Plant Biotechnol Rep (2014) 8:49–64
DOI 10.1007/s11816-013-0282-3

REVIEW ARTICLE

Genomics and bioinformatics resources for translational science in Rosaceae

Sook Jung · Dorrie Main

Received: 11 February 2013 / Accepted: 22 April 2013 / Published online: 21 May 2013

Abstract Recent technological advances in biology promise unprecedented opportunities for rapid and sustainable advancement of crop quality. Following this trend, the Rosaceae research community continues to generate large amounts of genomic, genetic and breeding data. These include annotated whole genome sequences, transcriptome and expression data, proteomic and metabolomic data, genotypic and phenotypic data, and genetic and physical maps. Analysis, storage, integration and dissemination of these data using bioinformatics tools and databases are essential to provide utility of the data for basic, translational and applied research. This review discusses the currently available genomics and bioinformatics resources for the Rosaceae family.

Keywords Rosaceae · Bioinformatics · Database · Genomics · Genetics · Breeding

Abbreviations

BAC	Bacterial artificial chromosome
BIND	Biomolecular interaction network database
CoGe	Place to compare genomes
EST	Expressed sequence tag
GDR	Genome database for Rosaceae
GEO	Gene expression omnibus
GO	Gene ontology
IGA	Istituto di Genomica Applicata
IPGI	International Peach Genome Initiative
JGI	Joint Genome Institute
KEGG	Kyoto Encyclopedia of Genes and Genomes
KOG	Eukaryotic Orthologous Groups
MASCP	Multinational *Arabidopsis* Steering Committee Proteomics Subcommittee
MTL	Mendelian trait loci
NCBI	National Center for Biotechnology
NMR	Nuclear magnetic resonance
ORF	Open reading frame
P3DB	Plant Protein Phosphorylation DataBase
PANTHER	Protein Analysis through Evolutionary Relationships
PGDD	Plant Genome Duplication Database
PPV	Plum pox virus
QTL	Quantitative trait loci
RCSB	Resource for Studying Biological Macromolecules
RosCOS	Rosaceae Conserved Orthologous Set
SCOP	Structural Classification of the Protein database
SGN	Solanaceae Genomics Network
SNP	Single nucleotide polymorphism
SRA	Short read archives
SSR	Simple sequence repeat
SUBA	*Arabidopsis* Subcellular Database
TAIR	*Arabidopsis* information resource
USRosEXEC	US Rosaceae Genomics, Genetics and Breeding Executive Committee

S. Jung (✉) · D. Main
Department of Horticulture, Washington State University, Pullman, WA 99164, USA
e-mail: sook_jung@wsu.edu

D. Main
e-mail: dorrie@wsu.edu

WTSS	Whole Transcriptome Shotgun Sequencing
WwPDB	Worldwide Protein Data Bank

Introduction

Rosaceae, comprised of over 100 genera and 3,000 species, contains a variety of crop species that are both biologically and economically important. Fruit-producing crops include apple (*Malus*), pear (*Pyrus*), raspberries/blackberries (*Rubus*), strawberries (*Fragaria*), and stone fruits (*Prunus*), such as peach/nectarine, apricot, plum, cherry and almond. Rosaceae also contains a wide variety of ornamental plants including roses, flowering cherry, crabapple, quince and pear. Crop improvement has traditionally been performed by incorporating desired traits from wild relatives through conventional breeding. Recent advances in high-throughput technology have revolutionized biology and provide unprecedented opportunities for rapid advancement of crop improvement through marker or genomics-assisted breeding.

The first spike in data generation occurred in the early 1990s when large-scale sequencing became available through the use of Sanger technology for EST and BAC sequencing (Adams et al. 1991; Shizuya et al. 1992). In the last few years, the advent of next generation technologies, such as 454 and Illumina, have significantly enhanced the ability to generate large-scale transcriptome and genome sequence at a fraction of the cost of Sanger sequencing. Similar advances in DNA array, nuclear magnetic resonance (NMR), Fourier transform infrared spectroscopy, Fourier transform ion cyclotron resonance mass spectroscopy, high performance liquid chromatography and mass spectrometry have generated large-scale gene expression, proteome and metabolome data for many species. Other data types include molecular marker data along with genetic mapping data and/or large-scale genotyping data of various varieties. More recently, high-throughput phenotypic and genotypic data are also being generated to study the interaction between genotype, phenotype and environment as well as for breeding purposes. All of these large-scale data require proper analysis, storage, integration and dissemination to enhance our understanding of biology and to be utilized in further research. Bioinformatics tools and methodologies, therefore, have become an essential and integral part of the new era of "information-driven" biological research.

Large-scale genome and transcriptome data were initially accumulated for model species, but are now available for a wide range of species. Crop plants with sequenced genomes include rice (*Oryza sativa*) (International Rice Genome Sequencing Project 2005), grapevine (*Vitis vinifera*) (Jaillon et al. 2007), sorghum (*Sorghum bicolor*) (Paterson et al. 2009), cucumber (*Cucumis sativus*) (Huang et al. 2009), maize (*Zea mays*) (Schnable et al. 2009), soybean (*Glycine max*) (Schmutz et al. 2010), cotton (Wang et al. 2012), sweet orange (Xu et al. 2012) and five species in Rosaceae: peach (*Prunus persica*), apple (*Malus domestica*) (Velasco et al. 2010), strawberry (*Fragaria vesca*) (Shulaev et al. 2011), pear (*Pyrus bretschneideri*) (Wu et al. 2013) and *Prunus mume* (Zhang et al. 2012), with red raspberry, black raspberry, apricot and plum, among others, currently being sequenced. In addition to the annotated whole genome sequences, a wealth of other genomic and genetic data is available for Rosaceae (Shulaev et al. 2008). These include BAC libraries, peach and apple physical maps, ESTs, numerous genetic maps in various species of Rosaceae and molecular markers that have been used for mapping and genotyping. Large-scale genotypic and phenotypic data are also being generated from various projects including HIDRAS (Gianfranceschi and Soglio 2004), ISAFRUIT (Audergon et al. 2009), GENBERRY (Diamanti et al. 2012) RosBREED (Iezzoni et al. 2010) and FruitBreedomics (http://www.fruitbreedomics.com). Currently, there are only limited proteomic and metabolomic data available for Rosaceae.

In this review, we discuss genomics and bioinformatics resources for the Rosaceae family with the corresponding database resources that can be utilized for Rosaceae researchers across disciplines (Fig. 1; Table 1). Resources available in other model species are also discussed, since these present a valuable tool in conducting and future planning of research in Rosaceae.

Sequence resources

The availability of extensive sequence data forms an essential genomics resource in designing various research platforms to understand the biology of crops and to apply the knowledge in their improvements. Three primary sequence databases, to one of which researchers seeking peer-reviewed publication of new genomic data are required to submit their genome or gene sequences, are GenBank (Benson et al. 2013), European Nucleotide Archive (Cochrane et al. 2013) and the DNA Data Bank of Japan (Kaminuma et al. 2011). These databases synchronize daily, and, consequently, are generally the most up-to-date source of genomic data for any species. A caveat to this rule is where genome sequences are released ahead of publication through other portals, as in the case of the peach, sweet orange, mandarin and cotton genomes, for

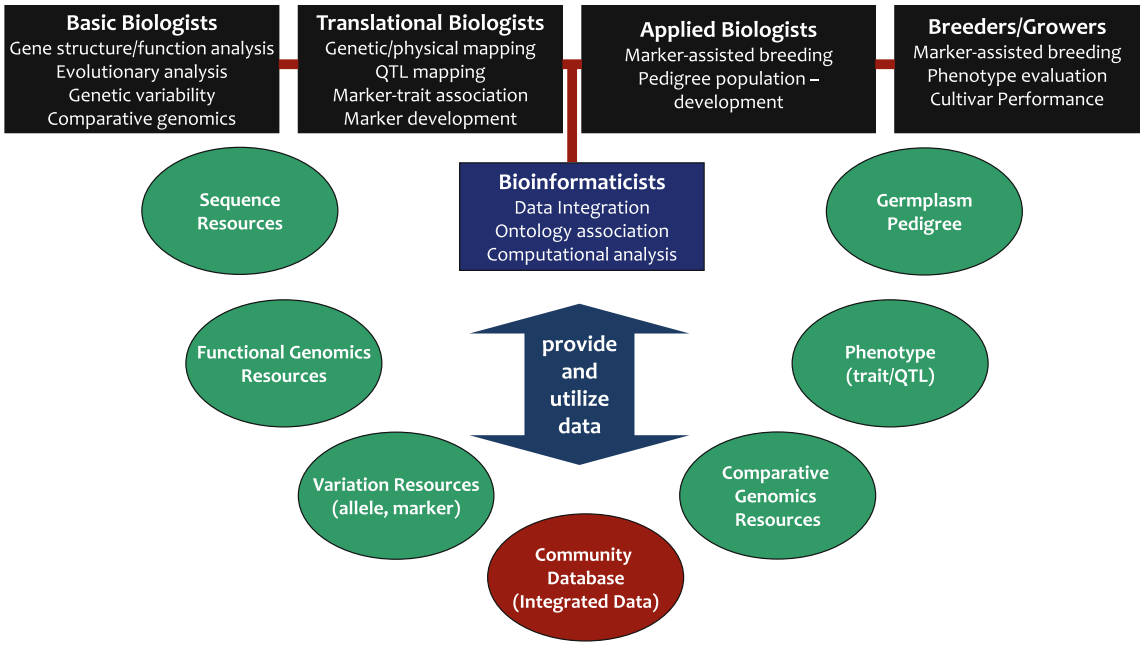

Fig. 1 Genomics and bioinformatics resources that can be utilized by Rosaceae researchers across disciplines

Table 1 Database resources for the Rosaceae family

Name	URL	Reference
GenBank	http://www.ncbi.nlm.nih.gov/genbank/	Benson et al. (2013)
European Nucleotide Archive (ENA)	http://www.ebi.ac.uk/ena/	Cochrane et al. (2013)
DNA Data Bank of Japan (DDBJ)	http://www.ddbj.nig.ac.jp/	Kaminuma et al. (2011)
Phytozome	http://www.phytozome.org/	
CoGe (The Place to Compare Genomes)	http://synteny.cnr.berkeley.edu/CoGe/	Lyons and Freeling (2008)
The plant genome duplication database (PGDD)	http://chibba.agtec.uga.edu/duplication/	Lee et al. (2013)
Plaza	http://bioinformatics.psb.ugent.be/plaza/	Van Bel et al. (2012)
GreenPhylD	http://greenphyl.cirad.fr/v2/cgi-bin/index.cgi	Rouard et al. (2011)
SALAD database	http://salad.dna.affrc.go.jp/salad/en/	Mihara et al. (2010)
NCBI's Gene Expression Omnibus (GEO)	http://www.ncbi.nlm.nih.gov/geo/	Wheeler et al. (2007)
ESTree DB	http://www.itb.cnr.it/estree/	Lazzari et al. (2008)
UniProt	http://www.uniprot.org/	The UniProt Consortium (2012)
The Worldwide Protein Data Bank (wwPDB)	http://www.wwpdb.org/	Berman et al. (2007)
CATH database	http://www.cathdb.info/	Cuff et al. (2011)
SUPERFAMILY	http://supfam.cs.bris.ac.uk/SUPERFAMILY/	Wilson et al. (2009)
ARAMEMNON	http://aramemnon.uni-koeln.de/	Schwacke et al. (2003)
The Arabidopsis Interactions Viewer	http://bar.utoronto.ca/interactions/cgi-bin/arabidopsis_interactions_viewer.cgi	Geisler-Lee et al. (2007)
The Biomolecular Interaction Network Database (BIND)	http://download.baderlab.org/BINDTranslation/	Isserlin et al. (2011)
The Arabidopsis Subcellular Database (SUBA)	http://suba.plantenergy.uwa.edu.au/	Heazlewood et al. (2007)
The Plant Protein Phosphorylation DataBase (P3DB)	http://p3db.org/	Yao et al. (2012)
AtMetExpress	http://prime.psc.riken.jp/lcms/AtMetExpress/	Matsuda et al. (2010)
Golm Metabolome Database (GMD)	http://gmd.mpimp-golm.mpg.de/	Hummel et al. (2010)
Genome Database for Rosaceae (GDR)	http://www.rosaceae.org/	Jung et al. (2008)

