PCP
PLANT & CELL PHYSIOLOGY

PeachVar-DB: A Curated Collection of Genetic Variations for the Interactive Analysis of Peach Genome Data

Marco Cirilli[1,7], Tiziano Flati[2,3,7], Silvia Gioiosa[2,3], Ilario Tagliaferri[2], Angelo Ciacciulli[1], Zhongshan Gao[4], Stefano Gattolin[5], Filippo Geuna[1], Francesco Maggi[6], Paolo Bottoni[6], Laura Rossini[1], Daniele Bassi[1,*], Tiziana Castrignanò[2,*] and Giovanni Chillemi[2]

[1]Department of Agricultural Science (DISAA), University of Milan, Milan, Italy
[2]Cineca, HPC High Performance Computing Department, Rome, Italy
[3]IBIOM-CNR, Institute of Biomembranes, Bioenergetics and Molecular Biotechnologies, Bari, Italy
[4]Department of Horticulture, College of Agriculture and Biotechnology, Zhejiang University, 310058, Hangzhou, China
[5]Parco Tecnologico Padano, Via Einstein, Loc. C.na Codazza, Lodi, Italy
[6]Department of Computer Science, 'Sapienza' University of Rome, Via Salaria 113, 00198 Rome, Italy
[7]These authors contributed equally to this work
*Corresponding authors: Tiziana Castrignanò, E-mail: t.castrignano@cineca.it; Daniele Bassi, E-mail: daniele.bassi@unimi.it
(Received September 1, 2017; Accepted November 16, 2017)

Applying next-generation sequencing (NGS) technologies to species of agricultural interest has the potential to accelerate the understanding and exploration of genetic resources. The storage, availability and maintenance of huge quantities of NGS-generated data remains a major challenge. The PeachVar-DB portal, available at http://hpc-bioinformatics.cineca.it/peach, is an open-source catalog of genetic variants present in peach (*Prunus persica* L. Batsch) and wild-related species of *Prunus* genera, annotated from 146 samples publicly released on the Sequence Read Archive (SRA). We designed a user-friendly web-based interface of the database, providing search tools to retrieve single nucleotide polymorphism (SNP) and InDel variants, along with useful statistics and information. PeachVar-DB results are linked to the Genome Database for Rosaceae (GDR) and the Phytozome database to allow easy access to other external useful plant-oriented resources. In order to extend the genetic diversity covered by the PeachVar-DB further, and to allow increasingly powerful comparative analysis, we will progressively integrate newly released data.

Keywords: Peach • Genomic variants • Database • NGS • Genome resequencing.

Abbreviations: **GDR**, Genome Database for Rosaceae; **NGS**, next-generation sequencing; **SNP**, single nucleotide polymorphism; **SRA**, Sequence Read Archive; **WGRS**, whole-genome re-sequencing.

Introduction

Peach (*Prunus persica* L. Batsch) is the most economically important fruit tree species of *Prunus* genera. According to archaeological evidence, peach was domesticated in China from an unknown ancestor and is related to other wild species of the *Amygdalus* subgenus, including *P. mira*, *P. davidiana* and *P. kansuensis* (Faust and Timon 1995). Wild-relatives bear fruits of poor eating quality, although they could be a valuable source either for the introgression of disease resistance traits or as rootstocks (Pascal et al. 1998).

Intense breeding activities in peach allowed a progressive improvement of important fruit quality characteristics and technological attributes (Infante et al. 2008, Monet and Bassi 2008). Peach genetics also achieved remarkable progress, developing a wide number and types of molecular markers, allowing several linkage maps to be built and the mapping of important traits, as well as the estimation of genetic diversity and domestication paths (Byrne et al. 2012, Li et al. 2013, Akagi et al. 2016). However, relevant quantitative characters, particularly those related to fruit nutritional properties, plant environmental adaptation and disease resistance, have a complex pattern of inheritance, so the understanding of the molecular genetic bases of such traits requires a more detailed knowledge of the gene network expressing the phenotype (Foulongne et al. 2003, Cirilli et al. 2016).

The relatively small size of its genome, estimated at about 265×10^6 bp, as well as the collinearity with other diploid *Prunus* species, makes peach an ideal model for comparative and functional genomics within the Rosaceae family (Abbott et al. 2002). The completion of the peach genome project made available a high-quality reference assembly of the double-haploid cultivar 'Lovell' (Verde et al. 2013, Verde et al. 2017), enabling high-throughput discovery of genetic markers through the application of next-generation sequencing (NGS) technologies. The whole-genome re-sequencing (WGRS) approach has the potential to reveal most of, if not all, the genomic variations in a target individual in comparison with the reference genome (Mochida and Shinozaki 2010). Sequencing efforts on peach and wild relatives are rapidly growing in light of the potential usefulness of high-density genetic information for germplasm characterization, functional genomics and evolutionary studies (Ahmad et al. 2011, Cao et al. 2016, Velasco et al. 2016).

Given their potential to overcome the well-known limited resolution of the biparental linkage mapping approach, genome-wide association studies are becoming a popular and effective tool for dissecting the genetic architecture of monogenic or polygenic traits in peach (Cao et al. 2014, Micheletti et al. 2015).

Plant Cell Physiol. 59(1): e2(1–9) (2018) doi:10.1093/pcp/pcx183, Advance Access publication on 4 December 2017, available online at www.pcp.oxfordjournals.org

Also, dense marker information is essential for the adoption of marker-assisted breeding and genome selection, which appears among the most promising strategies to boost breeding programs and trait introgression, as recently shown with peach varieties (Biscarini et al. 2017). The reliability and effectiveness of these novel genomics tools will largely depend on the continuous sharing and integration of massive amounts of genetic and phenotypic data coming from diversified sources (Kang et al. 2016). The optimal management of high-throughput biological information requires the implementation of dedicated bioinformatics facilities.

