromosome length or misassembles. Unplaced sequences (n = 2,853) spanned only 5.7% of the total assembly (Table 1).

Genome visualization displayed Illumina read coverage, SNP phase, and correlation of genetic and physical order for each chromosome pair on the assembly, as well as links between both alleles of each SNP marker (Genome visualization, supplementary data). With the exception of a region on chr_s10, all collapsed regions (with an Illumina read coverage of about 130×) were present in MDGGph_v1.0. Our assembly show collinearity with the genetic map of 'Fuji' × 'Gala'. Phase switches were observed along the pseudomolecules and, as expected, much lower SNP phase information was found in these collapsed regions (see Genome visualization, supplementary data).

The complete diploid assembly of 'Gala Galaxy' (MDGGdi_v1.0) obtained by combining primary and secondary haploid assemblies is available in Genbank as Bioproject XXXXXX. Genome assembly statistics are shown in Table 1, chromosome sizes are shown in supplementary Table 1, while analysis results of BUSCOs identified in MDGGdiv1.0, MDGGph_V1.0 and GDDH13 are shown in supplementary Table 2. The number of complete BUSCOs identified in MDGGdi_v1.0 was 1,387 (96.3%), of which 467 (32.4%) were present as single-copy and 920 (63.9%) as duplicated (Table 3). The same analysis on the primary haploid assembly MDGGph_v1.0 identified 1,347 (93.6%) complete BUSCOs, with 943 (65.5%) being single-copy and 404 (28.1%) were duplicated (supplementary Table 2).

Gene annotation: Using RNAseq data, a total of 74,900 leaf transcripts encoded by 41,264 genes were identified and functionally annotated as described in Knorst et al. (2019). Of the 74,900 proteins searched, 64,765 had an annotation assigned, of which 59,039 were assigned at least one GO term. The *de novo* annotated primary assembly MDGGph_v1.0 and the GDDH13 assembly were used to generate a synthenic dotplot, confirming a high collinearity between the two assemblies (Supplementary Figure 1).

K-mer analysis: Kat spectra plots were used to investigate k-mer multiplicity distribution and counting the occurrence of each k-mer in the investigated assemblies. The distribution of the k-mer counts in the KAT spectra plot revealed two peaks: the first peak at ×=27 represents the heterozygous content (and hence unique sequences) and the second peak at ×=58 the homozygous (Figure 1). The color of the area under the curve indicated how often k-mers were found in the investigated reference assemblies: black corresponds to k-mers missing in the assembly, while red and violet correspond to k-mers counted one and twice, respectively. The black area in the spectrum generated using GDDH13 is clearly larger than the one generated using MDGGdi_v1.0 (Figure 1A and 1B). The spectrum generated using the phased graph-based assembly (based on MDGGph_v1.0) shows a large black area, but also a violet area under the peak for heterozygous content (Figure 1C).

Discussion

Here, we present a diploid assembly of the heterozygous apple cultivar 'Gala Galaxy' (MDGGdi_v1.0). Optical mapping was essential in order to increase the contiguity of the assembly, as shown by the 60-fold increase of

N50 values, reaching 34.28 Mb for the hybrid scaffolds only (Table 1) and corresponding to a two-fold increase compared to the GDDH13 (Daccord, et al. 2017) and HFTH1 (Zhang, et al. 2019) apple genome assemblies. Despite being diploid, this assembly is unphased, as phase switches assessed by SNP markers were observed (Supplementary data). Phase switches are probably due to the unzip approach in FALCON, where the choice of which variant to include in the primary contigs is based on length – resulting in chimeric contigs. Due to the Mb resolution of the optical map, this data was not helpful in rearranging chimeric scaffolds into correctly phased scaffolds. BUSCO analyses confirmed that the assembly MDGGdi_v1.0 is complete, with only 3.7% missing BUSCOs. The diploid assembly presented in this study represents an advancement over existing resources. Further improvements might be achieved by separating the collapsed haplotypes and resolving the phase switches in to a correctly phased genome. Moreover, gene annotation could further be improved by integrating transcriptome data from different tissue types (e.g. fruits or flowers).