example, which were released through Phytozome (http://www.phytozome) ahead of publication. GenBank is built and distributed by the National Center for Biotechnology (NCBI). They collect sequence data by submission from authors and bulk submission from high-throughput sequencing centers. The Entrez search site on the homepage of NCBI (http://www.ncbi.nih.gov) allows researchers to search all the different data that it houses. Searching for information on *Prunus* species, by typing Prunus[ORGN] in the search box, returns a results page categorized by the different databases it houses. As of February 1, 2013, this search returned 131 genes, 110,815 ESTs, 7,481 unigenes, 4,931 proteins, 48,261 genome survey sequences, 46,915 nucleotide sequences, 14,705 SNPs and 89 large-scale datasets in the short read archives (SRA). For *Malus*, 61 genes, 336,190 ESTs, 22,493 unigenes, 3,523 proteins, 76 genome survey sequences, 6,410 nucleotide sequences, 7,907 SNPs and 11 large-scale datasets in the SRA. For *Fragaria*, 609 genes, 58,795 ESTs, 3,055 proteins, 17 genome survey sequences, 3,212 nucleotide sequences and 18 large-scale datasets in the SRA. Users can also download the whole genome data from their FTP site and BLAST against the whole genome. NCBI, is therefore, a good source of sequence data for Rosaceae, although access to the whole genome data is somewhat limited, and some of the larger datasets such as the SRA are not generally very useful for biologists without access to bioinformatics expertise to help with analysis of data. Researchers can keep up to date on new data submitted to NCBI for their species of interest by creating a My NCBI account which saves search history and filters data, among other features. For more information on the data and tools available at NCBI (GenBank), DDJB or EMBL (ENA), readers should refer to the excellent online tutorials at these sites and the yearly update published in the Nucleic Acids Research Database Addition published every January.

Recently, several genomes of Rosaceae have been sequenced and some of these are available from the Genome Database for Rosaceae (GDR, http://www.rosaceae.org/peach/genome, Jung et al. 2008) and Phytozome (http://www.phytozome.org/peach.php). The peach genome v1.0 is a high quality plant genome. It currently consists of eight pseudomolecules (scaffolds) representing the eight chromosomes of peach, and are numbered according to their corresponding linkage groups. The genome sequencing of the double haploid cultivar 'Lovell' consisted of approximately 7.7-fold whole genome shotgun sequencing employing the highly accurate Sanger methodology, and was assembled using Arachne (Batzoglou et al. 2002). The assembled peach scaffolds cover nearly 99 % of the peach genome, with over 92 % having confirmed orientation. To further validate the quality of the assembly, 74,757 *Prunus* ESTs were queried against the genome at 90 % identity and 85 % coverage, and it was found that only ~2 % were missing. While gene prediction and annotation is an ongoing process, current estimates indicate that peach has 28,689 transcripts and 27,852 genes. The whole genome sequence of apple, *Malus* × *domestica*, was reported by sequencing and assembly of the 'Golden Delicious' apple genome followed the whole genome shotgun approach (Velasco et al. 2010). Of the 16.9-fold genome coverage, 26 % was provided by Sanger dye primer sequencing of paired reads, and the remaining 74 % was from 454 sequencing by synthesis of paired and unpaired reads. The sequence data showed that a relatively recent (>50 million years ago) genome-wide duplication event occurred in Pyreae lineage. The domesticated apple genotypes are all highly heterozygous and the assembly produced overlapping contigs. Of 122,146 contigs, 103,076 were assembled into 1,629 metacontigs. Anchoring of metacontigs (598.3 Mb, or 71.2 % of genome) was based on the high-quality genetic map with 1,643 markers. In total, 17 linkage groups, or chromosomes, were reconstructed. The total number of genes predicted for the apple genome is 57,386, including some genes that may be present only in one of the two chromosomes of a pair. Efforts are on-going in the apple community to improve the assembly and annotation of the apple genome. A diploid strawberry, *Fragaria vesca*, has also been sequenced as a reference genome for the genus that contains the cultivated tetraploid strawberry, *Fragaria* × *ananassa* (Shulaev et al. 2011). The genome was sequenced to 39× coverage using second-generation technology, assembled de novo and then anchored to the genetic linkage map into seven pseudochromosomes. A total of 34,809 genes were predicted, with most being supported by transcript evidence. Newer version v1.1, which contains an updated pseudomolecule assembly of the original v1.0 is also available. Recently, whole genome sequencing of pear (*Pyrus bretschneideri*) and *Prunus mume* have also been reported. The pear genome was sequenced to 194× coverage using a combination of BAC-by-BAC and next generation sequencing (Wu et al. 2013). A 512.0 Mb sequence covered 97.1 % of the estimated genome size of this highly heterozygous species and 75.5 % of the sequence has been anchored to all 17 chromosomes by high density genetic maps comprising of 2,005 SNP markers. A total of 42,812 genes have been predicted with about 28.5 % encoding multiple isoforms. For *Prunus mume*, approximately 84.6 % (237 Mb) of its genome was assembled by combining 101× next-generation sequencing and optical mapping data (Zhang et al. 2012), while 83.9 % of scaffolds have been anchored to the eight chromosomes with genetic map constructed by restriction-site-associated DNA sequencing.

The whole genome sequence and annotation data of three Rosaceae species, peach, strawberry and apple, are

available from GDR and Phytozome. Details on the resources and tools for the genomes in GDR are provided in the "Community database resources" section of this review. Phytozome is a joint project of the Department of Energy's Joint Genome Institute (JGI) and the Center for Integrative Genomics to facilitate comparative genomic studies amongst green plants. Families of orthologous and paralogous genes that represent the modern descendents of ancestral gene sets are constructed at key phylogenetic nodes. These families allow easy access to clade specific orthology/paralogy relationships as well as clade specific genes and gene expansions. As of release v9.0, Phytozome provides access to 31 sequenced and annotated green plant genomes which have been clustered into gene families at ten evolutionarily significant nodes. Where possible, each gene has been annotated with PFAM (Finn et al. 2010), KOG (Tatusov et al. 2003), KEGG (Aoki and Kanehisa 2005), and PANTHER (Mi et al. 2010) assignments, and publicly available annotations from RefSeq (Pruitt et al. 2007), UniProt (The UniProt Consortium 2012) and TAIR (Lamesch et al. 2012). Similar to the GDR genome sites, users can view the genome though the genome browser GBrowse, which provides access to transcripts, alternative transcripts, mapped ESTs, aligned plant peptides and genetic markers. Users can also search their sequences against the peach genome sequence, transcripts and proteome though BLAST and BLAT servers, and download all the genome annotation through a bulk data site. The Istituto di Genomica ApplicataIGA peach site (IGA, http://services.appliedgenomics.org/projects/drupomics/intro/) also contains a peach GBrowse and BLAST tools. The IGA peach GBrowse provides access to the same tracks as phytozome but in addition they have RNA-seq Illumina profiles from tissues sequenced from cotyledon and embryo, fruit, leaf and root; repetitive sequences and microsatellites; and a profile of exact genome 20-mers.

Resources for variation analysis

High-density genome scanning is now possible through use array platforms developed using comprising high-throughput and low-cost next generation sequencing (NGS) to identify single nucleotide polymorphisms (SNPs). The International RosBREED SNP Consortium (IRSC) *Malus* research community developed an Infinium® II WGG genotyping array (IRSC array, Chagne et al. 2012) for *Malus* and *Pyrus* using data from the re-sequencing of 27 *Malus* genotypes along with data from the 'Golden Delicious' genome sequence. The IRSC array contains a total of 7,867 *Malus* SNPs in addition to 921 *Pyrus* SNPs. A GoldenGate-based assay platform (Khan et al. 2012) is also available for *Malus* containing 12,299 SNPs identified from EST data from 14 genotypes. Of these, 1,411 SNPs were validated using four apple genotypes. The IRSC also developed Infinium® II WGG genotyping arrays for peach (Verde et al. 2012) and cherry (Peace et al. 2012). The cherry array used data from accessions of sweet cherry and of tart cherry, along with the peach genome sequence to develop a 6,000 SNP array, of which approximately one-third were informative for each crop (Peace et al. 2012).

The IRSC peach array comprised of 8,144 SNPs developed from re-sequencing of 53 peach genotypes along with data from the peach genome sequence (Verde et al. 2012).

Functional genomics resources

EST resources

ESTs are valuable resources for marker development, genome annotation, co-expression studies and comparative analyses as well as for gene discovery. Currently, GenBank contains 527,240 ESTs for Rosaceae: 336,190 from *Malus*, 110,815 from *Prunus*, 58,795 from *Fragaria* and the rest from other genus. *Malus* has the largest collection of ESTs and the majority have been produced to aid discovery of genes involved in important agricultural traits with National Science Foundation funding (award no. 0321701; Schuyler Korban, principal investigator) and by HortResearch in New Zealand (Newcomb et al. 2006). Transcriptome analyses in *Prunus* focused on generating ESTs to identify candidate genes involved in different tissues, different stages of plant development and responding to abiotic and biotic stresses (Georgi et al. 2002; Horn et al. 2005; Vecchietti et al. 2009). Transcriptomics analyses have also been used to investigate fruit ripening and post-harvest physiology in *Prunus* (Trainotti et al. 2003; Vizoso et al. 2009). In *Fragaria*, 50,627 transcripts are from the diploid *Fragaria vesca* and 10,855 are from the cultivated tetraploid *Fragaria* × *ananassa*. Out of 143 SRA datasets for Rosaceae, 79 are from RNA sequences. The majority, 42 datasets, are from *Prunus* with other datasets from other genera: 12 from *Fragaria*, 10 from *Pyrus*, 9 from *Malus* and 6 from *Rosa*.

GDR contains all the publicly available Rosaceae ESTs and the unigene sets for the family Rosaceae and each genus. With the availability of the whole genome sequences, the ESTs from *Malus*, *Prunus* and *Fragaria* are anchored to the predicted gene transcripts from whole genome sequencing. The best matches between the ESTs and the predicted genes are available in a spreadsheet with links to the predicted genes in the GDR GBrowse. PlantGDB (http://www.plantttgdb.org) also contains EST unigenes, assembled from GenBank ESTs, from various plant

species. Unigene data is available from *Prunus persica*, *Prunus armeniaca*, *Malus domestica*, *Fragaria vesca* and *Fragaria ananassa*. ESTree DB (Lazzari et al. 2008) has a collection of ESTs from *Prunus persica* and *Prunus amygdalus*. The ESTs are from 12 peach libraries and 3 almond libraries produced in nine different laboratories. The peach unigene in the sixth release of ESTree DB contains 28,391 unigenes with 7,709 contigs and 20,682 singlets, assembled from 75,404 ESTs. The ESTree mapping project is underway to map about 200 additional ESTs on the peach transcriptome.