For this reason, integrative databases were built for several model and non-model species, such as TAIR, Gramene, OryzaBase and the SOL Genomics Network (Mochida and Shinozaki 2010). An important central repository, the Genome Database for Rosaceae (GDR), is available for peach and other species of the *Rosaceae* family (Jung et al. 2013), but it does not currently host high-density genomic data or the relative annotations obtained from re-sequencing projects of a massive number of accessions. Therefore, whole-genome information about SNPs (single nucleotide polymorphisms) and InDels (small insertions and deletions) is not easily accessible, remaining de facto a powerful but largely unused resource. The availability of an interactive database to explore genomic variability in peach would facilitate research studies, allowing access to information regarding genetic variants even to scientists who have no access to bioinformatics platforms for data analysis.

The present article introduces the first release of PeachVar-DB, an open-source catalog of annotated genomic variants (SNP and InDels) found both in peach (*P. persica* L. Batsch) and wild-related species of *Prunus* genera. Variant discovery was achieved by applying an imputation-free joint variant-calling procedure on 146 accessions publicly available on the Sequence Read Archive (SRA). Such a procedure improves variant discovery by leveraging population-wide information from a cohort of multiple samples (Liu et al. 2013). Our objective is to provide easy access to information from WGRS for a range of purposes, from genotyping to functional genomics. The web interface of the database provides several tools to search, visualize and download results, along with other useful annotation statistics and information.

Results

Database content and web access

Whole-genome sequencing data of 125 peach (*P. persica* L. Batsch) accessions and 21 wild relatives of the *Amygdalus* subgenus have been downloaded from the NCBI SRA (Leinonen et al. 2011). Following identification of genetic variants for each accession (see the Materials and Methods), data were stored in PeachVar-DB, a NoSQL graph database (Neo4j).

The current database release contains a total of 4,630,814 and 461,785 high-quality SNPs and InDels, respectively, with 42.8% SNPs and 42.5% InDels being exclusively present in wild relatives. About 11% of the total number of SNPs have a minor allele frequency of <0.05, and 5.1% are located in exonic regions (**Table 1**). The PeachVar-DB web portal provides a user-friendly interface and is available at the http://hpc-bioinformatics.cineca.it/peach. The home page contains a list of all the input accessions with links to the SRA (**Fig. 1A**), an interactive pie chart of the number of variants per accession (**Fig. 1B**), a variant summary table by algorithm (**Fig. 1D**) and an overall summary of accessions and variants found (**Fig. 1C**).

Table 1 Number of genetic variants in the Peach-DB database, grouped by type and genomic features

	Variant type	
	SNP	InDel
	4,630,814	461,785
Region (%)		
Intergenic	16.0	15.2
Exon	5.1	0.9
Intron	13.1	14.6
5′-UTR	1.8	2.1
3′-UTR	2.6	3.2
Upstream	30.8	31.5
Downstream	30.1	31.7

UTR, untranslated region.

Search tools

The 'Search' menu provides several tools to explore genomic variants in PeachVar-DB (**Fig. 2**). All the query output can be easily downloaded in both text and VCF format, allowing users to perform downstream analysis on a customized subset of variants. Whenever a genetic variant present in PeachVar-DB falls into a known genetic marker interval, the output table displays a button named 'Genetic Marker'.

The 'Search by region' function (**Fig. 2A**) runs the query on the genomic co-ordinates specified by the user. Furthermore, the selected genomic windows are hyperlinked to the 'Synteny Viewer' tool available under GDR, allowing the visualization of genome conservation and rearrangement patterns. Users interested in the variants occurring on a specific gene can use the 'Search by gene ID' function (**Fig. 2B**) to retrieve this information. Genes can be queried either by typing their name (based on the JGI nomenclature for peach genome v2.1 transcript annotations) or through the provided full list of transcripts. The results can be visualized via JBrowse and are linked to a detailed functional annotation description in Phytozome (Goodstein et al., 2012). Additionally, Gene Ontology identifiers, annotations and descriptions are displayed for the selected gene by clicking on the 'see GOterms' button. In addition, the search can be restricted to specific gene regions (e.g. CDS, untranslated regions or mRNA sequence), using the 'Search by gene features' function (**Fig. 2C**). Users interested in accession-specific variants can run the 'Search by accession' function (**Fig. 2D**). The results table reports an extra column indicating whether the listed variants are common to other accessions as well. Users can also directly paste custom nucleotide sequences, retrieving BLAST results as full-text alignment through the 'Search by similarity' function (**Fig. 2E**). Finally, the 'Pairwise accession comparison' function (**Fig. 2F**) gives the user the ability to