Our results clearly show that using a haploid reference is suboptimal to analyse genomic data derived from the heterozygous apple and might lead to sequence information loss, as indicated by the black area in Figure 1A. This occurred despite the closeness of these two genotypes (as 'Gala' is an offspring of 'Golden Delicious', from which the doubled-haploid GDDH13 was derived). Furthermore, the k-mer analysis indicated that a graph-based approach generated an assembly in which some allelic regions were absent while other were artificially duplicated (black and violet area in Figure 1C, respectively. Thus, by running WhatsHap on a haploid reference, the second allele could not be recovered as well as the *de novo* assembly by FALCON Unzip did. This may be a consequence of a high divergence between allelic regions and therefore, haplotype homology was not resolved correctly by this approach.

This diploid genome will thus be the appropriate reference for achieving the most information from resequencing projects of sport mutants of this important apple cultivar. Understanding the genome organization and including genomic information in the breeding process will enable the rapid development of resilient cultivars allowing a more sustainable production of this important fruit.

Availability of supporting data

The Illumina sequencing reads of each sequencing library and the RNA-seq data have been deposited at SRA with the accessions numbers XXXXXXX. The genome assembly MDGGdi_v1.0 was deposited at NCBI as Bioproject XXXXXXXX. The genome visualization is available as supplementary data.

Funding

This work was supported by the Swiss National Science Foundation Grant 31003A_163386.

Authors' contributions

GALB, BS, and AP designed the study. GALB and GR assembled the genome. IS and GALB collected leaf samples and extracted DNA. GALB, GR, SAY, and DC analyzed the data. All authors participated to the writing of the manuscript and approved the final version.

Competing interests

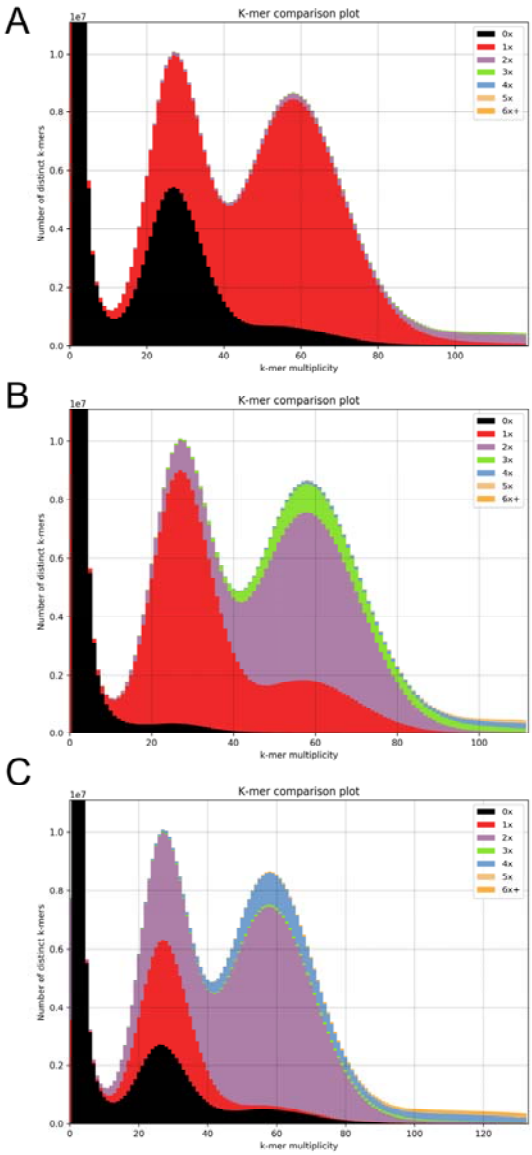
No competing interests declared.

