Bioinformatics Resources for Plant Genomics: Opportunities and Bottlenecks in the -omics Era

Luca Ambrosino[†], Chiara Colantuono[†], Francesco Monticolo and Maria Luisa Chiusano*

Department of Agricultural Sciences, University of Naples Federico II, Portici, Italy.

*Correspondence: chiusano@unina.it

[†]These authors contributed equally

https://doi.org/10.21775/cimb.027.071

Abstract

The sudden exponential increase of biological data concerning genome structure and functionalities, also fostered by the advent of next-generation sequencing technologies, while expanding the opportunity to highlight still uncovered molecular aspects, challenges bioinformatics in several respects. Data management, processing, updating, dissemination and integration are the major areas of concern.

The rapid increase in various omics technologies causes two major issues, which may even appear contrasting: the dissemination of poorly curated datasets, still in the form of raw collections or preliminary draft results, and the fast updating of information that, as a consequence, affects the establishment of stable reliable resources. These issues are mainly caused by the lower rate of bioinformatics in extracting added value information from the large number of data, when compared to the faster technologies involved in data production.

This review describes main bioinformatics resources for plants genomics to underline the heterogeneity of the available collections, coherent with the multifaceted complexity of plant sciences. It aims to provide an in-depth report highlighting bottlenecks that may significantly affect a fluent progress in the field and attempts to suggest possible solutions to the various issues.

The impact of plant genomics

The deciphering of the molecular mechanisms which determine plant diversity and adaptation to different environments and the impact that this knowledge may provide for sustainable productivity in the food industry, energy production or biotechnological applications (Blanchfield, 2004; Ma *et al.*, 2003; Wilson and Roberts, 2014; Yuan *et al.*, 2008) can be strongly supported by structural and functional genomics. These are among the main reasons why the scientific community is increasingly demanding fully sequenced plant genomes with the aim to exploit the advantages and the opportunities that genomics may offer in plant sciences. The release of the genome of *Arabidopsis thaliana* (The Arabidopsis Genome Initiative, 2000), represents a milestone in the field, making this relevant species in plant biology a model reference for plant genomics (Becker *et al.*, 2011; Gan *et al.*, 2011; Kim *et al.*, 2007). Nevertheless, the expanding of genomics has further highlighted the multifaceted complexity of plants, where complex genomes, often larger than those of mammals (Mayer *et al.*, 2012), present peculiar organizations and extensive duplications and reshuffling (Bowers *et al.*, 2003; Cui *et al.*, 2006; Flagel and Wendel, 2009; Hughes, 2005; Jiao *et al.*, 2012; Lynch and Conery, 2000; Maere *et al.*, 2005; Magadum *et al.*, 2013), revealing complex evolutionary histories often involving poliploidization events followed by diploidization and gene reduction (Blanc *et al.*, 2003; Jaillon *et al.*, 2007; Jiao *et al.*, 2011; Moniz de Sa and Drouin, 1996; Wolfe, 2001).

The recent establishment of novel, low-cost and easily accessible technologies is further encouraging the increase in genome sequencing efforts. Indeed, the recent introduction of next-generation sequencing (NGS) technologies, which replaced the conventional Sanger strategy (Sanger *et al.*, 1977), deeply reshaped the omics research areas (Bateman and Quackenbush, 2009; Esposito *et al.*, 2016; Wang *et al.*, 2015), contributing to several species-specific efforts usually curated by dedicated consortia (Fig. 4.1). The result is usually in the form of assembled draft genome sequences with quality levels established by the consortia requirements and by funding opportunities. It is common practice to accompany new genome assemblies with consortium curated gene annotations, which benefit from specific competences from the interested scientific community that is contributing to the consortium. Moreover, dedicated web resources are usually made available to convey the Consortia efforts into a unified framework. These efforts are consistently contributing to the research of some of the most relevant crops (Brenchley *et al.*, 2012; Choulet *et al.*, 2010; Jia *et al.*, 2013; Ling *et al.*, 2013; Nystedt *et al.*, 2013; Wang *et al.*, 2011). Besides, the reference database for the model plant *Arabidopsis thaliana* represented by The Arabidopsis Information Resource (TAIR) (Lamesch *et al.*, 2012), numerous consortia gave rise to species-specific platforms including results from the genome sequencing and/or gene annotations efforts. As an example, among the major efforts for relevant crops, the MSU Rice Genome Annotation Project (Kawahara *et al.*, 2013), funded by the National Science Foundation, provides sequence and annotation data for the rice genome, which was released as a first version in 2002 (Goff *et al.*, 2002). The grapevine genome was released by an Italian-French consortium and made available through two main websites maintained by its members: the Genoscope Institute website (Jaillon *et al.*, 2007) (www.genoscope.cns.fr/spip/) and the CRIBI website (http://genomes.cribi.unipd.it/grape/). CRIBI recently released an updated gene annotation (Vitulo *et al.*, 2014) on the same genome version. Worthy to note that this novel annotation version is not reported in the Genoscope website. On the other hand, the genome analysis of a heterozygous grapevine variety was also published (Velasco *et al.*, 2007), though the genome assembly was never publicly released. The *Sorghum bicolor* genome sequencing and annotation (Paterson *et al.*, 2009) was released by the Joint Genome

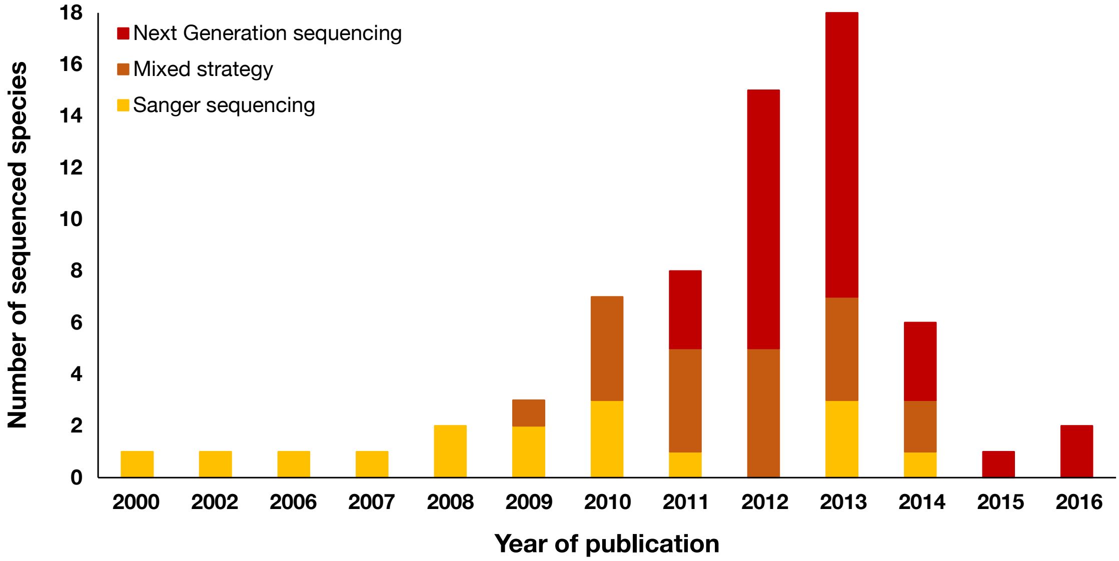

Figure 4.1 Number of plant genomes sequenced from 2000 (publication year of *Arabidopsis thaliana*) until today. The sequencing strategy is also highlighted.