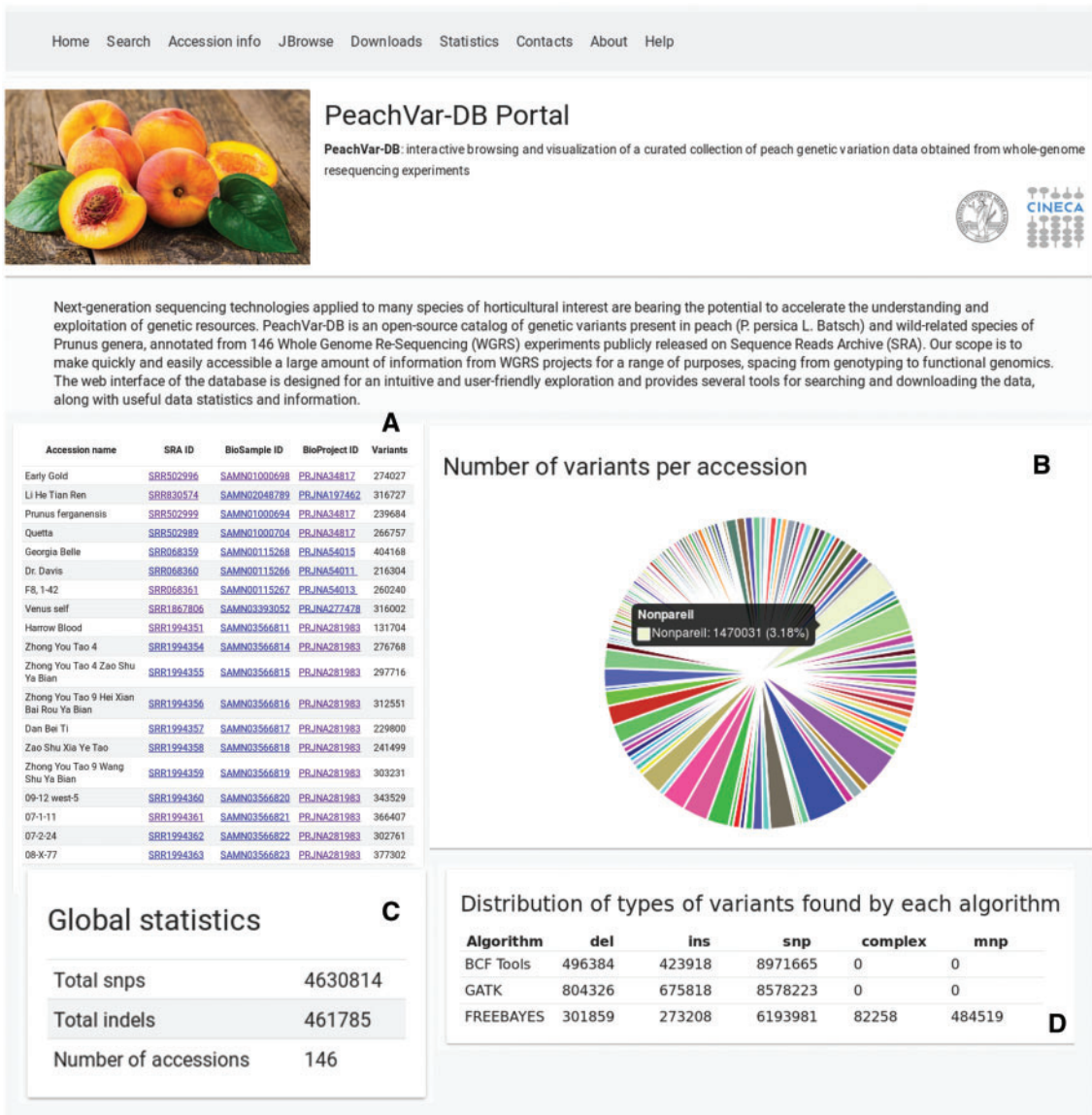

Fig. 1 Overview of the PeachVar-DB portal homepage. (A) The accessions included in the input dataset, each one linked to SRA through its SRA/BioProject/BioSample ID. (B) An interactive pie chart, giving an overview of the number of genetic variants identified for each accession. Two summary tables (C and D) show the genetic variants detected by applying three variant-calling algorithms.

make direct genotype comparisons between two selected accessions.

Accession info

The 'Accession info' link gives access to four menu options named 'dataset information', 'population structure', 'phylogenesis' and 'phenotypic data' (**Fig. 3**). The 'dataset information' page provides an overview of the input files (e.g. SRA accession names and Bioproject hyperlinks) subdivided into two categories: *P. persica* and wild relatives (**Fig. 3A**). The 'population structure' page describes the population structure of the 146 accessions inferred using a subset of 50,000 randomly selected SNPs in ADMIXTURE v1.22 (Alexander et al. 2009). A value of $K = 4$ explains most of the ancestry within accessions. It differentiates the cluster of wild relatives (P4) and ornamental peach (P2) from edible peach (P1 and P3), the latter showing various degrees of admixture. For higher values of K, landraces and accessions derived from Oriental and Occidental breedings tend to separate. Accession names and membership probabilities can be viewed in pop-up windows by hovering over the histograms (**Fig. 3B**). The 'phylogenesis' page displays the genetic relationships estimated using SNPhylo (Lee et al. 2014) and PhyML 3.0 (Guindon et al. 2010) using an interactive phylogenetic tree. By clicking on the branch of interest, a pop-up menu allows users to apply several display modification functions (e.g. subtree collapsing; descendent/internal/incident branches visualization; path to root selection as shown in **Fig. 3C**). 'Phenotypic data' encloses a range of typical peach Mendelian traits, mainly related to quality attributes, such as fruit hairiness (peach or nectarine), shape (round or flat), texture (melting,