Microarrays

Malus has the largest collection of microarrays as well as ESTs which have been used to study environmental effects on tree-to-tree variability in the orchard, fruit aroma production (Schaffer et al. 2007), early development of apple fruit (Lee et al. 2007), apple fruit development from the floral bud to ripe fruit (Janssen et al. 2008), fruit development and ripening (Costa et al. 2010), fruitlet abscission (Botton et al. 2011), role of Ring finger gene family in fruit development (Li et al. 2011), fire blight susceptibility (Jensen et al. 2012) and apple fruit maturation and texture attributes (Zhu et al. 2012).

High-throughput transcriptome analysis using microarrays have been recently incorporated in various *Prunus* research projects. One of the platforms, microPEACH1.0, consists of 4,806 70-mer oligonucleotides designed from *Prunus persica* (peach) EST unigene clusters, mainly in the 3′ end terminal region. It was used to investigate transcriptome profiling of ripening nectarine (Ziliotto et al. 2008), stone formation in peach fruit (Dardick et al. 2010) and the role of Jasmonate in fruit ripening (Ziosi et al. 2008). GFNChile_Peach_0.9K_v1.0 contains 847 cDNAs from a cv. Loring ripe peach fruit (*Prunus persica*) cDNA library and used to study cold-responsive genes in peach fruit (Ogundiwin et al. 2008). Website (http://bioinfo.ibmcp.upv.es/genomics/ChillPeachDB) holding detailed information on the ChillPeach database was also created. Other investigations that used microarrays include the identification of key genes in almond adventitious shoot regeneration (Santos et al. 2009) and evaluation of expression of genes mostly engaged in fruit development between *Prunus mume* and *Prunus armeniaca* (Li et al. 2012). Recently, a 60-mer oligo-DNA microarray, constructed using *Prunus* Unigene V4 from GDR, was used to study fruit softening in peach (Tatsuki et al. 2013). Some of the results are available from NCBI's Gene Expression Omnibus (GEO) (Wheeler et al. 2007).

In *Fragaria*, cDNA microarray have been used to study maturation and non-climacteric ripening (Aharoni et al. 2002) and the evolution of fruit flavor compounds (Aharoni et al. 2004) in the cultivated octoploid *Fragaria* × *ananassa*. An oligo-based microarray with sequences from both diploid and octoploid has also been used to compare the transcriptome of the ripe receptacle in these species (Bombarely et al. 2010).

RNA-seq technology

The latest and most efficient tool for transcriptome analysis involves the use of high-throughput sequencing technologies to sequence cDNA. This technology is called Whole Transcriptome Shotgun Sequencing (WTSS) or RNA-seq. RNA-seq allows not only the assessment of the expression level of specific genes but also the detection of less-represented transcripts, allelic-specific expression of transcripts, post-transcriptional mutations and the expression of splice-variants. Even though it has been available for only a couple of years, this technology has been used extensively in humans, mammals and yeast and it is beginning to be used in plant species. *Arabidopsis* was the first plant species to be studied using this technology (Weber et al. 2007). Comparison of deep transcriptome sequencing with the EST database confirmed most of the annotated introns and identified thousands of novel alternatively spliced mRNA isoforms, suggesting at least 42 % of intron-containing genes of *Arabidopsis* are alternatively spliced (Filichkin et al. 2010). These RNA-seq data (Lister et al. 2008; Filichkin et al. 2010), along with proteomics data (Baerenfaller et al. 2008; Castellana et al. 2008), have been used to revise the *Arabidopsis* gene models and are incorporated in the TAIR10 release (Lamesch et al. 2012).

In *Malus*, RNA-seq experiments were performed to study genes associated with columnar phenotype, characterized by a compact growth habit, of apple trees (Krost et al. 2013). In *Prunus*, RNA-seq technology is currently being utilized in projects to study PPV (Plum Pox Virus) resistance, graft incompatibility fruit quality and flowering time in peach, apricot and plum, as summarized in Martínez-Gómez et al. (2011). The results are expected to be incorporated in GDR when the results are publicly available.

Resources for proteomics and metabolomics

Proteomics and metabolomics refer to the large-scale study of proteins and metabolites. When transcriptomic, proteomic and metabolomic data are integrated, it can help to give us a more complete picture of a living organism in a specific condition. Due to the enormous complexity of the proteome, many different approaches are being taken to generate and catalogue proteomics data, such as structural and functional annotation of proteins, protein expression and dynamics, stress and developmental responses, post-translational

protein modifications and protein interactions. UniProt (The UniProt Consortium 2012) provides a comprehensive, high-quality and freely accessible resource of protein sequence and functional information. Currently (searched in January 2013), 3,695, 2,495 and 1,982 protein entries are available for *Prunus, Malus* and *Fragaria*, respectively. Only 70, 45 and 34 entries of them are in Swiss-Prot and the rest (3,625, 2,450 and 1,948) are in TrEMBL.

The Worldwide Protein Data Bank (WwPDB) (Berman et al. 2007) is the main structural protein database. It includes RCSB (A Resource for Studying Biological Macromolecules) PDB in USA, PDBe in Europe (Velankar et al. 2011) and PDBj in Japan. Currently, PDB contains the coordinates and related information of more than 76,000 structures of proteins, nucleic acids, protein/nucleic acid complexes and other macromolecules that have been determined using X-ray crystallography, NMR and electron microscopy techniques. Only limited numbers of structures have been deposited in wwPDB. For *Prunus*, 11 structures are currently available from PDB including proteins from *Prunus persica*, *Prunus avium*, *Prunus dulcis* and *Prunus mume*. Other entries for the Rosaceae family include six from *Malus domestica* and one from *Fragaria* × *ananassa*. Other protein structure databases include CATH (Cuff et al. 2011) and SUPERFAMILIY (Wilson et al. 2009). The CATH (class, architecture, topology and homology) database provides a hierarchical classification of protein domain structures obtained from PDB. The classification class reflects the amino acid composition, architecture of the general shape of the protein domain and topology of the way in which the protein folds into this architecture. SUPERFAMILY provides the prediction of protein domains of known structure in amino acid sequences. The classification of domains is hierarchical, based on nature of the similarity (sequence, evolutionary and structural), class, fold, superfamily and family, following the structural classification of the protein (SCOP) database (Andreeva et al. 2008). SUPERFAMILY currently includes data for 2,476 distinct organisms, including 373 Eucaryotes. Data from *Prunus persica*, *Malus domestica* and *Fragaria vesca* are available from SUPERFAMILY.

Various web-based plant proteome-related databases are summarized in the database section of the proteomics subcommittee of the Multinational *Arabidopsis* Steering Committee Proteomics Subcommittee (MASCP) Web site (http://www.masc-proteomics.org/). Some of the databases are illustrated below. ARAMEMNON (Schwacke et al. 2003) is a database of plant membrane proteins with *Arabidopsis thaliana* as the reference model plant. Currently, the database holds all putative membrane proteins of five other plant species: grape (*Vitis vinifera*), poplar (*Populus trichocarpa*), rice (*Oryza sativa*), maize (*Zea mays*) and brachypodium (*Brachypodium distachyon*). The *Arabidopsis* Interactions Viewer (Geisler-Lee et al. 2007) is an interaction database for *Arabidopsis thaliana* predicted from interacting orthologs in yeast (*Saccharomyces cerevisiae*), nematode worm (*Caenorhabditis elegans*), fruit fly (*Drosophila melanogaster*) and human (*Homo sapiens*). The database includes 70,944 predicted and 22,156 confirmed *Arabidopsis* interacting proteins. The confirmed *Arabidopsis* interacting proteins come from BIND, the Biomolecular Interaction Network Database (Isserlin et al. 2011), high-density *Arabidopsis* protein microarrays (Popescu et al. 2007, 2009) and other literature sources. The interactions in BIND were identified using several different methods, such as yeast two-hybrid screens, but also via traditional biochemical methods. All subcellular localization data in the *Arabidopsis* Interactions Viewer are from SUBA, the *Arabidopsis* Subcellular Database (Heazlewood et al. 2007). Subcellular localization data in SUBA are brought together from various sources such as studies using chimeric fluorescent fusion proteins, proteomic surveys using mass spectrometry and literature. It also contains precompiled bioinformatic predictions for protein subcellular localizations from a set of ten different prediction tools. Complex relational queries can be performed between these experimental and predicted datasets to find and collate evidence for the subcellular location of *Arabidopsis* proteins. The Plant Protein Phosphorylation DataBase (P3DB) (Yao et al. 2012) hosts protein phosphorylation data for eight species: *Arabidopsis thaliana*, *Brassica napus*, *Glycine max*, *Medicago truncatula*, *Oryza sativa*, *Solanum tuberosum*, *Zea mays* and *Nicotiana tabacum*. AtMetExpress (Matsuda et al. 2010) and the Golm Metabolome Database (Hummel et al. 2007, 2010) are metabolome databases. AtMetExpress contains data from a study of phytochemical accumulation during development of the model plant *Arabidopsis thaliana* using liquid chromatography–mass spectrometry in samples covering many growth stages and organs. The Golm Database contains data of mass spectra from biologically active metabolites quantified using gas chromatography coupled to mass spectrometry. It covers data from mammals, yeast, corynebacterium, model plants, such as crop plants and related wild species, as well as required non-sample controls.

Comparative genomics resources

Comparative genomics is an area of study where the structure and function of genomes from different species or varieties are compared. An important goal of comparative genomics is to obtain an insight into how genomes have evolved. Detection of duplicated regions within species and syntenic regions among species plays an important role in studies of genome evolution. The comparison of genomes and their contents among species allows us to identify

genes and other sequence features that are conserved among species and/or genes that are specific to certain clades. These results help us to infer what genes and other sequence features are responsible for the similarities and differences among species. Comparative genomics also allows us to transfer knowledge from well-studied species to less-studied species.

With the increasing number of species with whole genome sequences, several web-based databases are available to compare whole genome data. A few contain the whole genome sequences of one or more species of the Rosaceae family: *Prunus persica*, *Malus domestica* and *Fragaria vesca*.

CoGe (The Place to Compare Genomes) (Lyons and Freeling 2008), contains whole genome data of 15,037 species across all domains of life, including *Prunus persica*, *Fragaria vesca* and *Malus domestica*. CoGe contains a series of web tools where users can select species of interest to view the genome data and/or perform some comparative analysis. The tools include OrganismView for searching organisms, CoGeBlast for blasting sequences against the genomes of interest, FeatView for searching genomic features by name, SynMap for whole genome syntenic dotplot analyses, SynFind for identifying syntenic regions across multiple genomes and GEvo for high-resolution sequence analysis of genomic regions. SynMap is the tool for comparative genomics which uses DAGChainer (Haas et al. 2004) as the underlying software for synteny detection.

The plant genome duplication database (PGDD) (Lee et al. 2013) contains 26 plant genomes including *Prunus persica*, *Malus domestica* and *Fragraria vesca*. PGDD contains similar tools to those in CoGE, where users can explore plant genes in terms of intra-genome or cross-genome syntenic relationships. Users can choose two species to view syntenic blocks in Dot Plot, can select a locus to view the regions that contain the locus along with the syntenic regions in multiple species, and can use BLAST to map the sequence of interest to selected genomes. PGDD uses the MCscan package (Tang et al. 2008), which uses DAGchainer to predict pairwise segments, as the underlying software for synteny detection.