Institute (JGI), which maintains a protected website. Thanks to the Maize Genome Sequencing Project, funded by the National Science Foundation, the complete genome sequence of *Zea mays* cv. B73 (Schnable *et al.*, 2009) was made available in MaizeGDB (Andorf *et al.*, 2016), and included in the collection available in Gramene (Tello-Ruiz *et al.*, 2016). The Potato Genome Sequencing Consortium (PGSC) released the first draft of the potato genome (Xu *et al.*, 2011), that was made available on the SpudDB website (Hirsch *et al.*, 2014) and in the Solanaceae Genomics Network collection. The International Tomato Genome Sequencing Project and the International Tomato Annotation Group (ITAG) have defined the sequence and the annotation of the tomato genome (The Tomato Genome Consortium, 2012), respectively, both released and maintained by the leading website of the Solanaceae Genomics Networks (SGN) (Fernandez-Pozo *et al.*, 2015), though offered also through several parallel dedicated platforms (Chiusano *et al.*, 2008; Hirsch *et al.*, 2014). The same consortium also defined the genome of *Solanum pimpinellifolium*, a wild species of the domesticated tomato, available in the form of genome scaffolds in the SGN platform (https://solgenomics.net/organism/Solanum_pimpinellifolium/genome). An international group of scientists from Korea, Israel and the USA sequenced and annotated the hot pepper genome (Kim *et al.*, 2014), which is available in the Pepper Genome Database (http://peppersequence.genomics.cn/page/species/index.jsp) and also in the SGN platform.

From genome structure to function

The rise of several independent projects for the sequencing of diverse plant genomes offers hints to understand their organization and functionality, revealing unknown molecular information and supporting scientific knowledge and the technological transfer of useful information (Esposito *et al.*, 2016). The understanding of genome functionalities, previously mainly supported by EST sequencing (Blair *et al.*, 2011; Izzah *et al.*, 2014; Wang *et al.*, 2005) and microarray technologies (Bülow *et al.*, 2007; Mukherjee *et al.*, 2005; Page and Coulibaly, 2008), is being unexpectedly favoured by the evolution of parallel -omics efforts providing enriched information to support the definition of genome structure organization and the investigation on its functionality (Bateman and Quackenbush, 2009). To this aim, advances in transcriptomics, epigenomics, proteomics and metabolomics, are consistently contributing novel data sources, useful to unravel hidden molecular aspects.

In particular, the amount of information provided by novel RNA sequencing technologies (RNA-seq), beyond contributing a deeper and expanded overview of gene expression levels, even for poorly expressed genes, supports their function profiling in different tissues and developmental stages, as well as in stress and pathological conditions. This improves the gene annotation, the identification of variants and the definition of expression patterns, useful to highlight specificities and peculiarities of the control and regulation of gene expression, providing an enriched snapshot of the transcriptome plasticity. Moreover, novel applications from NGS technologies in plants support the characterization of small and microRNAs, of genome methylated regions or of chromatin organization, also depicting protein binding sites (Becker *et al.*, 2003; Bokszczanin *et al.*, 2015; Horner *et al.*, 2010; MacLean *et al.*, 2009; Mardis, 2008a,b, 2009; Morozova and Marra, 2008a,b; Morrissy *et al.*, 2009; Schuster, 2008). These approaches (Adams *et al.*, 1991; Brenner *et al.*, 2000; Kodzius *et al.*, 2006; Velculescu *et al.*, 1995) generally accompanied the flourishing of genome sequencing projects of reference plant species, such as *Arabidopsis thaliana* (The Arabidopsis Genome Initiative, 2000). Similar efforts on other model or non-model species of agricultural interest soon followed. Some examples are rice (Goff *et al.*, 2002), grapevine (Jaillon *et al.*, 2007), maize (Schnable *et al.*, 2009), potato (Xu *et al.*, 2011), tomato (The Tomato Genome Consortium, 2012) genome sequencing which expanded the number of plant genomes currently available to the scientific community with enriched information favoured by alternative -omics approaches from transcriptomics, proteomics, epigenomics and metagenomics projects (Esposito *et al.*, 2016).

Bioinformatics data resources for plant genomics

The -omics efforts must be necessarily accompanied by bioinformatics (Schatz *et al.*, 2012) to translate data into knowledge. However, bioinformatics, followed by human curated result interpretations, represent the slowest phases of -omics research, when compared to the fast sequencing rate. The management, the analysis, the integration and the comparison of the large data number under production are, indeed, a major challenge in bioinformatics and in plant genomics as well (De Luca *et al.*, 2012). Computational methods and suitable bioinformatics are being evolving to accompany the evolution of the technologies and to face the need of added-value information, continuously improving and adapting to the increase of the amount of biological data. However, further novel solutions are still required (Esposito *et al.*, 2016).

Bioinformatics has been always fundamental also for the organization and the integration of molecular data collections, to offer access and suitable data views to all the interested scientific community, even to non-experts in -omics data management. Indeed, the release of integrated molecular information through dedicated web-based services or platforms has been always pushing the evolution of -omics research representing, since the beginning, an essential source of information in science. This is why general reference bioinformatics resources were attempted since the initial production of molecular information (Dayhoff *et al.*, 1965) and policies to unify and share the data were established to benefit the whole scientific community (Brunak *et al.*, 2002). Since then, scientists have been well aware of the responsibility and the opportunities offered by the release of their published data in general reference databases (Cochrane *et al.*, 2016; Brunak *et al.*, 2002). Indeed, conveying comprehensive collections in a common computational platform establishes references for all scientists for a one-stop shop, independently from the main scientific interests. Specialized, secondary resources, on the other hand, aim to offer curated (UniProt Consortium, 2015; Kanehisa and Goto, 2000) and dedicated collections (Goodstein *et al.*, 2012; Dong *et al.*, 2004). Table 4.1 summarizes some of

Table 4.1 List of the major bioinformatics resources available for plant genomics. Description and website are specified. General databases include also collections from non-plant species

Database	Description	Website
General		
INSDC	Unified DDBJ, EMBL-EBI and NCBI repository	www.insdc.org/
UniProt	Database of functional annotated protein sequences	www.uniprot.org/
Protein Data Bank	3D structure of proteins and nucleic acids	www.rcsb.org/pdb/home/home.do
RFAM	RNA family collections	http://rfam.xfam.org/
Gene Ontology	Database of ontologies and gene annotations	http://geneontology.org/page/go-database
KEGG	Metabolic pathway database	www.kegg.jp/
EggNOG	Comparative genomics	http://eggnog.embl.de/version_4.0.beta/
InParanoid	Comparative genomics	http://inparanoid.sbc.su.se/cgi-bin/index.cgi
Plant specific		
Ensembl Plants	Plant genomics database	http://plants.ensembl.org/index.html
Phytozome	Plant genomics database	www.phytozome.net/
PlantGDB	Plant genomics database	www.plantgdb.org/
Plant Metabolic Network	Plant metabolic pathway database	www.plantcyc.org/
Plant Reactome	Plant metabolic pathway database	http://plantreactome.gramene.org/
GreenPhyl	Plant comparative genomics	www.greenphyl.org/cgi-bin/index.cgi
Plaza	Plant comparative genomics	http://bioinformatics.psb.ugent.be/plaza/

the major resources, general or specialized, offering comprehensive collections for plant genomics.