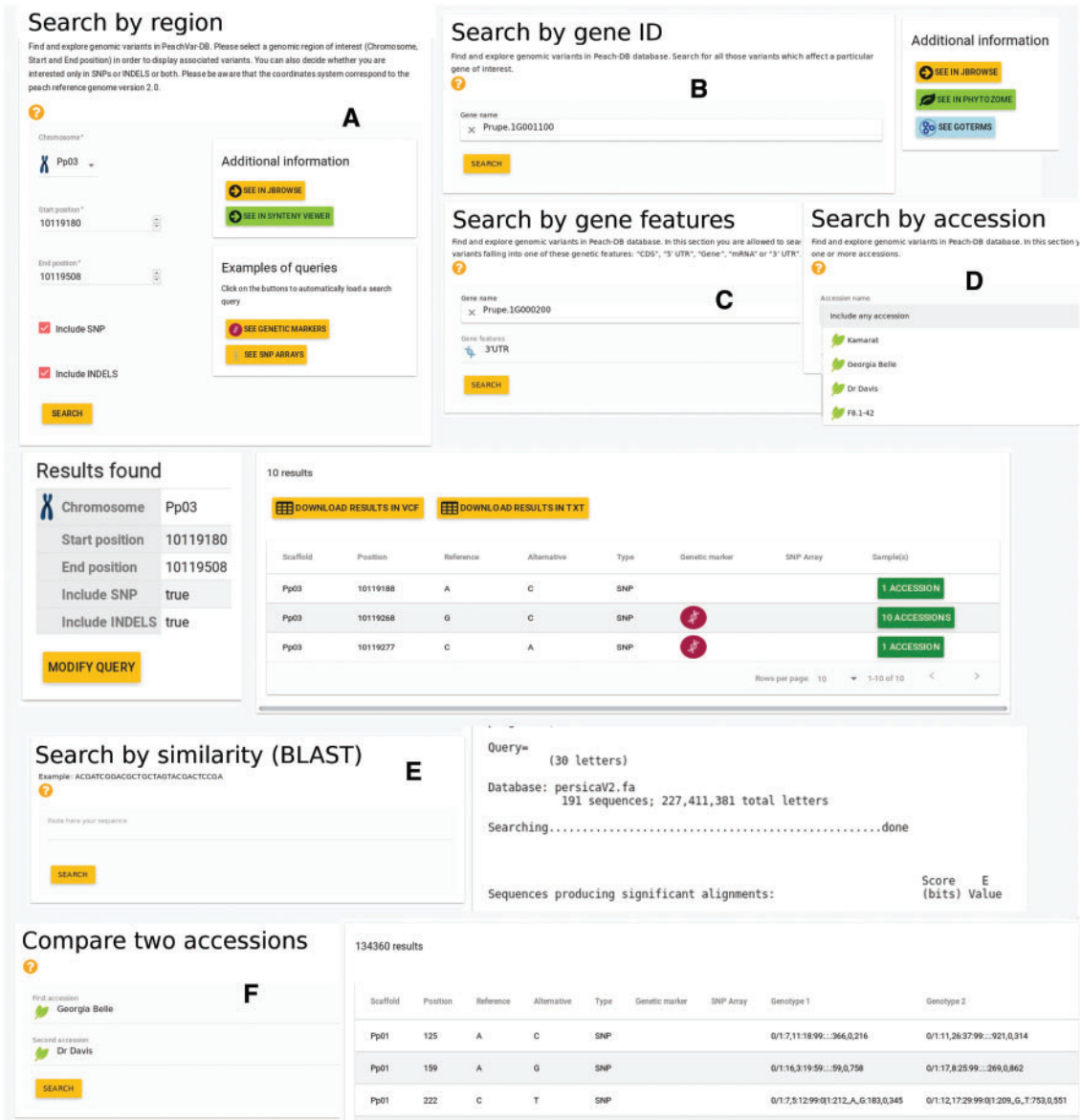

Fig. 2 Examples of search functions implemented in PeachVar-DB: 'Search by region' (A), allowing delimitation of the queries to specific genomic co-ordinates; 'Search by gene ID' (B) allowing the search in a single gene; 'Search by gene feature' (C), allowing the search of variants in specific regions (CDS, intron); 'Search by accession' (D), allowing the search of variants in a single accession; 'Search by similarity' (E), allowing BLAST search against the peach genome; 'Pairwise accession comparison' (F) allowing retrieval of genetic variants in common between two accessions.

non-melting and stony hard), flesh color (yellow or white) and taste (acid or sub-acid). The dynamical pie charts provide an overview of the phenotypic variability in peach fruit (**Fig. 3D**).

The JBrowse Genome Browser

PeachVar-DB integrates the Ajax-based genome browser JBrowse v1.11.3 (Skinner et al. 2009) for an easy to use panning and zooming navigation of the peach reference genome (**Fig. 4**). In addition to gene models and putative orthologous genes from other species (*Arabidopsis thaliana*, *Prunus avium* and *Oryza sativa*), JBrowse allows the direct visualization of the location of SNPs/InDels. After selecting a variant of interest, users can visualize the complete list of annotations, such as allele frequency and distribution, missing data or assembling statistics of each accession. Moreover, through the native 'Open track file' tool, user-provided files, such as VCF or BAM, can be privately visualized and explored using the reference genome features present in the JBrowse environment. To increase usability, we added the latest release of *Prunus* genetic markers mapped to the peach genome v2.0.a1 as an additional track.

Statistics

This page reports a range of useful statistics (see **Fig. 5**). Information regarding whole-genome linkage disequilibrium patterns, SNP density (**Fig. 5A, B**), gene density (**Fig. 5C**), nucleotide diversity (**Fig. 5D**) and mean read depth, among others, are displayed. For example, as shown in **Fig. 5A**,