Plaza (Van Bel et al. 2012) is another platform where users can access plant genome data to perform comparative analyses. Plaza contains 25 species including *Malus domestica* and *Fragraria vesca*. Data annotation includes primary gene annotation, gene family, orthologous genes, and functional annotation such as GO terms, InterPro Domains, and Reactome (pathway) data. In the Plaza Synteny Plot tool, users can start from a gene or a gene family to view the gene organization of all homologs of a gene family in selected species. In the Skyline plot, users can enter a gene locus to view the collinear regions that exist within a set of selected species. In WGDotplot, users can select two species to view syntenic blocks or select the same species twice to view all duplicated blocks within species. The collinear regions within species can also be viewed with Circle Plot. If available, the age of the collinear blocks, determined using Ks, is reported using a color code. From the Skyline plot and WGDotplot, users can access Multiplicon View to see the aligned gene strings of a set of homologous segments. The collinear regions are detected by i-ADHoRe (Simillion et al. 2008). In addition to those described above, Plaza provides other tools for genome evolution and collinearity research. The WGMapping tool allows users to choose all genes or selection of genes to display their location on the chromosomes along with the gene type. The functional clustering visualization tool provides an overview of the location and content of each functional cluster, detected using C-Hunter (Yi et al. 2007), on a chromosome-wide scale. Genomic sequences and genomic features can be viewed using Genome Browser. Tools such as Similarity heatmap, orthologous gene tool, Tree explorer and Gene family finder allows users to further explore gene family evolution.

The GreenPhylDB (Rouard et al. 2011) and SALAD databases (Mihara et al. 2010) are related databases. The GreenPhylDB contains data from *Malus domestica*, but the SALAD database does not yet contain any Rosaceae genome data. The difference between these and the comparative genome databases mentioned above is that GreenPhylDB and SALAD do not contain data or tools for synteny analysis, but focus on gene family, phylogeny and ortholog/paralog analyses. GreenPhylDB is a web resource for plant comparative and functional genomics. GreenPhylDB v3.0 contains 22 full genomes from the major phylum of plant evolution. The data include various lists of gene families, such as plant-, phylum- and species-specific lists, and tools to facilitate the comparisons. The SALAD database is a web-based resource for genome-wide comparative analysis of annotated protein sequences in plants. In SALAD, users can search for genes and view phylogenetic trees constructed from sequence alignment for a selected single motif or multiple motifs. Another functionality called sequence logo provides users with graphical representation of the sequence conservation of amino acids made from alignment of the conserved motif in each node. Users can also compare the sequence logos among members of distinct nodes to evaluate conservation and diversity of amino acid at any sites in the conserved motif.

GDR (Jung et al. 2008) uses GBrowse_Syn (McKay et al. 2010), a synteny browser, to show the orthologous regions among the three sequenced genomes of Rosaceae (Jung et al. 2012), as detected by the Mercator program (Dewey 2007). GBrowse_Syn is hyperlinked to GBrowse so that users can access various genome annotation data including markers from the conserved syntenic regions

shown in GBrowse_Syn. Comparative genomics data made available in GDR thus allow users to explore other data, such as genomic features, anchored trait loci and genetic markers, in the orthologous regions.

Community database resources

Plant community databases provide access to all or most of the datasets for individual or closely related species. Such databases include The *Arabidopsis* Information Resource (TAIR, Lamesch et al. 2012), the Genome Database for Rosaceae (GDR, Jung et al. 2008), the Solanaceae Genomics Network (SGN, Bombarely et al. 2011), Gramene (Youens-Clark et al. 2011), TreeGenes (Wegrzyn et al. 2008) and MaizeGDB (Schaeffer et al. 2011). Community databases generally store comprehensively integrated data such as annotated sequences of genomes and transcriptomes, genetic data, and molecular and phenotypic diversity data. As a result of data integration, the value of individual types of data increases exponentially, providing essential resources to accelerate the molecular understanding of phenotypic traits and the use of DNA information in crop improvement.

Initiated in 2002, GDR is the sole community database for the Rosaceae family, which includes economically important genera such as *Prunus*, *Fragaria*, *Malus*, *Pyrus*, *Rosa* and *Rubus*. The Rosaceae research community has generated data for the annotated genome sequences, physical maps, a large collection of ESTs, transcriptome map, numerous genetic maps, genetically mapped traits (MTL and QTL), genotypic diversity data, publicly available breeding data with both phenotypic and genotypic data, and cultivar evaluation data for growers. The integration and standardization of the data is crucial for the data to be utilized by different types of users including genomicists, geneticists, molecular biologists, evolutionary biologists, bioinformaticists, breeders and growers. The purpose of GDR is to collect, curate, analyze and integrate the data and provide efficient interfaces for user access to allow these numerous and complex data to be efficiently utilized. In the sections below, we describe the data and web interface, analysis tools and community tools available from GDR, together with work in progress and future plans. GDR tutorials are available at http://www.rosaceae.org/tutorials.

Data and web interface

Annotated whole genome sequence

The whole genome sequences of various versions of the three species, *Prunus persica* genome v1.0, *Malus domestica* genome v1.0 and v1.0p and *Fragaria vesca* genome v1.0 and v1.1, are available in GDR. *Malus domestica* genome v1.0 is represented as metacontigs, composed of assembled overlapping contigs that have been produced by the assembly of heterozygous genotypes, which have been anchored to the chromosomes. The v1.0p is the primary pseudo-haplogype assembly, which is composed of chromosome-anchored contigs that are non-overlapping. Users can access all the annotated genome data from each genome page. Through the graphic interface GBrowse (Donlin 2007; Stein et al. 2002), users can view various genomic features aligned to the genome, such as gene models, repeats, and SNPs, as well as alignments of ESTs, repeats, genetic markers and genes from other plant model species. Each feature has hyperlinks that lead to a page with sequences and other information, with further hyperlinks to external databases where applicable. The genetic marker feature in GBrowse is linked to the marker page in GDR where all the detailed marker information is available, such as primers, mapped positions, references, and link to CMap, the comparative map viewer in GDR. The mapped ESTs are also linked to the EST page in GDR, where all the detailed EST information is available. The genome pages in GDR contains various downloadable files, including the fasta files of predicted peach gene transcripts, CDS (coding sequences), and predicted gene peptides. Excel files of gene transcripts with homologs to *Arabidopsis* genes, Swiss-Prot entries and TrEMBL entries are also available with hyperlinks to external databases. Other Excel downladable files include various Rosaceae ESTs and genetic markers that map to the whole genome sequences and SNPs with hyperlinks to GBrowse in GDR. The orthologous regions among three species, detected by the Mercator program (Dewey 2007), are displayed using GBrowse_Syn (McKay et al. 2010). The whole genome data are also available to search by various categories, such as by name, interpro protein domain name or KEGG pathway terms, so that users can directly access the genes by querying. The predicted genes from the whole genome sequences have also been utilized in the construction of PlantCyc databases. Currently, three PlantCyc databases, peachCyc, appleCyc and FragariaCyc, are available in GDR for users to explore the pathway data.

Annotated EST unigene data

GDR contains all the publicly available Rosaceae ESTs, downloaded from the dbEST at NCBI (Gibney and Baxevanis 2011). Routine processing in GDR occurs in three stages: sequence filtering and trimming to obtain high-quality sequences, assembly into contigs to reduce the inherent redundancy and building unigene sets from the combined contigs and singlets, and sequence annotation. A unigene is available for *Prunus*, as well as *Malus*, *Fragaria*, *Rosa* and *Pyrus*. The assembled contigs and singlets for the four genera were assembled together to generate a

putative unigene set for the entire Rosaceae ESTs. Other annotation includes putative function and Gene Ontology (The Gene Ontology Consortium 2013) association to contigs and ESTs by homology with SwissProt, TREMBL and InterPRO proteins (Hunter et al. 2012). Plant Structure Ontology (Ilic et al. 2007) is also utilized to annotate the ESTs with the tissue from which the ESTs are generated.

The unigene page is a good starting point for an overview of the various annotated data for the unigenes. It displays the overall results of the project with a sidebar containing links to the library information, putative homology, KEGG analysis, microsatellite analysis and downloads. A link to the gene ontology (GO) classification is also available for the genera assemblies as well as the Rosaceae family assembly. Downloadable data includes batch sequence in fasta format, homology results file in Excel format with links into GDR and external databases and SSR/ORF/primers results in Excel format.

The EST search site is for those users who are interested in a subset of ESTs. They can choose to search ESTs of the entire Rosaceae or the genus of interest by selecting the appropriate tabs. Users can also search either ESTs or contigs. In each search page, ESTs or contigs can be searched by their name(s), assembly results, sequence features such as SSR or SNP, taxonomy, tissue type and putative function including match description, match organism and GO term. Users can also perform a batch search by uploading a file with EST names. Previous unigene versions are also available for search to help those who have been using an older version in their research. The results can be downloaded in fasta format or as a tab-delimited file with SWISS-PROT homology results containing hyperlinks back into the data for each sequence retrieved. Instead of displaying all the details on one page, the EST details page initially displays the clone information and the sequence with a sidebar containing links to library details, unigene information, sequence homology, SSR/ORF information, map position and anchored BACs when applicable. A unigene information page provides the contig name and hyperlink for both the genus and family unigenes. The contig page gives similar annotation data for the contig with additional links to the SNP results and the comprising ESTs. For the ESTs anchored to peach BACs and/or to Rosaceae genetic maps, the EST detail page provides a link to view the ESTs' map positions using the GDR Map Viewer or CMap.

Genetic map

GDR currently contains data for 54 genetic maps for Rosaceae species. GDR uses CMap, the web-based comparative map tool, to allow users to compare maps from different cultivars and species. The comparative mapping facilitates the data transfer from well-studied species to less-studied ones. For example, the GDR map collection includes the TxE map (Dirlewanger et al. 2004; Howad et al. 2005), which is recognized as the reference map for *Prunus.* The TxE map, constructed from an almond × peach F2 population, contains 826 markers with a total distance of 524 cM. The TxE map contain many markers that are used in the construction of maps of other *Prunus* species such as peach, apricot, sour cherry, plum × almond–peach hybrid and almond × peach, but also other Rosaceae species such as apple and pear. The essential collinearity of the anchored markers in the *Prunus* maps and the presence of large collinear blocks among different genera in Rosaceae, such as *Prunus* and *Malus* (Dirlewanger et al. 2004), enable comparative mapping, an invaluable tool for cross-utilization of data in Rosaceae. In addition to the directly-mapped genetic markers, the TxE map in GDR-CMap displays peach transcriptome map data, major trait loci affecting agronomic characters found in various *Prunus* species (Dirlewanger et al. 2004), pathogen resistance loci (Lalli et al. 2005) and Rosaceae Conserved Orthologous Set (RosCOS) (Cabrera et al. 2009). CMap also contains the apple integrated map which was developed for anchoring metacontigs from whole genome sequencing. The integrated map were derived from six F1 populations totaling 720 individuals. The FV × FB diploid *Fragaria* reference map (Sargent et al. 2011), which played important role in the scaffold ordering of the *F. vesca* genome sequence and rose integrated consensus map, built based on the information of four diploid populations and more than 1,000 initial markers (Spiller et al. 2011) are other important resources for the community. GDR-CMap serves as an integrative tool in the utilization of the data anchored to the maps in Rosaceae. The anchored features, such as marker and ESTs, in the map are also linked to the corresponding GDR sites so that all the relevant information for the features can be viewed. Markers that are anchored to various genomes have hyperlinks to GBrowse. Another important resource in GDR is the peach physical map data. The peach physical map (Zhebentyayeva et al. 2008) is constructed from two peach BAC libraries (Georgi et al. 2002) and the physical length of the map is estimated to be 303 Mb, which is 104.5 % of the peach genome. GDR uses two tools available from the WebAGCoL Package (Pampanwar et al. 2005) to display the current peach physical map.