The International Nucleotide Sequence Database Collaboration (INSDC) holds raw sequence data, accommodating also experimental design details and NGS reads (Cochrane *et al.*, 2016). Ensembl Plants is part of the Ensembl genome project, which includes genome sequences, protein annotations, transcriptional data, genetic variation and comparative results from different taxa (Kersey *et al.*, 2016). It contains reference genome assemblies from 33 plant species (Bolser *et al.*, 2016). Phytozome is a comparative platform for plants. It provides access to sequences and functional annotations of complete plant genomes (currently 65), and a view of the evolutionary history of every plant gene (Goodstein *et al.*, 2012). PlantGDB is a reference genomic database for plants (Dong *et al.*, 2004), including 50 plant species. The Universal Protein Resource (UniProt) is a general reference platform of protein sequences and their annotation. UniProt is divided in two sections: the Reviewed (Swiss-Prot) manually annotated database, in which proteins and their information are manually curated based also on literature data, and the Unreviewed (TrEMBL) computationally analysed database, containing automatically annotated proteins from general nucleotide databases (UniProt Consortium, 2015). The Protein Data Bank (PDB) is a database of tridimensional structure data. This database stores X-ray crystal structures, nuclear magnetic resonance (NMR) structures, cryo-electron microscopy and theoretical modelling (Berman *et al.*, 2000). Rfam is a collection of non-coding RNA families depicted by manually curated sequence alignments, annotation and consensus of secondary structures (Nawrocki *et al.*, 2015). Gene Ontology Consortium (GOC) is a project dedicated to the definition of consistent descriptions of gene products, incorporating many collections from plant, animal and microbial genomes (Ashburner *et al.*, 2000). KEGG (Kyoto Encyclopaedia of Genes and Genomes) is a database for systematic analysis of biological systems (pathways included) (Kanehisa and Goto, 2000). Plant Metabolic Network (PMN) is a database of metabolic pathways for plants (Dreher, 2014). PMN currently groups one multispecies reference database called PlantCyc and 22 species/taxon-specific databases. Plant Reactome is a database of pathways and reactions from plants (Fabregat *et al.*, 2016). It supplies molecular details of DNA replication, metabolism, signal transduction, gene expression, diseases, transmembrane transport of small molecules and vesicle-mediated transport. EggNOG is a database of orthologous groups of proteins from different taxonomic levels with functional annotations from 20 plant species. Furthermore, it provides scaffolds for quickly mapping novel sequences to orthologous groups based on HMM profiles (Huerta-Cepas *et al.*, 2016). InParanoid is a collection of pairwise orthologue groups including 20 plant species (O'Brien *et al.*, 2005). GreenPhyl is a web-based resource for comparative and functional genomics in plants (Rouard *et al.*, 2011), containing a catalogue of gene families based on gene predictions, covering a broad taxonomy of green plants. Plaza is a public resource containing 37 plant species genomes (Proost *et al.*, 2015), with the aim to facilitate inspection on structural and functional annotations, gene families, protein domains, and detailed information about genome organization and phylogenetic trees. All these collections resemble the amount of resources and facilities today available for plant genomics. All of them are endowed with peculiarities and specificities that should be appropriately addressed by interested users for an appropriate exploitation.

NGS-based resources

Next-generation sequencing, though fast and cheaper, is computationally expensive. Indeed, it requires suitable and dedicated software and platforms able to manage, organize, analyse and disseminate huge quantities of short nucleotide reads, reaching commonly at least one gigabase per analytical run (Bateman and Quackenbush, 2009; Horner *et al.*, 2010; Tang and Zhao, 2015; Yang *et al.*, 2009). Data management and -omics data integration has always represented a challenge for bioinformatics (Benson *et al.*, 2000; Bita *et al.*, 2011; Chiusano *et al.*, 2008; Dong *et al.*, 2004; Edgar *et al.*, 2002; Flicek *et al.*, 2013; Kodama *et al.*, 2012; Leinonen *et al.*, 2011; O'Leary *et al.*, 2015), and the spreading of NGS data made the challenge even harder, increasing the need for suitable storage, methods for the processing and mining, and platforms for immediate access to results from novel massive data generated by these technologies (Magi *et al.*, 2010; Wang *et al.*, 2015).

The Sequence Read Archive (SRA, available at www.ncbi.nlm.nih.gov/sra) (Kodama *et al.*, 2012), for example, is a public resource established by the NCBI with the aim to gather raw collections from NGS efforts, including data from Roche 454 GS System (Droege and Hill, 2008), Illumina Genome Analyzer (Bennett, 2004), Applied Biosystems SOLiD System (Porreca *et al.*, 2006) and Helicos Heliscope (Harris *et al.*, 2008). Another resource at the NCBI, GEO (Gene Expression Omnibus) (Edgar *et al.*, 2002), which was initially established to favour dissemination of microarray results, today also provides results from NGS-based collections, such as gene expression or methylation profiles. The 'plant' keyword was used in order to identify NGS data from plants from both resources (Table 4.2), resulting in 131,207 matches in SRA and 1829 matches in GEO. The table shows that the results in GEO are far distant from representing the whole NGS collections publicly available in SRA. This highlights the gap in general resources to access results from NGS approaches. Indeed, the inestimable source of information that transcriptomics, epigenomics and metagenomics (Esposito *et al.*, 2016) may provide to elucidate genome organization and functionalities still needs appropriate platforms to be made available to the scientific community, though results from dedicated efforts are somewhat made available**.** Table 4.3 lists general NGS resources and dedicated platforms that include plant RNA-seq collections.

Comparative genomics resources

A great opportunity for genomics is represented by comparative efforts. They drive investigations on differences and similarities among species, therefore contributing to the deciphering of the evolutionary forces that shaped genomic features, beyond supporting the transfer of information from model organisms to newly sequenced genomes. The detection of orthologue genes among different species is a key approach for comparative analyses (Altenhoff and Dessimoz, 2009, 2012; Altenhoff *et al.*, 2011; Ambrosino and Chiusano, 2013; Dessimoz *et al.*, 2012; Kristensen *et al.*, 2011; Trachana *et al.*, 2014). Collections of orthologues are organized in widely used comparative platforms that today include numerous plant species (Table 4.4).

It is evident from the overview provided here the high number of similar resources, including different collections and based on different approaches to define the respective results.

Table 4.2 Number of query matches (hits) per species searching for 'plant' keyword in NCBI's SRA (www.ncbi.nlm.nih.gov/sra) and GEO (www.ncbi.nlm.nih.gov/gds). In the case of GEO, a filter for high-throughput sequencing was applied. In both cases, hits associated to non-plant species were omitted

Database	Species	Number of hits
SRA	*Arabidopsis thaliana*	9824
	Soil metagenome	7918
	Oryza sativa	3561
	Manihot esculenta	3047
	Zea mays	2448
	Triticum aestivum	2334
	Plant metagenome	2280
	Hordeum vulgare	2138
	Solanum lycopersicum	2025
	Brassica napus	1931
	Glycine max	1826
	Root metagenome	1697
	Solanum tuberosum	1650
	Populus trichocarpa	1619
	Erythranthe guttata	1549
	Boechera stricta	1529
	Miscanthus sinensis	1480
	All other taxa	58,234
GEO	*Arabidopsis thaliana*	508
	Zea mays	98
	Oryza sativa	88
	Glycine max	68
	Solanum lycopersicum	41
	Chlamydomonas reinhardtii	35
	Vitis vinifera	31
	Gossypium hirsutum	25
	Brassica rapa	22
	Arabidopsis lyrata	21
	Triticum aestivum	18
	Brassica napus	17
	Medicago truncatula	17
	Manihot esculenta	15
	Capsicum annuum	14
	Physcomitrella patens	14

Table 4.3 Summary of general and species-specific RNA-seq resources. Description of contents and links are also included