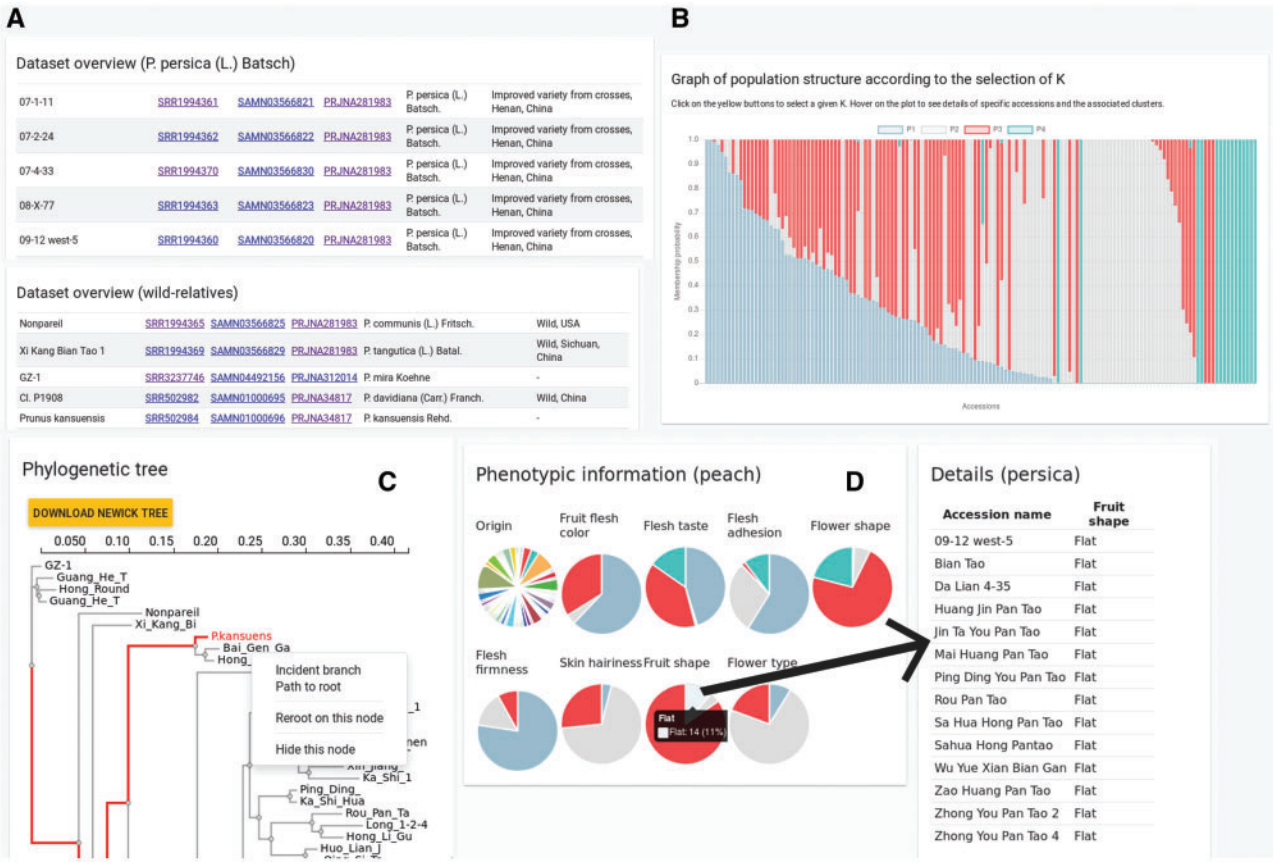

Fig. 3 An overview of the input dataset (A); population structure estimated for different values of *K* (B); phylogenetic relationships among accessions (C); phenotypic traits described with interactive pie charts (D).

responsive histograms provide more detail as the size of the display window increases (e.g. compare the two histograms in **Fig. 5A and B** in which the same information is dynamically displayed by using two different window sizes, about 2 and 7 million positions, respectively). Whenever a specific accession is selected, a dynamic annotated genetic variants summary is displayed at the top of the page (**Fig. 5E**).

Conclusions and Future Developments

Tremendous progress of -omics approaches is enabling high-throughput generation of biological data, opening the so called bio-data era. The development of bioinformatics facilities to manage such massive amounts of information is becoming a fundamental aspect of research, particularly in the field of functional genomics and population genetics. Integrative databases are already available for several species, including also those specifically intended for hosting high-density genome data.

PeachVar-DB is the first database specifically dedicated to the storing, sharing and querying of peach genomic variations identified by the analysis of WGRS data from a panel of peach and wild relative accessions. The final results presented in this database rely on the identification of a highly reliable and accurate consensus call-set, consisting of >5 millions peach genetic variants, detected by merging the results of multiple variant-calling algorithms. Genomic variants can be quickly and easily explored, visualized and retrieved by using a well-assorted set of tools ranging from BLAST similarity search to the JBrowse Genome Browser. In turn, this information can be used for a range of purposes, including germplasm characterization, SNP array design and experimental design for functional and evolutionary genomics.

As the amount of NGS data freely available in peach is expected to grow, we designed the PeachVar-DB portal architecture to support new data updates and the integration of new tools. The dataset of accessions and genomic variants will be updated at least every 2 years, depending on the amount of whole-genome sequencing data publicly released. Other aspects will also be improved, including hyperlinks to other important reference communities, particularly EVA (European Variations Archive).

The linking of genomic data with phenotypic traits remains one of the major challenges for crop improvement. In this perspective, the inclusion of phenotypic information has not been dictated solely by the need for a basic description of the input dataset, but rather it represents a first step towards the creation of a more comprehensive database devoted to functional genomics, which will be able to handle and analyze a plethora of both genomic and phenomic data.

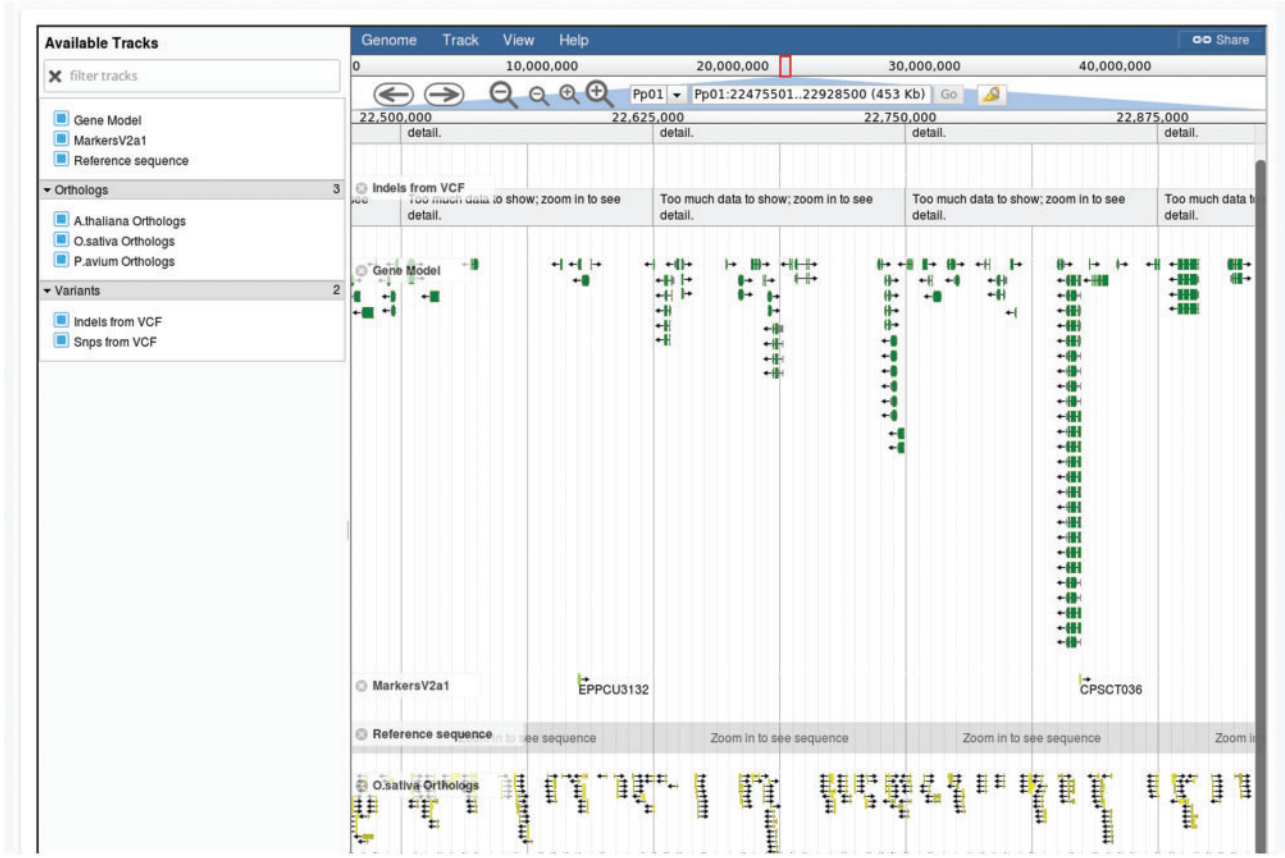

Fig. 4 The JBrowse interface, the genome browser implemented in PeachVar-DB.