Genetic markers and traits

To provide more details of the genetic markers and traits that have been used in genetic map development or genetic diversity studies, GDR contains an extensively annotated molecular marker database. Currently, over 44,439 extensively annotated markers, including SNPs from IRSC peach array v1 (Verde et al. 2012) and candidate SNPs, are

available from GDR search engine. The marker annotation includes marker aliases, source cultivar, source description, primer sequences, PCR conditions, references, map position, associated ESTs and associated BACs. While annotation of trait data is at an initial stage, the traits are annotated in GDR with aliases, published symbol, curated trait category, taxon, trait description, screening method, map position and references.

The marker search site allows both a simple search by name and an advanced search with various search categories. The search category includes marker type, the species from which the marker is developed, the species to which the marker is mapped, map position, markers with associated BAC clones and markers with associated ESTs. Users can also upload a file of names to get the detailed data. In the trait search site, users can search trait by name, symbol, taxon or curated trait category. Other SNP markers included in arrays, such as IRSC apple 9K (Chagne et al. 2012), cherry 6K (Peace et al. 2012) and UC Davis peach 6K (Ahmad et al. 2011), as well as those in IRSC peach 9K, are available to download and view in GBrowse.

DNA polymorphism

GDR contains DNA polymorphism data from various projects on molecular diversity studies of Rosaceae species. Currently, data from nine different projects are available, three from *Prunus* diversity studies and the rest from *Malus* and *Pyrus* studies. All the current data are from projects with SSR markers. Users can query by the marker name or species to view the details of diversity projects and the genotype of the varieties used in the analyses. SNP markers from the RosBREED project (http://www.rosbreed.org) are being added.

Breeding data

One of the newest components of GDR is the search/browse site for breeding data. GDR contains password-protected private breeding data as well as publicly available breeding data. Private breeding data includes data from the Washington Apple Breeding Program and Pacific North West Sweet Cherry Breeding Program. Public data includes data from the federally funded RosBreed project, a program designed to establish a sustainable marker-assisted breeding infrastructure for US Rosaceae crops. The data includes the varieties and their pedigrees, phenotyping and genotyping data and experimental metadata. The current search interfaces allows users to search by datasets, variety name, trait values and pedigree. From the result page, users can view detailed results for a variety or download an Excel file with all the phenotyping/genotyping results that has been specified. Users can also select a pedigree by selecting a variety and number of ancestral and progeny generations to generate an input file for breeding software such as Pedimap and FlexQTL (Bink and van Eeuwijk 2009). A breeding decision tool called Cross Assist, which produces a list of parents to cross and the number of seedlings to screen to obtain certain number of seedlings above/within user-specified trait thresholds, is also available.

Analysis tools

GDR web-based tools include a BLAST server, FASTA server, CAP3 Assembly server and SSR server. The FASTA/BLAST servers allow users to conduct sequence homology analyses against various sequence databases including annotated sequences in GDR. The databases include whole genome nucleotide and protein sequences of peach, strawberry and apple; ESTs of the Rosaceae or each genus from NCBI, genera-specific unigene sets, as well as a family-wide unigene; Rosaceae genomic or protein sequences from NCBI, peach, apple and cherry SNP sequences; and *Arabidopsis* protein sequences from TAIR and the ESTs, unigene sets, SSR-containing ESTs from individual cDNA libraries of peach mesocarp, almond, octoploid strawberry and diploid strawberry. Peach mesocarp ESTs that are anchored to the peach BACs are also available for sequence analysis. Batch sequences can be uploaded for analysis and the results are returned as both raw aligned output and parsed out in Excel. The output in Excel has hyperlinks to the GDR and NCBI sites. FASTA formatted library files of both the sequences with or without matches are provided to allow the user to easily conduct further batch searches in GDR and other databases. An EST assembly server using the CAP3 program is also available so that users can assemble their own EST sets. The server returns the raw output, a summary report and a fasta file containing the combined contig sequences and singlet sequences, which are also available as individual files. The contig file lists the contig number and comprising clone names in the comment line for each assembled transcript. Also available is a SSR server that allows user-defined SSRs to be identified in uploaded sequences. Users can also choose to run Primer3 along with the SSR-detection program to generate primer sets for the SSRs. The results are returned in an Excel file containing all the SSRs, primers, ORFs in sequence and product size with summary information on motif number and type.

Community resources

GDR provides access to community-based news on various pages under the ‘community’ header bar, such as Rosaceae genomics, USRosEXEC, conferences, meetings, funding, employment, mailing lists and message

boards. USRosEXEC stands for US Rosaceae Genomics, Genetics and Breeding Executive Committee, which serves as a communication and coordination focal point for the community. The USRosEXEC page provides the official documents, meeting minutes, membership and subcommittee information. Several mailing lists, in addition to the GDR mailing list, are available to serve the community with information for specific interests or purposes, and the archives can be viewed through the message board sites. All the publications in Rosaceae genomics and genetics are also available in GDR through the publication search site.

Future directions

With the availability of the whole genome sequences of apple, peach and strawberry, and other Rosaceae genomes and re-sequencing data to be added, the future direction of GDR will include further integrating the annotated whole genome data with other genomics, genetics and breeding data to accelerate the usage of DNA information in crop improvement as well as to improve our knowledge on various aspects of Rosaceae biology. The genetic markers will also be queried by the anchored genome position and the neighboring trait locus. The effort toward collecting and curating trait data, including MTL and QTL, will also be continued. When data for genetic markers, DNA polymorphism and trait data are integrated and easily searchable, users will be able to query for markers that are close to the trait of interest and polymorphic between the varieties of interest. The Genome Sequence Annotation Server (GenSAS) (Lee et al. 2011), an online annotation tool that provides a customizable automated pipeline for whole genome structural annotation, will be available for community for further curate the genome. In collaboration with RosBreed project, more breeding decision tools will be developed. Growers' gateway that allows growers to get information about cultivars and compare the cultivar performances will also be developed.

Conclusion

The recent addition of publicly available whole genome sequences of three Rosaceae species to the already existing wealth of genetic data has brought great opportunity for researchers to accelerate Rosaceae research. Multiple whole genome sequences in one family allow genome level comparative analysis to gain evolutionary insight as well as transfer knowledge among species. The second and third generations of high-throughput sequencing technology now allows resequencing of multiple varieties to catalogue sequence variations and quantitative gene expression analyses. The proteome and metabolome analyses are still in their infancy in Rosaceae, but the advances of tools and methodology with other model species will aid greatly in planning and adopting those analyses in Rosaceae research. Integration of different types data is critical in the interpretation and utilization of these data and hence the role of the community databases to bring together genomics, genetics, breeders and growers data will become increasingly important.

References

Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B, Moreno RF et al (1991) Complementary DNA sequencing: expressed sequence tags and human genome project. Science 252:1651–1656

Aharoni A, Keizer LC, Van Den Broeck HC, Blanco-Portales R, Muñoz-Blanco J, Bois G, Smit P, De Vos RC, O'Connell AP (2002) Novel insight into vascular, stress, and auxin-dependent and -independent gene expression programs in strawberry, a non-climacteric fruit. Plant Physiol 129:1019–1031

Aharoni A, Giri AP, Verstappen FW, Bertea CM, Sevenier R, Sun Z, Jongsma MA, Schwab W, Bouwmeester HJ (2004) Gain and loss of fruit flavor compounds produced by wild and cultivated strawberry species. Plant Cell 16:3110–3131

Ahmad R, Parfitt DE, Fass J, Ogundiwin E, Dhingra A, Gradziel TA, Lin D, Joshi NA, Martinez-Garcia PJ, Crisosto CH (2011) Whole genome sequencing of peach (*Prunus persica* L.) for SNP identification and selection. BMC Genomics 12:569

Andreeva A, Howorth D, Chandonia JM, Brenner SE, Hubbard TJP, Chothia C, Murzin AG (2008) Data growth and its impact on the SCOP database: new developments. Nucleic Acids Res 36:D419–D425

Aoki KF, Kanehisa M (2005) Using the KEGG database resource. Curr Protoc Bioinforma Chapter 1:Unit 1.12

Audergon JM, Ruiz D, Bachellez A, Blanc A, Corre MN, Croset C, Ferréol AM, Lambert P, Pascal T, Poëssel JL, Signoret V, Quilot B, Gouble B, Grotte M, Bogé M, Reiling P, Reich M, Bureau S, Boudehri K, Renaud C, Dirlewanger E, Deborde C, Maucourt M, Moing A, Monllor S, Dondini L, Arús P (2009) ISAFRUIT: study of the genetic basis of *Prunus* fruit quality in two peach and two apricot populations. Acta Hortic 814:523–528

Baerenfaller K, Grossmann J, Grobei MA, Hull R, Hirsch-Hoffmann M, Yalovsky S, Zimmermann P, Grossniklaus U, Gruissem W, Baginsky S (2008) Genome-scale proteomics reveals *Arabidopsis thaliana* gene models and proteome dynamics. Science 320:938–941

Batzoglou S, Jaffe DB, Stanley K, Butler J, Gnerre S, Mauceli E, Berger B, Mesirov JP, Lander ES (2002) ARACHNE: a whole genome shotgun assembler. Genome Res 13:91–96

Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW (2013) GenBank. Nucleic Acids Res 41:D36–D42

Berman H, Henrick K, Nakamura H, Markley JL (2007) The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. Nucleic Acids Res 35:D301–D303

Bink MC, van Eeuwijk FA (2009) A Bayesian QTL linkage analysis of the common dataset from the 12th QTLMAS workshop. BMC Proc 3(Suppl 1):S4

Bombarely A, Merchante C, Csukasi F, Cruz-Rus E, Caballero JL, Medina-Escobar N, Blanco-Portales R, Botella MA, Muñoz-Blanco J, Sánchez-Sevilla JF, Valpuesta V (2010) Generation and analysis of ESTs from strawberry (*Fragaria* × *ananassa*) fruits and evaluation of their utility in genetic and molecular studies. BMC Genomics 11:503

Bombarely A, Menda N, Tecle IY, Buels RM, Strickler S, Fischer-York T, Pujar A, Leto J, Gosselin J, Mueller LA (2011) The sol genomics network (solgenomics.net): growing tomatoes using Perl. Nucleic Acids Res 39:D1149–D1155

Botton A, Eccher G, Forcato C, Ferrarini A, Begheldo M, Zermiani M, Moscatello S, Battistelli A, Velasco R, Ruperti B, Ramina A (2011) Signaling pathways mediating the induction of apple fruitlet abscission. Plant Physiol 155:185–208

Cabrera A, Kozik A, Howad W, Arus P, Iezzoni AF, van der Knaap E (2009) Development and bin mapping of a Rosaceae conserved ortholog set (COS) of markers. BMC Genomics 10:562

Castellana NE, Payne SH, Shen Z, Stanke M, Bafna V, Briggs SP (2008) Discovery and revision of *Arabidopsis* genes by proteogenomics. Proc Natl Acad Sci USA 105:21034–21038

Chagne D, Crowhurst R, Troggio M, Davey M, Gilmore B, Lawley C, Vanderzante S, Hellens R, Kumar S, Cestaro A, Velasco R, Main D, Rees J, Iezzoni A, Mockler T, Wilhelm L, van de Weg E, Gardiner S, Bassil N, Peace C (2012) Genome-wide SNP detection, validation, and development of an 8K SNP array for apple. PLoS ONE 7(2):e31745. doi:10.1371/journal.pone.0031745