Database	Description	Website
General		
SRA (Kodama *et al*., 2012)	Sequence Read Archive; includes raw data reads	www.ncbi.nlm.nih.gov/sra
ENA (Leinonen *et al*., 2011)	European Nucleotide Archive; includes raw data, assembly and functional annotation	www.ebi.ac.uk/ena
GEO (Edgar *et al*., 2002)	Gene Expression Omnibus; includes microarray and NGS data results	www.ncbi.nlm.nih.gov/gds
Expression Atlas (Petryszak *et al*., 2016)	Provides information about gene expression patterns; includes both microarray and RNA-seq data	www.ebi.ac.uk/gxa/home
Next-Gen Sequence Databases (Nakano *et al*., 2006)	NGS databases of 19 plant species; selection of small-RNA, RNA-seq, MethylC-seq and Chip-seq for plant species are made available	https://mpss.danforthcenter.org/index.php
MedPlant – RNAseq Database	Sequence Read Archive; includes raw data reads	www.medplantrnaseq.org/
Species specific		
AGED (O'Rourke *et al*., 2015)	Alfalfa expression atlas database; RNA-seq of two subspecies, *Medicago sativa* ssp. *sativa* and *Medicago sativa* ssp. *falcate*. Query per gene expression and per differentially expressed genes are allowed	http://plantgrn.noble.org/AGED/index.jsp
Genome Database for Rosaceae (Jung *et al*., 2013)	Database dedicated to Rosaceae family; provides links to SRA	www.rosaceae.org/
morexGenes	Barley gene expression levels database; includes results from RNA-seq from different tissues and developmental stages includes also microarray data	https://ics.hutton.ac.uk/morexGenes/
MOROKOSHI (Makita *et al*., 2015)	Sorghum transcriptome database; includes results from different tissues and developmental stages based on 26 RNA-seq samples. Links to the raw data used are also provided	http://sorghum.riken.jp/morokoshi/Home.html
NexGenEx-Tom (Bostan and Chiusano, 2015)	Tomato gene expression atlas; includes RNA-seq results from different tissues and developmental stages. The platform offers expression matrix, profiles and correlations and reads mapping onto the tomato genome	http://cab.unina.it/NexGenEx-Tom
PvGEA (O'Rourke *et al*., 2014)	Bean database; includes gene expression profiles in different tissues and developmental stages based on 24 RNA-seq samples. Links to the raw data used are also provided	http://plantgrn.noble.org/PvGEA/
Rice Gene Expression (Kawahara *et al*., 2013)	Rice database of gene expression profiles based on RNA-seq downloaded from SRA. Links to the raw data used are also provided	http://rice.plantbiology.msu.edu/expression.shtml
SGN (Fernandez-Pozo *et al*., 2015)	Solanaceae Genome Network; includes raw RNA-seq data for tomato and mapping of reads on the tomato genome	https://solgenomics.net/
SoyBase (Severin *et al*., 2010)	Soybean genomic database; includes RNA-seq from 14 tissues. The database offers the opportunity to query for differentially expressed genes between two tissues and to search for tissue-specific gene expression. Links to the raw data used are also provided	http://soybase.org/soyseq/
SpinachBase (Xu *et al*., 2015)	Spinach genomic database; provides links to RNA-seq data	www.spinachbase.org/cgi-bin/spinach/index.cgi
SpudDB (Hirsch *et al*., 2014)	Potato genomic database; includes raw RNA-seq data and mapping of reads on the potato genome	http://solanaceae.plantbiology.msu.edu/

Table 4.3 Continued

Database	Description	Website
TENOR (Kawahara *et al*., 2016)	Rice database of expression profiles; information of cis-regulatory elements in promoter regions and co-expressed transcript based on RNA-seq data from 140 environmental stress experiments and plant hormone treatments. Links to the raw data used are also provided	http://tenor.dna.affrc.go.jp/
TomExpress	Tomato database of gene expression profiles per different tissues and developmental stages, based on RNA-seq data downloaded from ENA and SRA	http://gbf.toulouse.inra.fr/tomexpress/www/welcomeTomExpress.php
TRAVA (Klepikova *et al*., 2015)	*Arabidopsis thaliana* database of gene expression profiles from different tissues and developmental stages based on 79 RNA-seq samples	http://travadb.org/
Vespucci (Moretto *et al*., 2015)	Grapevine expression compendium obtained by publicly available transcriptome experiments from RNA-seq and microarray data	http://vespucci.colombos.fmach.it/
WheatExp (Pearce *et al*., 2015)	Wheat database of gene expression profiles per different tissues and developmental stages, based on RNA-seq data downloaded from ENA, SRA and GEO. Links to the raw data used are also provided	http://wheat.pw.usda.gov/WheatExp/

Bottlenecks and challenges

Plant sciences are typically characterized by high heterogeneity, multiple different species, an incredible amount of crops and variants, and distinct and widespread scientific communities. Therefore, the easy accessibility to sequencing technologies drove far beyond the sequencing of reference species (Goff *et al.*, 2002; Jaillon *et al.*, 2007; Schnable *et al.*, 2009; The Arabidopsis Genome Initiative, 2000; The Tomato Genome Consortium, 2012; Xu *et al.*, 2011) and paved the way to the production of multiple genomes from variants, wild species and community-specific collections (Aflitos *et al.*, 2014; Aversano *et al.*, 2015; Ercolano *et al.*, 2014; Lam *et al.*, 2010; Weigel and Mott, 2009), giving rise to multifaceted genomics data sources.

Similar efforts are even more widespread when considering transcriptomes or other -omics approaches: physiological, stress or pathological conditions for all the possible variants enrich molecular databases of heterogeneous collections. However, these collections often need to be mapped on the genome sequences, and therefore they may suffer the drawback of not being appropriately exploitable by a reference genome sequence representing the genome of a distinct genotype. Moreover, the collections can derive from limited experimental design for contributing as suitable data sources for gene expression atlases and/or for gene co-expression analyses (Bostan and Chiusano, 2015; Di Salle *et al.*, 2016; Schmid *et al.*, 2005). These scientific trends, therefore, give rise to overwhelming data that need selection and reconciliation to contribute as consistent source of information in integrative analyses that could support structure and functional genomics.

The advent of revolutionary experimental technologies and novel computational approaches are evidently not accompanied by a comparable progress in genome characterizations of plant species. Indeed, although the attitude of the whole scientific community is being consistently affected by the interest for solving primary structures of genomes of different species, genotypes or cultivars (Aflitos *et al.*, 2014; Lam *et al.*, 2010; Weigel and Mott, 2009), generally driven by international consortiums and pushed by fast and low-cost technologies, only 10% of the genomes have been today confidently deciphered. This highlights that, despite 70 years having passed from the discovery of the DNA structure, the genomics era is still in its early stage, and extensive bioinformatics is still required in order to exploit molecular models of complex biological organisms (Esposito *et al.*, 2016).