Materials and Methods

Data retrieving and processing

Data were downloaded from the NCBI SRA, selecting only whole-genome sequencing libraries (125 peach accessions, *P. persica* L. Batsch, and 21 wild relatives of the *Amygdalus* subgenus), and excluding microRNA and RNaseq libraries or DNA samples derived from chromatin immunoprecipitation or other DNA treatments. The whole set of accessions is largely representative of the genetic variability present in peach germplasm, including ornamental varieties, landrace and improved accessions derived from oriental and occidental breeding. Wild relatives include accessions of *P. davidiana*, *P. mira*, *P. kansuensis*, *P. communis* (almond) and *P. ferganensis*, the latter being genetically similar to peach. A detailed list of accession numbers and relative hyperlinks is provided both in the homepage and in the 'Dataset information' page, accessible from the 'Accession info' menu. The full set of sequences provides an overall 1,601× coverage of the estimated peach genome size (227 Mbp), with a mean coverage for each accession varying from 2.65× to 109.93×. Raw reads were quality filtered, trimmed with Trimmomatic v. 0.32 (Bolger et al. 2014) and mapped onto the peach reference 'Lovell' genome V2.1 using the Burrows–Wheeler Aligner (BWA)-MEM algorithm with default parameters. After duplicate removal and indexing of mapped reads with Picard tools (http://broadinstitute.github.io/picard/), genomic variants were identified by executing three different variant-calling algorithms: GATK-Haplotype Caller (Van der Auwera et al. 2013), BCFtools v. 1.2 and Freebayes (Garrison and Marth, 2012). SNP and InDel discovery and genotyping were performed across all 146 samples simultaneously, with a joint-calling procedure, according to each algorithm's best practices. After filtering for low-quality variants, an intersection of results was performed using a custom Perl script, retaining only those in common to all the algorithms in order to get a more robust set of results (**Fig. 6**). Once obtained, the final joint VCF file has been split into individual samples, zipped and indexed with the bgzip utility from samtools suite (http://www.htslib.org/doc/tabix.html). Users can freely download all the above files via the 'Download' page of the PeachVar-DB portal.

VCF downstream analysis

The evolutionary relationships among accessions stored in the database were inferred using SNPhylo (Lee et al. 2014) for whole-genome SNP data filtering. Multiple sequence alignment was exerted with MUSCLE software (Edgar 2004), and PhyML 3.0 (Guindon et al. 2010) was used for estimating phylogenetic relationships through the maximum likelihood method and bootstrapping (100). Population structure was estimated through ADMIXTURE software v. 1.22 (Alexander et al. 2009) using 50,000 randomly selected SNPs and by inputting *K* values from 2 to 8 to identify the value of *K* (a priori genetic clusters) that maximizes the predictive accuracy based on a 10-fold cross-validation with 10 different fixed initial seeds. The 'compare two accessions' utility relies on the runtime execution of bedtools software (Quinlan and Hall 2010), showing only variants in common between two accessions of choice. The information shown under the 'Statistics' page has been pre-calculated using: the vcf-tools utility (Danecek et al. 2011) for SNP density, linkage disequilibrium and nucleotide diversity; samtools mpileup (Li et al. 2009, Li 2011) for 'Mean Depth'; and BCFtools (http://github.com/samtools/bcftools) for 'Gene density by region' and 'Gene density by chromosome'.

Web architecture

PeachVar-DB is hosted on a Ubuntu 16.04 Linux operating system with the Apache 2.4.18 Web server. The front-end web-user interface was developed using the AngularJS framework (v. 1.6.4) which presents data dynamically within a single-page application (SPA). The graphical user interface was implemented using the Angular material library (v. 1.1.4), a reference implementation of Google's Material Design Specification, which provides a set of reusable, well-tested and accessible UI components for an optimal final user experience.