Acknowledgments

We acknowledge the Functional Genomic Centre Zurich, Switzerland, especially Dr. Lucy Poveda, Andrea Patrignani and Dr. Catharine Aquino-Fournier for the support in generating the genome sequencing data, the Fruit-Growing Extension Group of Agroscope, Switzerland, for taking care of the plant material and the Method Development and Analytics Group of Agroscope, Switzerland, for sharing laboratory infrastructures. Part of the analyses were performed on the CyVerse cyberinfrastructure, which was supported by the National Science Foundation under Award Numbers DBI-0735191, DBI-1265383, and DBI-1743442. URL: www.cyverse.org

Figures:

Figure 1: K-mers spectra generated using k-mer analysis toolkit (23-mers) with the Illumina reads of 'Gala Galaxy' against the double-haploid GDDH13 reference genome (A), the diploid genome reference MDGGdi_v1.0 (B) and the graph-based diploid phased assembly generated with WhatsHap (C). The overall curve shape indicates the frequency of the 23-mers in the original Illumina data (apple cv. 'Gala Galaxy'), while the colored area indicate if the 23-mers are found once (red), twice (violet) or more often (other colors) in the reference genome. The black area indicate reads that are present in the Illumina sequencing data but absent in reference assemblies.



233 **Tables:**

234 Table 1: Statistics of the different assemblies generated in this work.

	Falcon Unzip	**Two-enzymes hybrid assembly**	**MDGGph_v1.0 (primary haploid assembly)**	**MDGGsh_v1.0 (secondary haploid assembly)**	**MDGGdi_v1.0 (diploid assembly)**
Total number of contigs/scaffolds	2,061 primary contigs and 5,663 haplotigs	154 hybrid scaffolds and 6336 unscaffolded contigs	31 hybrid scaffolds in 17 pseudomolecules	3,662 hybrid scaffolds and contigs in 17 pseudomolecules and 2583 unanchored contigs	34 pseudomolecules from MDGGph_v1.0 and MDGGsh_v1.0 and 2583 unanchored contigs
Total length	1,308.969 Mb	1,414.649 Mb (999.96 Mb in scaffolds + 414.68 Mb unscaffolded	719.450 Mb	601.426 Mb + 80.621 Mb of unanchored sequences	1,401.497 Mb
N50 length	0.585 Mb	13.704 Mb (34.286 Mb for hybrid scaffolds only)			
Ns			29,755,402 bp	70,500,027 bp	100,255,429 bp

235

References:

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ 1990. Basic Local Alignment Search Tool. Journal of Molecular Biology 215: 403-410. doi: Doi 10.1016/S0022-2836(05)80360-2

Baumgartner IO, et al. 2016. Development of SNP-based assays for disease resistance and fruit quality traits in apple (Malus x domestica Borkh.) and validation in breeding pilot studies. Tree Genetics & Genomes 12. doi: 10.1007/s11295-016-0994-y

Bianco L, et al. 2016. Development and validation of the Axiom (R) Apple480K SNP genotyping array. Plant Journal 86: 62-74. doi: 10.1111/tpj.13145

Bianco L, et al. 2014. Development and Validation of a 20K Single Nucleotide Polymorphism (SNP) Whole Genome Genotyping Array for Apple (Malus x domestica Borkh). Plos One 9. doi: 10.1371/journal.pone.0110377

Chen S, Zhou Y, Chen Y, Gu J 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 34: i884-i890. doi: 10.1093/bioinformatics/bty560

Chin CS, et al. 2016. Phased diploid genome assembly with single-molecule real-time sequencing. Nature Methods 13: 1050-+. doi: 10.1038/Nmeth.4035

Daccord N, et al. 2017. High-quality de novo assembly of the apple genome and methylome dynamics of early fruit development. Nature Genetics 49: 1099-+. doi: 10.1038/ng.3886

Di Pierro EA, et al. 2016. A high-density, multi-parental SNP genetic map on apple validates a new mapping approach for outcrossing species. Horticulture Research 3. doi:10.1038/hortres.2016.57

Dickinson JP, White AG 1986. Red Color Distribution in the Skin of Gala Apple and Some of Its Sports. New Zealand Journal of Agricultural Research 29: 695-698.