Indeed, genome sequencing efforts handed to the scientific community an increasing number of newly sequenced plant genomes. However, several of them are still in the form of drafts with a still

Table 4.4 Summary of the major comparative genomics platform available for plants. The number of plant species out of total species and the methods for the detection of orthologues are included

Orthologues database	Plant species/total	Methods
EggNog (Huerta-Cepas *et al*., 2016)	20/2031	Orthologous groups inferred by the SIMAP (Similarity MAtrix of Proteins) approach and processed by the EggNOG orthology prediction pipeline
		Phylogenetic reconstruction for all groups was performed using the ETE toolkit
Ensembl Plants (Bolser *et al*., 2016)	44/44	Comparative genomics based on protein sequences providing gene trees and orthology information
		Whole genome alignments between selected genomes (based on LastZ and translated BLAT)
		Syntenies calculated from genome or peptide alignments
		Gene families constructed from classification of proteins
GreenPhyl (Rouard *et al*., 2011)	37/37	Clustering performed on the protein-coding gene using TribesMCL
		Phylogenetic analyses and ortholog inference based on MAFFT, PhyML and RAPGreen v54
InParanoid (O'Brien *et al*., 2005)	20/273	Homology detected by BLAST program
		Phylogenetic tree generated by UPGMA clustering of pairwise species distances
OrthoMCL database (Chen *et al*., 2006)	11/150	All-versus-all BLASTp of the protein sequences
		Putative inparalog, ortholog and co-ortholog pairs inferred using the OrthoMCL Pairs program
		MCL program to cluster the protein sequences pairs into groups
Plaza (Proost *et al*., 2015)	64/64	Orthologous gene families (ORTHO) inferred using OrthoMCL
		Tree-based orthologs (TROG) inferred using tree reconciliation of the phylogenetic tree of a gene family
		Best-Hits-and-Inparalogs (BHI) inferred from Blast hits against the PLAZA protein database
		Anchor points refer to gene-based colinearity between species

preliminary gene annotation. On the other hand, others among all may evolve faster, this depending on the quality of the first release and on the available opportunities in terms of support for dedicated efforts. Both aspects, however, affect the establishment of reliable reference resources that could drive associated -omics efforts (e.g. RNA-seq, proteomics, epigenomics) and their appropriate exploitation by non-expert users.

Main reasons for poorly annotated draft genomes are related to bottlenecks from bioinformatics and human curation, when compared to the faster data production. In particular, correct assembling of large amount of complex, redundant sequence data and sufficient in-depth studies on multiple heterogeneous collections require time. This slows down the reasonable deciphering of the major information content that could enable the understanding of the intricate aspects of genome functionality. Moreover, publication of novel genome sequences is often more attractive than the care for updates of already published information. Funding agencies may also consider more appealing the sequencing of a novel genome than the expansion and the integration of information contributing to the in-depth analyses of already sequenced ones, especially when they do not represent a reference international species. As a consequence, preliminary drafts often risk to remain in the preliminary status. On the counterpart, data production is also moving too fast, affecting the updating rate and the definition of stable reference resources that could support the whole interested community (Chiusano, 2015). Table 4.5 reports the genome annotation versions for different species of agronomic interest, as they are available in some of the most popular public resources.

With the exception of *Arabidopsis* and rice, whose genomes were sequenced more than 10 years ago (Goff *et al.*, 2002; The *Arabidopsis* Genome Initiative, 2000), there is an evident non-uniformity among the data included in public resources. This limit are related not only to recently sequenced genomes, such as tomato (The Tomato Genome Consortium, 2012) and potato (Xu *et al.*, 2011), but also to older genomes releases, such as the ones of maize (Schnable *et al.*, 2009) and grapevine (Jaillon *et al.*, 2007) (Table 4.5). This may be due to slower update rates, but also to the fast release of novel versions. As an example, the tuber crop potato, published in 2011 and initially released in the form of scaffolds, has been already endowed of six different genome annotation versions related to five different genome assemblies (Table 4.6), each with a level of heterogeneity further highlighted when considering differences in the number of predicted genes and alternative transcripts from each version (Table 4.6).

In addition, the dissemination of genome sequencing results is strongly affected also by other aspects. As an example, sequenced plant genomes, such as coffee (Denoeud *et al.*, 2014), available only in a dedicated resource (http://coffee-genome.org/), tobacco (Sierro *et al.*, 2014) or other genome releases (Ercolano *et al.*, 2014; The Tomato Genome Consortium, 2012; Velasco *et al.*, 2007) are not present in any of the reference databases here considered, NCBI and Ensembl Plants included. Undoubtedly, these limits and ambiguities do not support scientists, expert or non-expert users as well. The set-up of reference comprehensive public resources, supporting user friendly investigations for the whole scientific community, represents, indeed, the key strategy to make -omics really profitable. The recent story of molecular biology underlines that reference collections and data sharing thanks to bioinformatics resources have been fundamental, representing the backbone that moulded the progress in -omics sciences we are currently living. However, the flourishing of an increasing number of public databases with heterogeneous collections, providing different genome versions associated to varied independent annotations, as well as the lack of dissemination of published data, affect the establishment of well-accepted references and undoubtedly limit scientific applications.

Different gene annotations of the same genome slow down the work of the end user, often forced to carefully investigate among the available resources to assess the reliability of a collection, this also requiring appropriate expertise. Moreover, this may compromise the reproducibility of results, since they are affected by the materials and the methodologies employed (Di Salle *et al.*, 2016). As an example, well-known platforms for comparative genomics are based on different gene annotation versions from the same plant species (Fig. 4.2) and, as a consequence, on different results, compromising comparability among the different methodologies employed (as reported in Table 4.4) and reliability of analyses pending from the different efforts. Moreover, the heterogeneity of data, the lack of coordination to fix methodologies and/or collections, as well as the not clearly declared obsolescence of published resources, result in diversified information that limit non-expert users and do not facilitate uniform and reproducible scientific approaches. As an example, the presence of too many resources available in support of gene expression analyses for the reference species *A. thaliana* misleads the users also with discordant results (as reviewed in Di Salle *et al.*, 2016). Moreover, the establishment of platforms including species-specific selected collections of expression data is spreading for several plant species (Table 4.3), but these independent efforts, not always accompanied by evident coordination and international support, will not produce, presumably, stable, long term impact in plant sciences.

One last dramatic trend to consider, moreover, is the poor crosslinking of plant genome results with general databases worldwide recognized as well established resources. Dedicated annotations from independent consortium are precious, since they benefit from specific knowledge that typically comes from the scientific community supporting the consortium. However, many of the consortia results are not cross-linked with the general efforts undertaken by reference databases. As an example,

Table 4.5 List of plant species of relevant agronomical interest, sorted by the year of publication of their genomes. The annotation version available for these species in different reference databases is also reported

Species	Year of publication	Annotation version						
		NCBI	Ensembl Plants	PlantGDB	Phytozome	Plaza	Eggnog	KEGG
Arabidopsis	2000	TAIR 10	TAIR 10	TAIR 10	TAIR 10	TAIR 10	TAIR 10	TAIR 10
Rice	2002	RGAP 7	RGAP 7	RGAP 7	RGAP 7	RGAP 7	RGAP 7	RGAP 7
Grapevine	2007	GCF_000003745.3	V1 Cribi	V2 Genoscope	V2 Genoscope	V2 Genoscope	V1 Cribi	GCF_000003745.3
Maize	2009	B73_RefGen_v3	B73_RefGen_v4	B73_RefGen_v2	B73_RefGen_v3	B73_RefGen_v4	B73_RefGen_v4	B73_RefGen_v3
Potato	2011	GCF_000226075.1	GCF_000226075.1	PGSC v.3 2.1.10	PGSC v. 3.4	iTAG v. 1	GCF_000226075.1	GCF_000226075.1
Tomato	2012	GCF_000188115.3	iTAG v. 2.4	NA	iTAG v. 2.3	iTAG v. 2.3	iTAG v. 2.3	GCF_000188115.3
Sweet orange	2013	GCF_000317415.1	NA	NA	JGI v1	JGI v1	NA	GCF_000317415.1
Pepper	2014	GCF_000710875.1	NA	NA	NA	NA	NA	NA
Amborella	2013	GCF_000471905.1	GCF_000471905.1	NA	AmTr_v_0.10	AmTr_v_0.10	NA	GCF_000471905.1
Soybean	2006	GCF_000004515.4	GCF_000004515.1	Wm82.a2.v1	Wm82.a2.v1	Wm82.a2.v1	GCF_000004515.1	GCF_000004515.4
Apple	2010	GCF_000148765.1	NA	NA	M. Domestica v1.0	M. Domestica v1.0	NA	GCF_000148765.1
Sorghum	2009	GCF_000003195.2	GCF_000003195.2	Sbi1.4	Sbi3.1	NA	GCF_000003195.2	GCF_000003195.2
Barrel medic	2011	GCF_000219495.2	Mt4.0	Mt3.5	Mt4.0	Mt4.0	NA	GCF_000219495.2
Banana	2012	GCF_000313855.1	GCF_000313855.1	NA	M. acuminata v 1	NA	GCF_000313855.1	GCF_000313855.1
Cocoa	2011	GCF_000208745.1	GCF_000208745.1	NA	C. Matina v1.1	C. Matina v1.1	NA	GCF_000208745.1