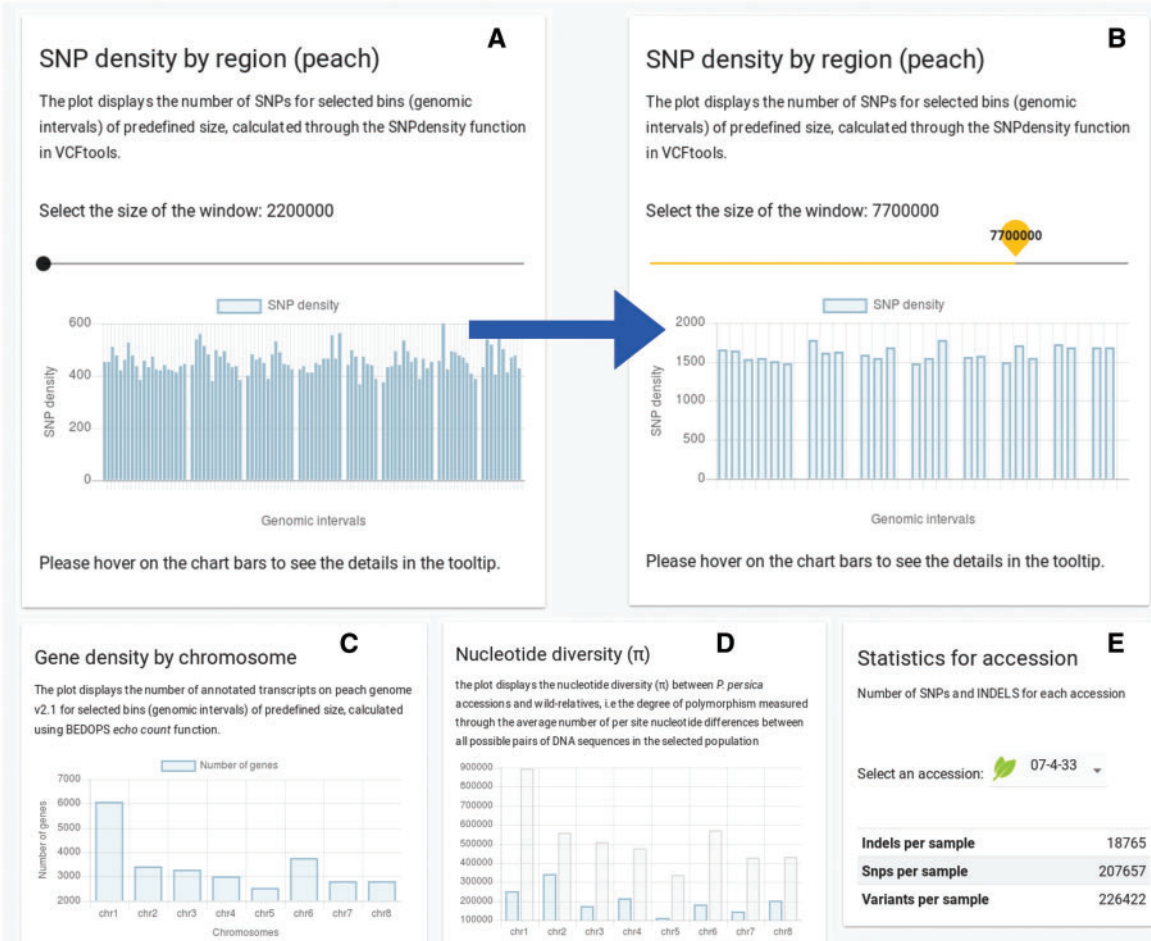

Fig. 5 The PeachVar-DB portal page reporting several data statistics. (A) The number of SNPs with a fixed size of about 2 million bases. By changing the value of the window size by means of a custom slider, the same information is plotted at a different resolution (i.e. about 7 million bases, see B). Information regarding gene density (C) and nucleotide diversity (D), among others, is also shown. A dynamical summary table of the number of annotated genetic variants is finally reported in (E), whenever a specific accession is selected.

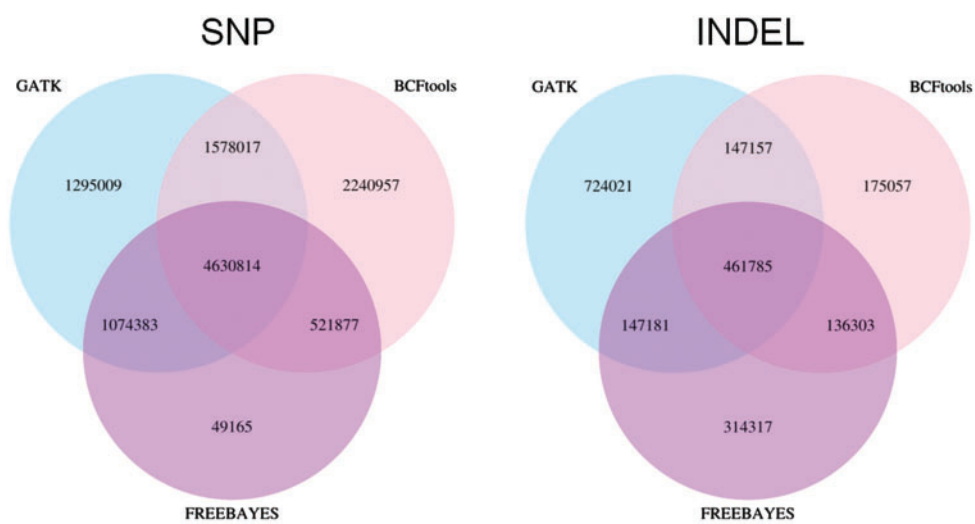

Fig. 6 Venn diagrams of genomic variants identified by each variant-calling algorithm (GATK-HC, BCFtools and Freebayes) and their intersection. Variants which were in common to two out of the three algorithms were not retained.

The front-end also relies on third-party graphical libraries, such as d3 (https://github.com/d3/d3), which allows easy display of custom graphic components (e.g., phylogenetic trees, Venn diagrams, etc.), angular-chart (http://jtblin.github.io/angular-chart.js/), which effectively renders data according to several chart types, and md-data-table (https://github.com/iamisti/mdDataTable) for visualization and pagination of search results. The front-end interface uses AJAX to retrieve data from a web server; domain data are stored in a huge graph-based database.

JBrowse additional tracks

Additional tracks displayed in the JBrowse section were downloaded from the following sources: *A. thaliana* and *O. sativa* from Phytozome database, *P. avium* and peach genetic markers from the GDR database.

Funding

This work was funded by the University of Milan [under a Transition Grant 2015/2017 to D.B. and a post-doctoral fellowship to M.C.].

Acknowledgments

The authors would like to thank Ms. Antonella Pintus for constructive criticism of the manuscript and also would like to thank the LISA initiative 2016–2018 [http://www.hpc.cineca.it/content/lisa-call] and the Italian Node of Elixir project [http://elixir-italy.org].

Disclosures

The authors have no conflicts of interest to declare.

References

Abbott, A., Georgi, L., Yvergniaux, D., Inigo, M., Sosinski, B., Wang, Y., et al. (2002) Peach: the model genome for rosaceae. *Acta Hortic.*1: 145–156.

Ahmad, R., Parfitt, D.E., Fass, J., Ogundiwin, E., Dhingra, A., Gradziel, T.M., et al. (2011) Whole-genome sequencing of peach (*Prunus persica* L.) for SNP identification and selection. *BMC Genomics* 12: 569.

Akagi, T., Hanada, T., Yaegaki, H., Gradziel, T.M., Tao, R. (2016) Genome-wide view of genetic diversity reveals paths of selection and cultivar differentiation in peach domestication. *DNA Res*. 23: 271–282.

Alexander, D.H., Novembre, J. and Lange, K. (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome Res*. 19: 1655–1664.