Cochrane G, Alako B, Amid C, Bower L, Cerdeño-Tárraga A, Cleland I, Gibson R, Goodgame N, Jang M, Kay S, Leinonen R, Lin X, Lopez R, McWilliam H, Oisel A, Pakseresht N, Pallreddy S, Park Y, Plaister S, Radhakrishnan R, Rivière S, Rossello M, Senf A, Silvester N, Smirnov D, Ten Hoopen P, Toribio A, Vaughan D, Zalunin V (2013) Facing growth in the European nucleotide archive. Nucleic Acids Res 41:D30–D35

Costa F, Alba R, Schouten H, Soglio V, Gianfranceschi L, Serra S, Musacchi S, Sansavini S, Costa G, Fei Z, Giovannoni J (2010) Use of homologous and heterologous gene expression profiling tools to characterize transcription dynamics during apple fruit maturation and ripening. BMC Plant Biol 10:229

Cuff AL, Sillitoe I, Lewis T, Clegg AB, Rentzsch R, Furnham N, Pellegrini-Calace M, Jones D, Thornton J, Orengo CA (2011) Extending CATH: increasing coverage of the protein structure universe and linking structure with function. Nucleic Acids Res 39:D420–D426

Dardick CD, Callahan AM, Chiozzotto R, Schaffer RJ, Piagnani MC, Scorza R (2010) Stone formation in peach fruit exhibits spatial coordination of the lignin and flavonoid pathways and similarity to Arabidopsis dehiscence. BMC Biol 8:13

Dewey CN (2007) Aligning multiple whole genomes with Mercator and MAVID. Methods Mol Biol 395:221–236

Diamant J, Sguigna V, Mezzetti B, Faedi W, Maltoni ML, Denoyes B, Chartier P, Petit A (2012) European small berries genetic resources, GENBERRY: testing a protocol for detecting fruit nutritional quality in EU strawberry germplasm collections. Acta Hortic 926:33–37

Dirlewanger E, Graziano E, Joobeur T, Garriga-Caldere F, Cosson P, Howad W, Arus P (2004) Comparative mapping and marker-assisted selection in Rosaceae fruit crops. Proc Natl Acad Sci USA 101:9891–9896

Donlin MJ (2007) Using the Generic Genome Browser (GBrowse). Curr Protoc Bioinforma Chapter 9:Unit 9.9

Filichkin SA, Priest HD, Givan SA, Shen RK, Bryant DW, Fox SE, Wong WK, Mockler TC (2010) Genome-wide mapping of alternative splicing in *Arabidopsis thaliana*. Gen Res 20:45–58

Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, Forslund K, Holm L, Sonnhammer EL, Eddy SR, Bateman A (2010) The Pfam protein families database. Nucleic Acids Res 38:D211–D222

Geisler-Lee J, O'Toole N, Ammar R, Provart NJ, Millar AH, Geisler M (2007) A predicted interactome for *Arabidopsis*. Plant Physiol 145:317–329

Georgi LL, Wang Y, Yverggniaux D, Ormsbee T, Inigo M, Reighard GL, Abbott AG (2002) Construction of a BAC library and its application to the identification of simple sequence repeats in peach [*Prunus* persica (L.) Batsch]. Theor Appl Genet 105:1151–1158

Gianfranceschi L, Soglio V (2004) The European project HiDRAS: innovative multidisciplinary approaches to breeding high quality disease resistant apples. Acta Hortic 663:327–330

Gibney G, Baxevanis AD (2011) Searching NCBI databases using Entrez. Curr Protoc Bioinforma Chapter 1:Unit 1.3

Haas BJ, Delcher AL, Wortman JR, Salzberg SL (2004) DAGchainer: a tool for mining segmental genome duplications and synteny. Bioinformatics 20:3643–3646

Heazlewood JL, Verboom RE, Tonti-Filippini J, Small I, Millar AH (2007) SUBA: the *Arabidopsis* subcellular database. Nucleic Acids Res 35:D213–D218

Horn R, Lecouls AC, Callahan A, Dandekar A, Garay L, McCord P, Howad W, Chan H, Verde I, Main D, Jung S, Georgi L, Forrest S, Mook J, Zhebentyayeva T, Yu Y, Kim HR, Jesudurai C, Sosinski B, Arús P, Baird V, Parfitt D, Reighard G, Scorza R, Tomkins J, Wing R, Abbott AG (2005) Candidate gene database and transcript map for peach, a model species for fruit trees. Theor Appl Genet 110:1419–1428

Howad W, Yamamoto T, Dirlewanger E, Testolin R, Cosson P, Cipriani G, Monforte AJ, Georgi L, Abbott AG, Arús P (2005) Mapping with a few plants: using selective mapping for microsatellite saturation of the *Prunus* reference map. Genetics 171:1305–1309

Huang S, Li R, Zhang Z, Li L, Gu X, Fan W, Lucas WJ, Wang X, Xie B, Ni P, Ren Y, Zhu H, Li J, Lin K, Jin W, Fei Z, Li G, Staub J, Kilian A, van der Vossen EA, Wu Y, Guo J, He J, Jia Z, Ren Y, Tian G, Lu Y, Ruan J, Qian W, Wang M, Huang Q, Li B, Xuan Z, Cao J, Asan WuZ, Zhang J, Cai Q, Bai Y, Zhao B, Han Y, Li Y, Li X, Wang S, Shi Q, Liu S, Cho WK, Kim JY, Xu Y, Heller-Uszynska K, Miao H, Cheng Z, Zhang S, Wu J, Yang Y, Kang H, Li M, Liang H, Ren X, Shi Z, Wen M, Jian M, Yang H, Zhang G, Yang Z, Chen R, Liu S, Li J, Ma L, Liu H, Zhou Y, Zhao J, Fang X, Li G, Fang L, Li Y, Liu D, Zheng H, Zhang Y, Qin N, Li Z, Yang G, Yang S, Bolund L, Kristiansen K, Zheng H, Li S, Zhang X, Yang H, Wang J, Sun R, Zhang B, Jiang S, Wang J, Du Y, Li S (2009) The genome of the cucumber, *Cucumis sativus* L. Nat Genet 41:1275–1281

Hummel J, Selbig J, Walther D, Kopka J (2007) The Golm metabolome database: a database for GC–MS based metabolite profiling. In: Nielsen J, Jewett MC (eds) Metabolomics. Springer, Berlin, pp 75–96

Hummel J, Strehmel N, Selbig J, Walther D, Kopka J (2010) Decision tree supported substructure prediction of metabolites from GC–MS profiles. Metabolomics 6:322–333

Hunter S, Jones P, Mitchell A, Apweiler R, Attwood TK, Bateman A, Bernard T, Binns D, Bork P, Burge S, de Castro E, Coggill P, Corbett M, Das U, Daugherty L, Duquenne L, Finn RD, Fraser M, Gough J, Haft D, Hulo N, Kahn D, Kelly E, Letunic I, Lonsdale D, Lopez R, Madera M, Maslen J, McAnulla C, McDowall J, McMenamin C, Mi H, Mutowo-Muellenet P, Mulder N, Natale D, Orengo C, Pesseat S, Punta M, Quinn AF, Rivoire C, Sangrador-Vegas A, Selengut JD, Sigrist CJ, Scheremetjew M, Tate J, Thimmajanarthanan M, Thomas PD, Wu CH, Yeats C, Yong SY (2012) InterPro in 2011: new developments in the family and domain prediction database. Nucleic Acids Res 40:D306–D312

Iezzoni A, Weebadde C, Luby J, Yue C, Peace C, Bassil N, McFerson J (2010) RosBREED: enabling marker-assisted breeding in Rosaceae. Acta Hortic 859:389–394

Ilic K, Kellogg EA, Jaiswal P, Zapata F, Stevens PF, Vincent LP, Avraham S, Reiser L, Pujar A, Sachs MM, Whitman NT, McCouch SR, Schaeffer ML, Ware DH, Stein LD, Rhee SY (2007) The plant structure ontology, a unified vocabulary of anatomy and morphology of a flowering plant. Plant Physiol 143:587–599

International Rice Genome Sequencing Project (2005) The map-based sequence of the rice genome. Nature 436:793–800

Isserlin R, El-Badrawi RA, Bader GD (2011) The biomolecular interaction network database in PSI-MI 2.5. Database 2011:baq037

Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C, Vezzi A, Legeai F, Hugueney P, Dasilva C, Horner D, Mica E, Jublot D, Poulain J, Bruyère C, Billault A, Segurens B, Gouyvenoux M, Ugarte E, Cattonaro F, Anthouard V, Vico V, Del Fabbro C, Alaux M, Di Gaspero G, Dumas V, Felice N, Paillard S, Juman I, Moroldo M, Scalabrin S, Canaguier A, Le Clainche I, Malacrida G, Durand E, Pesole G, Laucou V, Chatelet P, Merdinoglu D, Delledonne M, Pezzotti M, Lecharny A, Scarpelli C, Artiguenave F, Pè ME, Valle G, Morgante M, Caboche M, Adam-Blondon AF, Weissenbach J, Quétier F, Wincker P, French-Italian Public Consortium for Grapevine Genome Characterization (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature 449:463–467

Janssen BJ, Thodey K, Schaffer RJ, Alba R, Balakrishnan L, Bishop R, Bowen JH, Crowhurst RN, Gleave AP, Ledger S, McArtney S, Pichler FB, Snowden KC, Ward S (2008) Global gene expression analysis of apple fruit development from the floral bud to ripe fruit. BMC Plant Biol 8:16

Jensen PJ, Halbrendt N, Fazio G, Makalowska I, Altman N, Praul C, Maximova SN, Ngugi HK, Crassweller RM, Travis JW, McNellis TW (2012) Rootstock-regulated gene expression patterns associated with fire blight resistance in apple. BMC Genomics 13:9

Jung S, Staton M, Lee T, Blenda A, Svancara R, Abbott A, Main D (2008) GDR (Genome Database for Rosaceae): integrated web-database for Rosaceae genomics and genetics data. Nucleic Acids Res 36:D1034–D1040

Jung S, Cestaro A, Troggio M, Main D, Zheng P, Cho I, Folta KM, Sosinski B, Abbott A, Celton JM, Arús P, Shulaev V, Verde I, Morgante M, Rokhsar D, Velasco R, Sargent DJ (2012) Whole genome comparisons of *Fragaria*, *Prunus* and *Malus* reveal different modes of evolution between Rosaceous subfamilies. BMC Genomics 13:129

Kaminuma E, Kosuge T, Kodama Y, Aono H, Mashima J, Gojobori T, Sugawara H, Ogasawara O, Takagi T, Okubo K, Nakamura Y (2011) DDJB progress report. Nucleic Acids Res 39:D22–D27

Khan MA, Han Y, Zhao YF, Korban SS (2012) A high-throughput apple SNP genotyping platform using the GoldenGate™ assay. Gene 492:196–201

Krost C, Petersen R, Lokan S, Brauksiepe B, Braun P, Schmidt ER (2013) Evaluation of the hormonal state of columnar apple trees (*Malus* × *domestica*) based on high throughput gene expression studies. Plant Mol Biol 81:211–220

Lalli DA, Decroocq V, Blenda AV, Schurdi-Levraud V, Garay L, Le Gall O, Damsteegt V, Reighard GL, Abbott AG (2005) Identification and mapping of resistance gene analogs (RGAs) in *Prunus*: a resistance map for *Prunus*. Theor Appl Genet 111:1504–1513

Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, Muller R, Dreher K, Alexander DL, Garcia-Hernandez M, Karthikeyan AS, Lee CH, Nelson WD, Ploetz L, Singh S, Wensel A, Huala E (2012) The *Arabidopsis* information resource (TAIR): improved gene annotation and new tools. Nucleic Acids Res 40:D1202–D1210

Lazzari B, Caprera A, Vecchietti A, Merelli I, Barale F, Milanesi L, Stella A, Pozzi C (2008) Version VI of the ESTree db: an improved tool for peach transcriptome analysis. BMC Bioinforma 9(Suppl 2):S9

Lee YP, Yu GH, Seo YS, Han SE, Choi YO, Kim D, Mok IG, Kim WT, Sung SK (2007) Microarray analysis of apple gene expression engaged in early fruit development. Plant Cell Rep 26:917–926

Lee T, Cho I, Peace C, Jung S, Zheng P, Main D (2011) Development approach and architecture of GenSAS BIOCOMP'11, July 2011, Las Vegas

Lee TH, Tang H, Wang X, Paterson AH (2013) PGDD: a database of gene and genome duplication in plants. Nucleic Acids Res 41:D1152–D1158

Li Y, Wu B, Yu Y, Yang G, Wu C, Zheng C (2011) Genome-wide analysis of the RING finger gene family in apple. Mol Genet Genomics 286:81–94

Li X, Korir NK, Liu L, Shangguan L, Wang Y, Han J, Chen M, Fang J (2012) Microarray analysis of differentially expressed genes engaged in fruit development between *Prunus mume* and *Prunus armeniaca*. J Plant Physiol 169:1776–1788

Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, Ecker JR (2008) Highly integrated single-base resolution maps of the epigenome in Arabidopsis. Cell 133:523–536

Lyons E, Freeling M (2008) How to usefully compare homologous plant genes and chromosomes as DNA sequences. Plant J 53:661–673

Martínez-Gómez P, Crisosto CH, Bonghi C, Rubio M (2011) New approaches to *Prunus* transcriptome analysis. Genetica 139:755–769

Matsuda F, Hirai MY, Sasaki E, Akiyama K, Yonekura-Sakakibara K, Provart NJ, Sakurai T, Shimada Y, Saito K (2010) AtMetExpress development: a phytochemical atlas of Arabidopsis development. Plant Physiol 152:566–578

McKay SJ, Vergara IA, Stajich JE (2010) Using the Generic Synteny Browser (GBrowse_syn). Curr Protoc Bioinforma Chapter 9:Unit 9.12

Mi H, Dong Q, Muruganujan A, Gaudet P, Lewis S, Thomas PD (2010) PANTHER version 7: improved phylogenetic trees, orthologs and collaboration with the Gene Ontology Consortium. Nucleic Acids Res 38:D204–D210

Mihara M, Itoh T, Izawa T (2010) SALAD database: a motif-based database of protein annotations for plant comparative genomics. Nucleic Acids Res 38:D835–D842

Newcomb RD, Crowhurst RN, Gleave AP, Rikkerink EHA, Allan AC, Beuning LL, Bowen JH, Gera E, Jamieson KR, Janssen BJ et al (2006) Analyses of expressed sequence tags from apple. Plant Physiol 141:147–166

Ogundiwin EA, Martí C, Forment J, Pons C, Granell A, Gradziel TM, Peace CP, Crisosto CH (2008) Development of ChillPeach genomic tools and identification of cold-responsive genes in peach fruit. Plant Mol Biol 68:379–397

Pampanwar V, Engler F, Hatfield J, Blundy S, Gupta G, Soderlund C (2005) FPC web tools for rice, maize, and distribution. Plant Physiol 138:116–126

Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A, Schmutz J, Spannagl M, Tang H, Wang X, Wicker T, Bharti AK, Chapman J, Feltus FA, Gowik U, Grigoriev IV, Lyons E, Maher CA, Martis M, Narechania A, Otillar RP, Penning BW, Salamov AA, Wang Y, Zhang L, Carpita NC, Freeling M, Gingle AR, Hash CT, Keller B, Klein P, Kresovich S, McCann MC, Ming R,

Peterson DG, Mehboob-ur-Rahman, Ware D, Westhoff P, Mayer KF, Messing J, Rokhsar DS (2009) The Sorghum bicolor genome and the diversification of grasses. Nature 457:551–556

Peace C, Bassil N, Main D, Ficklin S, Rosyara UR, Stegmeir T, Sebolt A, Gilmore B, Lawley C, Mockler TC, Bryant DW, Wilhelm L, Iezzoni A (2012) Development and evaluation of a genome-wide 6K SNP array for diploid sweet cherry and tetraploid sour cherry. PLoS ONE 7:e48305

Popescu SC, Popescu GV, Bachan S, Zhang Z, Seay M, Gerstein M, Snyder M, Dinesh-Kumar SP (2007) Differential binding of calmodulin-related proteins to their targets revealed through high-density *Arabidopsis* protein microarrays. Proc Natl Acad Sci USA 104:4730–4735

Popescu SC, Popescu GV, Bachan S, Zhang Z, Gerstein M, Snyder M, Dinesh-Kumar SP (2009) MAPK target networks in *Arabidopsis thaliana* revealed using functional protein microarrays. Genes Dev 23:80–92

Pruitt KD, Tatusova T, Maglott DR (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 35:D61–D65

Rouard M, Guignon V, Aluome C, Laporte MA, Droc G, Walde C, Zmasek CM, Périn C, Conte MG (2011) GreenPhylDB v2.0: comparative and functional genomics in plants. Nucleic Acids Res 39:D1095–D1102

Santos AM, Oliver MJ, Sánchez AM, Payton PR, Gomes JP, Miguel C, Oliveira MM (2009) An integrated strategy to identify key genes in almond adventitious shoot regeneration. J Exp Bot 60:4159–4173

Sargent DJ, Kuchta P, Lopez Girona E, Zhang H, Davis TM, Celton JM, Marchese A, Korbin M, Folta KM, Shulaev V, Simpson DW (2011) Simple sequence repeat marker development and mapping targeted to previously unmapped regions of the strawberry genome sequence. Plant Genome 4:165–177

Schaeffer ML, Harper LC, Gardiner JM, Andorf CM, Campbell DA, Cannon EK, Sen TZ, Lawrence CJ (2011) MaizeGDB: curation and outreach go hand-in-hand. Database 2011:bar022

Schaffer RJ, Friel EN, Souleyre EJ, Bolitho K, Thodey K, Ledger S, Bowen JH, Ma JH, Nain B, Cohen D, Gleave AP, Crowhurst RN, Janssen BJ, Yao JL, Newcomb RD (2007) A genomics approach reveals that aroma production in apple is controlled by ethylene predominantly at the final step in each biosynthetic pathway. Plant Physiol 144:1899–1912

Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J, Xu D, Hellsten U, May GD, Yu Y, Sakurai T, Umezawa T, Bhattacharyya MK, Sandhu D, Valliyodan B, Lindquist E, Peto M, Grant D, Shu S, Goodstein D, Barry K, Futrell-Griggs M, Abernathy B, Du J, Tian Z, Zhu L, Gill N, Joshi T, Libault M, Sethuraman A, Zhang XC, Shinozaki K, Nguyen HT, Wing RA, Cregan P, Specht J, Grimwood J, Rokhsar D, Stacey G, Shoemaker RC, Jackson SA (2010) Genome sequence of the palaeopolyploid soybean. Nature 463:178–83. Erratum in: Nature 465:120

Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA, Minx P, Reily AD, Courtney L, Kruchowski SS, Tomlinson C, Strong C, Delehaunty K, Fronick C, Courtney B, Rock SM, Belter E, Du F, Kim K, Abbott RM, Cotton M, Levy A, Marchetto P, Ochoa K, Jackson SM, Gillam B, Chen W, Yan L, Higginbotham J, Cardenas M, Waligorski J, Applebaum E, Phelps L, Falcone J, Kanchi K, Thane T, Scimone A, Thane N, Henke J, Wang T, Ruppert J, Shah N, Rotter K, Hodges J, Ingenthron E, Cordes M, Kohlberg S, Sgro J, Delgado B, Mead K, Chinwalla A, Leonard S, Crouse K, Collura K, Kudrna D, Currie J, He R, Angelova A, Rajasekar S, Mueller T, Lomeli R, Scara G, Ko A, Delaney K, Wissotski M, Lopez G, Campos D, Braidotti M, Ashley E, Golser W, Kim H, Lee S, Lin J, Dujmic Z, Kim W, Talag J, Zuccolo A, Fan C, Sebastian A, Kramer M, Spiegel L, Nascimento L, Zutavern T, Miller B, Ambroise C, Muller S, Spooner W, Narechania A, Ren L, Wei S, Kumari S, Faga B, Levy MJ, McMahan L, Van Buren P, Vaughn MW, Ying K, Yeh CT, Emrich SJ, Jia Y, Kalyanaraman A, Hsia AP, Barbazuk WB, Baucom RS, Brutnell TP, Carpita NC, Chaparro C, Chia JM, Deragon JM, Estill JC, Fu Y, Jeddeloh JA, Han Y, Lee H, Li P, Lisch DR, Liu S, Liu Z, Nagel DH, McCann MC, SanMiguel P, Myers AM, Nettleton D, Nguyen J, Penning BW, Ponnala L, Schneider KL, Schwartz DC, Sharma A, Soderlund C, Springer NM, Sun Q, Wang H, Waterman M, Westerman R, Wolfgruber TK, Yang L, Yu Y, Zhang L, Zhou S, Zhu Q, Bennetzen JL, Dawe RK, Jiang J, Jiang N, Presting GG, Wessler SR, Aluru S, Martienssen RA, Clifton SW, McCombie WR, Wing RA, Wilson RK (2009) The B73 maize genome: complexity, diversity, and dynamics. Science 326:1112–1115

Schwacke R, Schneider A, van der Graaff E, Fischer K, Catoni E, Desimone M, Frommer WB, Flügge UI, Kunze R (2003) ARAMEMNON, a novel database for *Arabidopsis* integral membrane proteins. Plant Physiol 131:16–26

Shizuya H, Birren B, Kim UJ, Mancino V, Slepak T, Tachiiri Y, Simon M (1992) Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in *Escherichia coli* using an F-factor-based vector. Proc Natl Acad Sci USA 89:8794–8797

Shulaev V, Korban SS, Sosinski B, Abbott AG, Aldwinckle HS, Folta KM, Iezzoni A, Main D, Arús P, Dandekar AM, Lewers K, Brown SK, Davis TM, Gardiner SE, Potter D, Veilleux RE (2008) Multiple models for Rosaceae genomics. Plant Physiol 147:985–1003

Shulaev V, Sargent DJ, Crowhurst RN, Mockler TC, Folkerts O, Delcher AL, Jaiswal P, Mockaitis K, Liston A, Mane SP, Burns P, Davis TM, Slovin JP, Bassil N, Hellens RP, Evans C, Harkins T, Kodira C, Desany B, Crasta OR, Jensen RV, Allan AC, Michael TP, Setubal JC, Celton JM, Rees DJ, Williams KP, Holt SH, Ruiz Rojas JJ, Chatterjee M, Liu B, Silva H, Meisel L, Adato A, Filichkin SA, Troggio M, Viola R, Ashman TL, Wang H, Dharmawardhana P, Elser J, Raja R, Priest HD, Bryant DW Jr, Fox SE, Givan SA, Wilhelm LJ, Naithani S, Christoffels A, Salama DY, Carter J, Lopez Girona E, Zdepski A, Wang W, Kerstetter RA, Schwab W, Korban SS, Davik J, Monfort A, Denoyes-Rothan B, Arus P, Mittler R, Flinn B, Aharoni A, Bennetzen JL, Salzberg SL, Dickerman AW, Velasco R, Borodovsky M, Veilleux RE, Folta KM (2011) The genome of woodland strawberry (*Fragaria vesca*). Nat Genet 43:109–116