Table 4.6 List of different genome assemblies and annotation versions available for potato. Number of representative genes and alternative transcripts for each annotation version are also reported

Genome assembly version	Annotation version	Year of release	Number of representative genes	Number of alternative transcripts
v3 superscaffold	iTAG v.1	2011	35,004	NA
	PGSC 3.4	2011	39,031	56,218
	GCF_000226075.1	2011	33,608	37,885
v3 2.1.10 pseudomolecules	PGSC v. 3 2.1.10	2012	NA	52,228
v3 2.1.11 pseudomolecules	PGSC v. 3 2.1.11	2012	NA	52,228
v4.03 pseudomolecules	PGSC v. 4.03	2013	39,146	56,980
v4.04 pseudomolecules	NA	2016	NA	NA

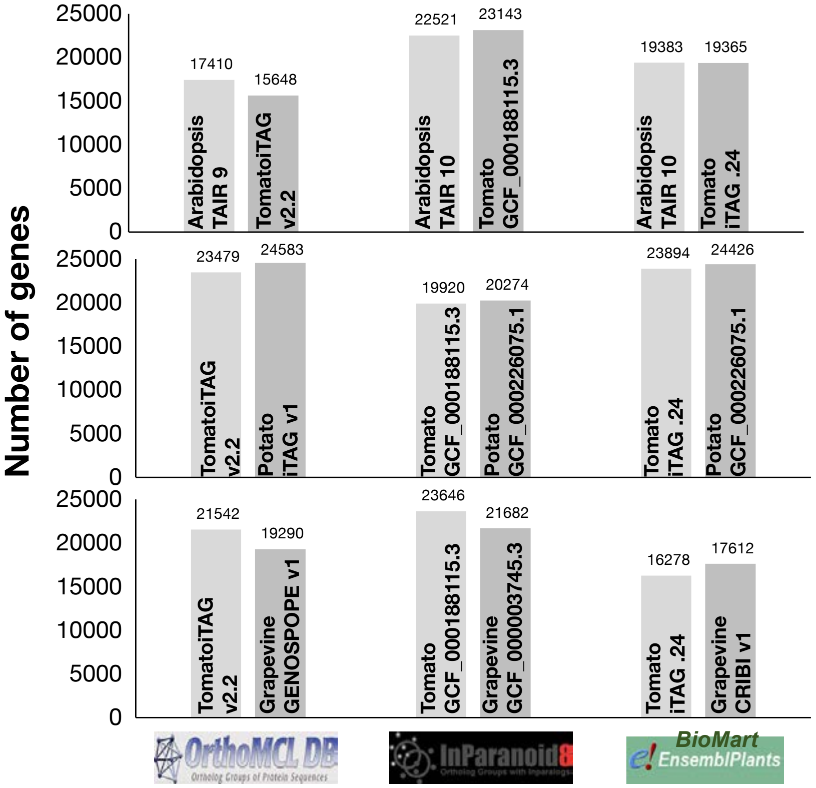

Figure 4.2 Pairs of orthologue genes stored in three widespread comparative genomics platforms. For each species (*Arabidopsis*, tomato, potato and grapevine), the annotation version and the number of genes that have an orthologue relationship with the compared counterpart is shown.

the RefSeq annotation (O'Leary *et al.*, 2015) comes out from a unified gene annotation pipeline for all public sequenced genomes provided by the NCBI. Refseq annotations are cross-linked to relevant resources such as UniProt (UniProt Consortium, 2015) and KEGG (Kanehisa and Goto, 2000). Very few genome platforms for plant species also include the RefSeq annotations. On the other hand, RefSeq analyses does not appear to integrate the community related gene annotation for a species. This is quite common in plant genomics and determines two main limits: the presence of added-value information in specific websites not shared with general reference database, and limited information access to non-plant users, that usually refer to general databases. This is evidently against the basic concept of sharing and global exchange of information established by the International Nucleotide Sequence Database Collaboration (INSDC) in 2002 (Brunak *et al.*, 2002) and definitely remarked in 2015 (Cochrane *et al.*, 2016). On the other hand, this lack of direct links with well-established reference collections makes plant users isolated from the general trends in bioinformatics, which are fast expanding for reference animal species. Furthermore, because of the multifaceted efforts in plant genomics, driven by different communities focused on specific species and usually growing in specific and specialized context, crosslinks are usually hard and difficult to be exploited. While interconnections and comparative efforts would surely support profitable science.

However, although these limitations, -omics in the NGS era together with bioinformatics innovations largely moulded the experimental design in plant molecular biology, consistently contributing to the scientific knowledge in plants molecular biology and positively affecting many applications of agriculture sciences (Esposito *et al.*, 2016). Diverse plants research fields, such as breeding, environmental sciences and microbiology, are benefiting from the available knowledge and are contributing data, favouring scientific advances, improving sustainability, products quality and strategies for stress or disease treatments (Deusch *et al.*, 2015; Tringe and Coleman-Derr, 2014).

Undoubtedly, NGS approaches are driving more efficient solutions from bioinformatics, continuously stimulating novel computational approaches. The need for innovative interfaces and even more intuitive tools to address the main biological questions emphasized by the novel technologies will not decline for many years to come (Horner *et al.*, 2010).

The most critical issues that bioinformatics should face to meet the major trends in plant biology and overcome some of the bottlenecks here highlighted are mainly related to a proper dissemination and stabilization of resources, favouring integration of information from different levels of the cell functionality and from different species. To this aim, the development of comprehensive collections and the favouring of in-depth knowledge through education and training in omics-based technologies and in bioinformatics, opportunely integrating experimental and computational efforts, and possibly different scientific communities, are essential (Chiusano, 2015; Esposito *et al.*, 2016).

Beyond all, coordinated efforts preserving data curation should be consistently supported, with the main purpose of avoiding that profitable -omics could be restricted to major experts and limited to major species.

The need for 'tailoring' bioinformatics to dedicated efforts that could improve the quality of draft results and make them appropriately accessible, usable and reconciled with related resources, is fundamental to reducing the risk that fast data production could compromise the sharing of reliable and updated information. Project reviewers and publication policies may also play a relevant role in facing the presented issues, since they can drive towards appropriate cross-referencing, data sharing and high quality standards in both software and resources.

The care for appropriate knowledge dissemination should remain the priority in Science. Enhanced bioinformatics is the strategy that can fulfil the scope.

Acknowledgements

This work was supported by the Genopom PRO and the GenHORT Projects [Ministero dell'Istruzione, dell'Università e della Ricerca (MIUR), Italy] and in the frame of the COST ACTION (FA1106).