El-Sharkawy I, Liang D, Xu KN 2015. Transcriptome analysis of an apple (Malus x domestica) yellow fruit somatic mutation identifies a gene network module highly associated with anthocyanin and epigenetic regulation. Journal of Experimental Botany 66: 7359-7376. doi: 10.1093/jxb/erv433

263 FAO statistical yearbook [Internet]. Rome: Food and Agriculture Organization of the United
264 Nations; 2004 9.1.2020]. Available from: http://www.fao.org/faostat/en/#data
265 Foster TM, Aranzana MJ 2018. Attention sports fans! The far-reaching contributions of bud
266 sport mutants to horticulture and plant biology. Horticulture Research 5.
267 doi:10.1038/s41438-018-0062-x

268 Gel B, Serra E 2017. karyoploteR: an R/Bioconductor package to plot customizable genomes
269 displaying arbitrary data. Bioinformatics 33: 3088-3090. doi:
270 10.1093/bioinformatics/btx346
271 Haug-Baltzell A, Stephens SA, Davey S, Scheidegger CE, Lyons E 2017. SynMap2 and
272 SynMap3D: web-based whole-genome synteny browsers. Bioinformatics 33: 2197-2198.
273 doi: 10.1093/bioinformatics/btx144
274 Iglesias I, Echeverria G, Soria Y 2008. Differences in fruit colour development, anthocyanin
275 content, fruit quality and consumer acceptability of eight 'Gala' apple strains. Scientia
276 Horticulturae 119: 32-40. doi: 10.1016/j.scienta.2008.07.004
277 Knorst V, et al. 2019. First assembly of the gene-space of Lolium multiflorum and
278 comparison to other Poaceae genomes. Grassland Science 65: 125-134. doi:
279 10.1111/grs.12225
280 Kost TD, et al. 2015. Development of the First Cisgenic Apple with Increased Resistance to
281 Fire Blight. Plos One 10. doi:10.1371/journal.pone.0143980

282 Kumar S, et al. 2012. Genomic Selection for Fruit Quality Traits in Apple (Malus x domestica
283 Borkh.). Plos One 7. doi: 10.1371/journal.pone.0036674

284 Langmead B 2010. Aligning short sequencing reads with Bowtie. Curr Protoc Bioinformatics
285 Chapter 11: Unit 11 17. doi: 10.1002/0471250953.bi1107s32
286 Laurens F, et al. 2018. An integrated approach for increasing breeding efficiency in apple
287 and peach in Europe. Horticulture Research 5. doi: 10.1038/s41438-018-0016-3

288 Lee HS, et al. 2016. Analysis of 'Fuji' apple somatic variants from next-generation
289 sequencing. Genetics and Molecular Research 15. doi: UNSP 15038185

10.4238/gmr.15038185

Li XW, et al. 2016. Improved hybrid de novo genome assembly of domesticated apple (Malus x domestica). Gigascience 5. doi: 10.1186/s13742-016-0139-0

Mapleson D, Garcia Accinelli G, Kettleborough G, Wright J, Clavijo BJ 2016. KAT: a K-mer analysis toolkit to quality control NGS datasets and genome assemblies. Bioinformatics 33: 574-576.

Marçais G, Kingsford C 2011. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics 27: 764-770.

Muranty H, et al. 2015. Accuracy and responses of genomic selection on key traits in apple breeding. Horticulture Research 2. doi:10.1038/hortres.2015.60

Peace CP, et al. 2019. Apple whole genome sequences: recent advances and new prospects. Horticulture Research 6. doi:10.1038/s41438-019-0141-7

Pryszcz LP, Gabaldon T 2016. Redundans: an assembly pipeline for highly heterozygous genomes. Nucleic Acids Research 44. doi:10.1093/nar/gkw294

Roth M, et al. 2019. Prediction of fruit texture with training population optimization for efficient genomic selection in apple. bioRxiv: 862193. doi: 10.1101/862193

Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31: 3210-3212. doi: 10.1093/bioinformatics/btv351

Tang HB, et al. 2015. ALLMAPS: robust scaffold ordering based on multiple maps. Genome Biology 16. doi:10.1186/s13059-014-0573-1

Trapnell C, et al. 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nature Protocols 7: 562-578. doi: 10.1038/nprot.2012.016

Turechek WW. 2004. Apple diseases and their management. In. Diseases of Fruits and Vegetables Volume I: Springer. p. 1-108.

316 Velasco R, et al. 2010. The genome of the domesticated apple (Malus x domestica Borkh.).
317 Nature Genetics 42: 833-+. doi: 10.1038/ng.654
318 Zhang L, et al. 2019. A high-quality apple genome assembly reveals the association of a
319 retrotransposon and red fruit colour. Nature Communications 10: 1494. doi:
320 10.1038/s41467-019-09518-x