Biscarini, F., Nazzicari, N., Bink, M., Arús, P., Aranzana, M.J., Verde, I., et al. (2017) Genome-enabled predictions for fruit weight and quality from repeated records in European peach progenies. *BMC Genomics* 18: 432.

Bolger, A.M., Lohse, M. and Usadel, B. (2014)Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*;30: 2114–2120.

Byrne, D.H., Raseira, M.B., Bassi, D., Piagnani, M.C., Gasik, K., Reighard, G.L., et al. (2012) Peach. *In* Fruit Breeding. Edited by Badenes, M.L. and Byrne, D.H. pp. 505–506. Springer, New York.

Cao, K., Zheng, Z., Wang, L., Liu, X., Zhu, G., Fang, W., et al. (2014) Comparative population genomics reveals the domestication history of the peach, *Prunus persica*, and human influences on perennial fruit crops. *Genome Biol.* 15: 415.

Cao, K., Zhou, Z., Wang, Q., Guo, J., Zhao, P., Zhu, G., et al. (2016) Genome-wide association study of 12 agronomic traits in peach. *Nat. Commun*. 7: 13246.

Cirilli, M., Geuna, F., Babini, A.R., Bozhkova, V., Catalano, L., Cavagna, B., et al. (2016) Fighting sharka in peach: current limitations and future perspectives. *Front. Plant Sci*. 7: 1290.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., et al. (2011) The variant call format and VCFtools. *Bioinformatics* 27: 2156–2158.

Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32: 1792–1797.

Faust, M. and Timon, B. (1995) Origin and dissemination of peach. *Hort. Rev*. 17: 331–379.

Foulongne, M., Pascal, T., Arús, P. and Kervella, J. (2003) The potential of *Prunus davidiana* for introgression into peach [*Prunus persica* (L.) Batsch] assessed by comparative mapping. *Theor. Appl. Genet*. 107: 227–238.

Garrison, E. and Marth, G. (2012) Haplotype-based variant detection from short-read sequencing. *arXiv* preprint arXiv: 1207.3907 [q-bio.GN].

Goodstein, D.M., Shu, S., Howson, R., Neupane, R., Hayes, R.D., Fazo, J., et al. (2012) Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res*. 40: D1178–D1186.

Guindon, S., Dufayard, J.F., Lefort, V., Anisimova ,M., Hordijk, W. and Gascuel, O. (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol*. 59: 307–321.

Infante, R., Martínez Gómez, P. and Predieri, S. (2008) Quality oriented fruit breeding: peach [*Prunus persica* (L.) batsch]. *J. Food Agric. Environ*. 6: 342–356.

Jung, S., Ficklin, S.P., Lee ,T., Cheng, C.H., Blenda, A., Zheng, P., et al. (2013) The Genome Database for Rosaceae (GDR): year 10 update. *Nucleic Acids Res*. 42: D1237–D1244.

Kang, Y.J., Lee, T., Lee, J., Shim, S., Jeong, H., Satyawan, D., et al. (2016) Translational genomics for plant breeding with the genome sequence explosion. *Plant Biotechnol. J*. 14: 1057–1069.

Lee, T.H., Guo, H., Wang, X., Kim, C. and Paterson, A.H. (2014) SNPhylo: a pipeline to construct a phylogenetic tree from huge SNP data. *BMC Genomics* 15: 162.

Leinonen, R., Sugawara, H., Shumway, M.; International Nucleotide Sequence Database Collaboration. (2011) The Sequence Read Archive. *Nucleic Acids Res*. 39: D19–D21.

Li, H. (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 27: 2987–2993.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009) The Sequence Alignment/Map (SAM) format and SAMtools. *Bioinformatics* 25: 2078–2079.

Li, X.W., Meng, X.Q., Jia, H.J., Yu, M.L., Ma, R.J., Wang, L.R., et al. (2013) Peach genetic resources: diversity, population structure and linkage disequilibrium. *BMC Genet*. 14: 84.

Liu, X., Han, S., Wang, Z., Gelernter, J. and Yang, B.-Z. (2013) Variant callers for next-generation sequencing data: a comparison study. *PLoS One* 8: e75619.

Micheletti, D., Dettori, M.T., Micali, S., Aramini, V., Pacheco, I., Linge, C.D.S., et al. (2015) Whole-genome analysis of diversity and SNP-major gene association in peach germplasm. *PLoS One* 10: 0136803.

Monet, R. and Bassi, D. (2008) Classical genetics and breeding. *In* The Peach: Botany, Production and Uses. Edited byLayne, D. and Bassi, D. pp. 61–84. CAB International, Wallingford, UK.

Mochida, K. and Shinozaki, K. (2010) Genomics and bioinformatics resources for crop improvement. *Plant Cell Physiol*. 51: 497–523.

Pascal, T., Kervella, J., Pfeiffer, F.G., Sauge, M.H. and Esmenjaud, D. (1998) Evaluation of the interspecific progeny *Prunus persica* cv Summergrand × *Prunus davidiana* for disease resistance and some agronomic features. *Acta Hort*. 465: 185–192

Quinlan, A.R. and Hall, I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26: 841–842.

Skinner, E.I., Uzilov, A.V., Stein, L.D., Mungall, C.J. and Holmes, I.H. (2009) JBrowse: a next-generation genome browser. *Genome Res*. 19: 1630–1638.

Van der Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., et al. (2013) From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinformatics* 11: 11.10.1–11.10.33.

Velasco, D., Hough, J., Aradhya, M. and Ross-Ibara, J. (2016) Evolutionary genomics of peach and almond domestication. *G3 (Baltimore)* 6: 3985–3993.

Verde, I., Abbott, A.G., Scalabrin, S., Jung, S., Shu, S., Marroni, F., et al. (2013) The high-quality draft genome of peach (*Prunus persica*) identifies unique patterns of genetic diversity, domestication and genome evolution. *Nat. Genet*. 45: 487–494.

Verde, I., Jenkins, J., Dondini, L., Micali, S., Pagliarani, G., Vendramin, E., et al. (2017) The Peach v2.0 release: high-resolution linkage mapping and deep resequencing improve chromosome-scale assembly and contiguity. *BMC Genomics* 18: 225.