References

Adams, M.D., Kelley, J.M., Gocayne, J.D., Dubnick, M., Polymeropoulos, M.H., Xiao, H., Merril, C.R., Wu, A., Olde, B., Moreno, R.F., *et al.* (1991). Complementary DNA sequencing: expressed sequence tags and human genome project. Science *252*, 1651–1656. http://dx.doi.org/10.1126/science.2047873

Aflitos, S., Schijlen, E., de Jong, H., de Ridder, D., Smit, S., Finkers, R., Wang, J., Zhang, G., Li, N., Mao, L., *et al.* (2014). Exploring genetic variation in the tomato (Solanum section Lycopersicon) clade by whole-genome sequencing. Plant J. *80*, 136–148. http://dx.doi.org/10.1111/tpj.12616

Altenhoff, A.M., and Dessimoz, C. (2009). Phylogenetic and functional assessment of orthologs inference projects and methods. PLoS Comput Biol 5, e1000262. http://dx.doi.org/10.1371/journal.pcbi.1000262

Altenhoff, A.M., and Dessimoz, C. (2012). Inferring orthology and paralogy. Methods Mol. Biol. *855*, 259–279. http://dx.doi.org/10.1007/978-1-61779-582-4_9

Altenhoff, A.M., Schneider, A., Gonnet, G.H., and Dessimoz, C. (2011). OMA 2011: orthology inference among 1000 complete genomes. Nucleic Acids Res. *39*, D289–294. http://dx.doi.org/10.1093/nar/gkq1238

Ambrosino, L., and Chiusano, M.L. (2013). In quest of orthologs. In BBCC 2013 (Avellino).

Andorf, C.M., Cannon, E.K., Portwood, J.L., 2nd, Gardiner, J.M., Harper, L.C., Schaeffer, M.L., Braun, B.L., Campbell, D.A., Vinnakota, A.G., Sribalusu, V.V., *et al.* (2016). MaizeGDB update: new tools, data and interface for the maize model organism database. Nucleic Acids Res. *44*, D1195–1201. http://dx.doi.org/10.1093/nar/gkv1007

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., *et al.* (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. *25*, 25–29. http://dx.doi.org/10.1038/75556

Aversano, R., Contaldi, F., and Ercolano, M.R. (2015). The *Solanum commersonii* genome sequence provides insights into adaptation to stress conditions and genome evolution of wild potato relatives. Plant Cell *27*, 954–968. http://dx.doi.org/10.1105/tpc.114.135954

Bateman, A., and Quackenbush, J. (2009). Bioinformatics for next generation sequencing. Bioinformatics *25*, 429. http://dx.doi.org/10.1093/bioinformatics/btp037

Becker, C., Hagmann, J., Muller, J., Koenig, D., Stegle, O., Borgwardt, K., and Weigel, D. (2011). Spontaneous epigenetic variation in the *Arabidopsis thaliana* methylome. Nature *480*, 245–249. http://dx.doi.org/10.1038/nature10555

Becker, J.D., Boavida, L.C., Carneiro, J., Haury, M., and Feijó, J.A. (2003). Transcriptional profiling of *Arabidopsis* tissues reveals the unique characteristics of the pollen transcriptome. Plant Physiol. *133*, 713-725. http://dx.doi.org/10.1104/pp.103.028241

Bennett, S. (2004). Solexa Ltd. Pharmacogenomics *5*, 433–438. http://dx.doi.org/10.1517/14622416.5.4.433

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Rapp, B.A., and Wheeler, D.L. (2000). GenBank. Nucleic Acids Res. *28*, 15–18.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). The Protein Data Bank. Nucleic Acids Res. *28*, 235–242. http://dx.doi.org/10.1093/nar/28.1.235

Betryszak, R., Keays, M., Tang, Y.A., Fonseca, N.A., Barrera, E., Burdett, T., Füllgrabe, A., Fuentes, A.M.-P., Jupp, S., Koskinen, S., *et al.* (2016). Expression Atlas update—an integrated database of gene and protein expression in humans, animals and plants. Nucleic Acids Res. *44*, D746–D752.

Bita, C.E., Zenoni, S., Vriezen, W.H., Mariani, C., Pezzotti, M., and Gerats, T. (2011). Temperature stress differentially modulates transcription in meiotic anthers of heat-tolerant and heat-sensitive tomato plants. BMC Genomics *12*, 384. http://dx.doi.org/10.1186/1471-2164-12-384

Blair, M.W., Fernandez, A.C., Ishitani, M., Moreta, D., Seki, M., Ayling, S., and Shinozaki, K. (2011). Construction and EST sequencing of full-length, drought stress cDNA libraries for common beans (*Phaseolus vulgaris* L.). BMC Plant Biol. *11*, 171. http://dx.doi.org/10.1186/1471-2229-11-171

Blanc, G., Hokamp, K., and Wolfe, K.H. (2003). A recent polyploidy superimposed on older large-scale duplications in the *Arabidopsis* genome. Genome Res. *13*, 137–144. http://dx.doi.org/10.1101/gr.751803

Blanchfield, J.R. (2004). Genetically modified food crops and their contribution to human nutrition and food quality. J. Food Sci.*69*, CRH28-CRH30. http://dx.doi.org/10.1111/j.1365-2621.2004.tb17846.x

Bokszczanin, K., Krezdorn, N., Fragkostefanakis, S., Muller, S., Rycak, L., Chen, Y., Hoffmeier, K., Kreutz, J., Paupiere, M., Chaturvedi, P., *et al.* (2015). Identification of novel small ncRNAs in pollen of tomato. BMC Genomics *16*, 714. http://dx.doi.org/10.1186/s12864-015-1901-x

Bolser, D., Staines, D.M., Pritchard, E., and Kersey, P. (2016). Ensembl plants: integrating tools for visualizing, mining, and analyzing plant genomics data. Methods Mol. Biol. *1374*, 115–140. http://dx.doi.org/10.1007/978-1-4939-3167-5_6

Bostan, H., and Chiusano, M.L. (2015). NexGenEx-Tom: a gene expression platform to investigate the functionalities of the tomato genome. BMC Plant Biol. *15*, 48. http://dx.doi.org/10.1186/s12870-014-0412-2

Bowers, J.E., Chapman, B.A., Rong, J., and Paterson, A.H. (2003). Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. Nature *422*, 433-438. http://dx.doi.org/10.1038/nature01521

Brenchley, R., Spannagl, M., Pfeifer, M., Barker, G.L., D'Amore, R., Allen, A.M., McKenzie, N., Kramer, M., Kerhornou, A., Bolser, D., *et al.* (2012). Analysis of the bread wheat genome using whole-genome shotgun sequencing. Nature *491*, 705–710. http://dx.doi.org/10.1038/nature11650

Brenner, S., Johnson, M., Bridgham, J., Golda, G., Lloyd, D.H., Johnson, D., Luo, S., McCurdy, S., Foy, M., Ewan, M., *et al.* (2000). Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. Nat. Biotechnol. *18*, 630–634. http://dx.doi.org/10.1038/76469

Brunak, S., Danchin, A., Hattori, M., Nakamura, H., Shinozaki, K., Matise, T., and Preuss, D. (2002). Nucleotide Sequence Database Policies. Science *298*, 1333.

Bülow, L., Schindler, M., and Hehl, R. (2007). PathoPlant®: a platform for microarray expression data to analyze co-regulated genes involved in plant defense responses. Nucleic Acids Res. *35*, D841–D845. http://dx.doi.org/10.1093/nar/gkl835

Chen, F., Mackey, A.J., Stoeckert, C.J., Jr., and Roos, D.S. (2006). OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. Nucleic Acids Res. *34*, D363–368.