Simillion C, Janssens K, Sterck L, Van de Peer Y (2008) i-ADHoRe 2.0: an improved tool to detect degenerated genomic homology using genomic profiles. Bioinformatics 24:127–128

Spiller M, Linde M, Hibrand-Saint Oyant L, Tsai CJ, Byrne DH, Smulders MJ, Foucher F, Debener T (2011) Towards a unified genetic map for diploid roses. Theor Appl Genet 122: 489–500

Stein LD, Mungall C, Shu S, Caudy M, Mangone M, Day A, Nickerson E, Stajich JE, Harris TW, Arva A, Lewis S (2002) The Generic Genome Browser: a building block for a model organism system database. Genome Res 12:1599–1610

Tang H, Bowers JE, Wang X, Ming R, Alam M, Paterson AH (2008) Synteny and collinearity in plant genomes. Science 320:486–488

Tatsuki M, Nakajima N, Fujii H, Shimada T, Nakano M, Hayashi KI, Hayama H, Yoshioka H, Nakamura Y (2013) Increased levels of IAA are required for system 2 ethylene synthesis causing fruit softening in peach (*Prunus persica* L. Batsch). J Exp Bot 64(4):1049–1059. doi:10.1093/jxb/ers381

Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf

YI, Yin JJ, Natale DA (2003) The COG database: an updated version includes eukaryotes. BMC Bioinforma 4:41

The Gene Ontology Consortium (2013) Gene ontology annotations and resources. Nucleic Acids Res 41:D530–D535

The UniProt Consortium (2012) Reorganizing the protein space at the Universal Protein Resource (UniProt). Nucleic Acids Res 40:D71–D75

Trainotti L, Zanin D, Casadoro G (2003) A cell wall-oriented genomic approach reveals a new and unexpected complexity of the softening in peaches. J Exp Bot 54:1821–1831

Van Bel M, Proost S, Wischnitzki E, Movahedi S, Scheerlinck C, Van de Peer Y, Vandepoele K (2012) Dissecting plant genomes with the PLAZA comparative genomics platform. Plant Physiol 158:590–600

Vecchietti A, Lazzari B, Ortugno C, Bianchi F, Malinverni R, Caprera A, Mignani I, Pozzi C (2009) Comparative analysis of expressed sequence tags from tissues in ripening stages of peach (*Prunus persica* L. Batsch). Tree Genet Genome 5:377–391

Velankar S, Alhroub Y, Alili A, Best C, Boutselakis HC, Caboche S, Conroy MJ, Dana JM, van Ginkel G, Golovin A, Gore SP, Gutmanas A, Haslam P, Hirshberg M, John M, Lagerstedt I, Mir S, Newman LE, Oldfield TJ, Penkett CJ, Pineda-Castillo J, Rinaldi L, Sahni G, Sawka G, Sen S, Slowley R, Sousa da Silva AW, Suarez-Uruena A, Swaminathan GJ, Symmons MF, Vranken WF, Wainwright M, Kleywegt GJ (2011) PDBe: Protein Data Bank in Europe. Nucleic Acids Res 39(Suppl 1):D402–D410

Velasco R, Zharkikh A, Affourtit J, Dhingra A, Cestaro A, Kalyanaraman A, Fontana P, Bhatnagar SK, Troggio M, Pruss D, Salvi S, Pindo M, Baldi P, Castelletti S, Cavaiuolo M, Coppola G, Costa F, Cova V, Dal Ri A, Goremykin V, Komjanc M, Longhi S, Magnago P, Malacarne G, Malnoy M, Micheletti D, Moretto M, Perazzolli M, Si-Ammour A, Vezzulli S, Zini E, Eldredge G, Fitzgerald LM, Gutin N, Lanchbury J, Macalma T, Mitchell JT, Reid J, Wardell B, Kodira C, Chen Z, Desany B, Niazi F, Palmer M, Koepke T, Jiwan D, Schaeffer S, Krishnan V, Wu C, Chu VT, King ST, Vick J, Tao Q, Mraz A, Stormo A, Stormo K, Bogden R, Ederle D, Stella A, Vecchietti A, Kater MM, Masiero S, Lasserre P, Lespinasse Y, Allan AC, Bus V, Chagné D, Crowhurst RN, Gleave AP, Lavezzo E, Fawcett JA, Proost S, Rouzé P, Sterck L, Toppo S, Lazzari B, Hellens RP, Durel CE, Gutin A, Bumgarner RE, Gardiner SE, Skolnick M, Egholm M, Van de Peer Y, Salamini F, Viola R (2010) The genome of the domesticated apple (*Malus* × *domestica* Borkh.). Nat Genet 42:833–839

Verde I, Bassil N, Scalabrin S, Gilmore B, Lawley CT, Gasic K, Micheletti D, Rosyara UR, Cattonaro F, Vendramin E, Main D, Aramini V, Blas AL, Mockler TC, Bryant DW, Wilhelm L, Troggio M, Sosinski B, Aranzana MJ, Arús P, Iezzoni A, Morgante M, Peace C (2012) Development and evaluation of a 9 K SNP array for peach by internationally coordinated SNP detection and validation in breeding germplasm. PLoS ONE 7(4):e35668. doi:10.1371/journal.pone.0035668

Vizoso P, Meisel LA, Tittarelli A, Latorre M, Saba J, Caroca R, Maldonado J, Cambiazo V, Campos-Vargas R, Gonzalez M, Orellana A, Silva H (2009) Comparative EST transcript profiling of peach fruits under different post-harvest conditions reveals candidate genes associated with peach fruit quality. BMC Genomics 10:423

Wang K, Wang Z, Li F, Ye W, Wang J, Song G, Yue Z, Cong L, Shang H, Zhu S, Zou C, Li Q, Yuan Y, Lu C, Wei H, Gou C, Zheng Z, Yin Y, Zhang X, Liu K, Wang B, Song C, Shi N, Kohel RJ, Percy RG, Yu JZ, Zhu YX, Wang J, Yu S (2012) The draft genome of a diploid cotton *Gossypium raimondii*. Nat Genet 44:1098–1103

Weber APM, Weber KL, Carr K, Wilkerson C, Ohlrogge JB (2007) Sampling the *Arabidopsis* transcriptome with massively parallel pyrosequecing. Plant Physiol 144:32–42

Wegrzyn JL, Lee JM, Tearse BR, Neale DB (2008) TreeGenes: a forest tree genome database. Intl J Plant Genomics 2008:412875. doi:10.1155/2008/412875

Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, Federhen S, Geer LY, Kapustin Y, Khovayko O, Landsman D, Lipman DJ, Madden TL, Maglott DR, Ostell J, Miller V, Pruitt KD, Schuler GD, Sequeira E, Sherry ST, Sirotkin K, Souvorov A, Starchenko G, Tatusov RL, Tatusova TA, Wagner L, Yaschenko E (2007) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 35:D5–D12

Wilson D, Pethica R, Zhou Y, Talbot C, Vogel C, Madera M, Chothia C, Gough J (2009) SUPERFAMILY: sophisticated comparative genomics, data mining, visualization and phylogeny. Nucleic Acids Res 37:D380–D386

Wu J, Wang Z, Shi Z, Zhang S, Ming R, Zhu S, Khan MA, Tao S, Korban SS, Wang H, Chen NJ, Nishio T, Xu X, Cong L, Qi K, Huang X, Wang Y, Zhao X, Wu J, Deng C, Gou C, Zhou W, Yin H, Qin G, Sha Y, Tao Y, Chen H, Yang Y, Song Y, Zhan D, Wang J, Li L, Dai M, Gu C, Wang Y, Shi D, Wang X, Zhang H, Zeng L, Zheng D, Wang C, Chen M, Wang G, Xie L, Sovero V, Sha S, Huang W, Zhang S, Zhang M, Sun J, Xu L, Li Y, Liu X, Li Q, Shen J, Wang J, Paull RE, Bennetzen JL, Wang J, Zhang S (2013) The genome of the pear (*Pyrus bretschneideri* Rehd.). Genome Res 23:396–408

Xu Q, Chen LL, Ruan X, Chen D, Zhu A, Chen C, Bertrand D, Jiao WB, Hao BH, Lyon MP, Chen J, Gao S, Xing F, Lan H, Chang JW, Ge X, Lei Y, Hu Q, Miao Y, Wang L, Xiao S, Biswas MK, Zeng W, Guo F, Cao H, Yang X, Xu XW, Cheng YJ, Xu J, Liu JH, Luo OJ, Tang Z, Guo WW, Kuang H, Zhang HY, Roose ML, Nagarajan N, Deng XX, Ruan Y (2012) The draft genome of sweet orange (*Citrus sinensis*). Nat Genet 45:59–66

Yao Q, Bollinger C, Gao J, Xu D, Thelen JJ (2012) P3DB: an integrated database for plant protein phosphorylation. Front Plant Sci 3:206

Yi G, Sze SH, Thon MR (2007) Identifying clusters of functionally related genes in genome. Bioinformatics 23:1053–1060

Youens-Clark K, Buckler E, Casstevens T, Chen C, Declerck G, Derwent P, Dharmawardhana P, Jaiswal P, Kersey P, Karthikeyan AS, Lu J, McCouch SR, Ren L, Spooner W, Stein JC, Thomason J, Wei S, Ware D (2011) Gramene database in 2010: updates and extensions. Nucleic Acids Res 39:D1085–D1094

Zhang Q, Chen W, Sun L, Zhao F, Huang B, Yang W, Tao Y, Wang J, Yuan Z, Fan G, Xing Z, Han C, Pan H, Zhong X, Shi W, Liang X, Du D, Sun F, Xu Z, Hao R, Lv T, Lv Y, Zheng Z, Sun M, Luo L, Cai M, Gao Y, Wang J, Yin Y, Xu X, Cheng T, Wang J (2012) The genome of Prunus mume. Nat Commun 3:1318

Zhebentyayeva T, Swire-Clark G, Georgi L, Garay L, Juns S, Forrest S, Blenda A, Blackmon B, Mook J, Horn R, Howad W, Arus P, Main D, Tomkins J, Sosinski B, Baird W, Reighard G, Abbott A (2008) A framework physical map for peach, a model *Rosaceae* species. Tree Genet Genome 4:745–756

Zhu Y, Varanasi V, Zheng P, Main D, Curry E, Mattheis J (2012) Multiple plant hormones and cell wall metabolism regulate apple fruit maturation patterns and texture attributes. Tree Genet Genome 8:1389–1406

Ziliotto F, Begheldo M, Rasori A, Bonghi C, Tonutti P (2008) Transcriptome profiling of ripening nectarine (*Prunus persica* L. Batsch) fruit treated with 1-MCP. J Exp Bot 59:2781–2791

Ziosi V, Bonghi C, Bregoli AM, Trainotti L, Biondi S, Sutthiwal S, Kondo S, Costa G, Torrigiani P (2008) Jasmonate-induced transcriptional changes suggest a negative interference with the ripening syndrome in peach fruit. J Exp Bot 59:563–573