Chiusano, M.L. (2015). On the multifaceted aspects of bioinformatics in the next generation era: the run that must keep the quality. Next Generat Sequenc & Applic. e106. http://dx.doi.org/10.4172/2469-9853.1000e106

Chiusano, M.L., D'Agostino, N., Traini, A., Licciardello, C., Raimondo, E., Aversano, M., Frusciante, L., and Monti, L. (2008). ISOL@: an Italian SOLAnaceae genomics resource. BMC Bioinformatics *9 Suppl 2*, S7. http://dx.doi.org/10.1186/1471-2105-9-s2-s7

Choulet, F., Wicker, T., Rustenholz, C., Paux, E., Salse, J., Leroy, P., Schlub, S., Le Paslier, M.C., Magdelenat, G., Gonthier, C., *et al.* (2010). Megabase level sequencing reveals contrasted organization and evolution patterns of the wheat gene and transposable element spaces. Plant Cell *22*, 1686–1701. http://dx.doi.org/10.1105/tpc.110.074187

Cochrane, G., Karsch-Mizrachi, I., and Takagi, T. (2016). The International Nucleotide Sequence Database Collaboration. Nucleic Acids Res. *44*, D48-50. http://dx.doi.org/10.1093/nar/gkv1323

Cui, L., Wall, P.K., Leebens-Mack, J.H., Lindsay, B.G., Soltis, D.E., Doyle, J.J., Soltis, P.S., Carlson, J.E., Arumuganathan, K., Barakat, A., *et al.* (2006). Widespread genome duplications throughout the history of flowering plants. Genome Res. *16*, 738–749. http://dx.doi.org/10.1101/gr.4825606

Dayhoff, M., Eck, R., Chang, M., and Sochard, M. (1965). Atlas of Protein Sequence and Structure, Vol. 1 (MD: Silver Spring).

De Luca, V., Salim, V., Atsumi, S.M., and Yu, F. (2012). Mining the biodiversity of plants: a revolution in the making. Science *336*, 1658–1661. http://dx.doi.org/10.1126/science.1217410

Denoeud, F., Carretero-Paulet, L., Dereeper, A., Droc, G., Guyot, R., Pietrella, M., Zheng, C., Alberti, A., Anthony, F., Aprea, G., *et al.* (2014). The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. Science *345*, 1181–1184. http://dx.doi.org/10.1126/science.1255274

Dessimoz, C., Gabaldon, T., Roos, D.S., Sonnhammer, E.L., and Herrero, J. (2012). Toward community standards in the quest for orthologs. Bioinformatics *28*, 900–904. http://dx.doi.org/10.1093/bioinformatics/bts050

Deusch, S., Tilocca, B., Camarinha-Silva, A., and Seifert, J. (2015). News in livestock research – use of Omics-technologies to study the microbiota in the gastrointestinal tract of farm animals. Comput Struct Biotechnol J *13*, 55–63. http://dx.doi.org/10.1016/j.csbj.2014.12.005

Di Salle, P., Incerti, G., Colantuono, C., and Chiusano, M.L. (2016). Gene co-expression analyses: an overview from microarray collections in *Arabidopsis thaliana*. Brief Bioinform. http://dx.doi.org/10.1093/bib/bbw002

Dong, Q., Schlueter, S.D., and Brendel, V. (2004). PlantGDB, plant genome database and analysis tools. Nucleic Acids Res. *32*, D354-359. http://dx.doi.org/10.1093/nar/gkh046

Dreher, K. (2014). Putting the Plant Metabolic Network Pathway Databases to Work: Going Offline to Gain New Capabilities. In Plant Metabolism: Methods and Protocols, G. Sriram, ed. (Totowa, NJ: Humana Press), pp. 151-171. http://dx.doi.org/10.1007/978-1-62703-661-0_10

Droege, M., and Hill, B. (2008). The Genome Sequencer FLX™ System – Longer reads, more applications, straightforward bioinformatics and more complete data sets. J. Biotechnol. *136*, 3-10. http://dx.doi.org/http://dx.doi.org/10.1016/j.jbiotec.2008.03.021

Edgar, R., Domrachev, M., and Lash, A.E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. *30*, 207-210. http://www.ncbi.nlm.nih.gov/geo

Ercolano, M.R., Sacco, A., Ferriello, F., D'Alessandro, R., Tononi, P., Traini, A., Barone, A., Zago, E., Chiusano, M.L., Buson, G., *et al.* (2014). Patchwork sequencing of tomato San Marzano and Vesuviano varieties highlights genome-wide variations. BMC Genomics *15*, 138. http://dx.doi.org/10.1186/1471-2164-15-138

Esposito, A., Colantuono, C., Ruggieri, V., and Chiusano, M.L. (2016). Bioinformatics for agriculture in the Next-Generation sequencing era. Chemical and Biological Technologies in Agriculture *3*, 1-12. http://dx.doi.org/10.1186/s40538-016-0054-8

Fabregat, A., Sidiropoulos, K., Garapati, P., Gillespie, M., Hausmann, K., Haw, R., Jassal, B., Jupe, S., Korninger, F., McKay, S., *et al.* (2016). The Reactome pathway Knowledgebase. Nucleic Acids Res. *44*, D481–487. http://dx.doi.org/10.1093/nar/gkv1351

Fernandez-Pozo, N., Menda, N., Edwards, J.D., Saha, S., Tecle, I.Y., Strickler, S.R., Bombarely, A., Fisher-York, T., Pujar, A., Foerster, H., *et al.* (2015). The Sol Genomics Network (SGN) – from genotype to phenotype to breeding. Nucleic Acids Res. *43*, D1036–1041. http://dx.doi.org/10.1093/nar/gku1195

Flagel, L.E., and Wendel, J.F. (2009). Gene duplication and evolutionary novelty in plants. New Phytol *183*, 557–564. http://dx.doi.org/10.1111/j.1469-8137.2009.02923.x

Flicek, P., Ahmed, I., Amode, M.R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fairley, S., *et al.* (2013). Ensembl 2013. Nucleic Acids Res. *41*, D48–55. http://dx.doi.org/10.1093/nar/gks1236

Gan, X., Stegle, O., Behr, J., Steffen, J.G., Drewe, P., Hildebrand, K.L., Lyngsoe, R., Schultheiss, S.J., Osborne, E.J., Sreedharan, V.T., *et al.* (2011). Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*. Nature *477*, 419–423. http://dx.doi.org/10.1038/nature10414

Goff, S.A., Ricke, D., Lan, T.-H., Presting, G., Wang, R., Dunn, M., Glazebrook, J., Sessions, A., Oeller, P., Varma, H., *et al.* (2002). A Draft Sequence of the Rice Genome (*Oryza sativa* L. ssp. *japonica*). Science *296*, 92–100. http://dx.doi.org/10.1126/science.1068275

Goodstein, D.M., Shu, S., Howson, R., Neupane, R., Hayes, R.D., Fazo, J., Mitros, T., Dirks, W., Hellsten, U., Putnam, N., *et al.* (2012). Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res. *40*, D1178–1186. http://dx.doi.org/10.1093/nar/gkr944

Harris, T.D., Buzby, P.R., Babcock, H., Beer, E., Bowers, J., Braslavsky, I., Causey, M., Colonell, J., DiMeo, J., Efcavitch, J.W., *et al.* (2008). Single-molecule DNA sequencing of a viral genome. Science *320*, 106. http://dx.doi.org/10.1126/science.1150427

Hirsch, C.D., Hamilton, J.P., Childs, K.L., Cepela, J., Crisovan, E., Vaillancourt, B., Hirsch, C.N., Habermann, M., Neal, B., and Buell, C.R. (2014). Spud DB: a resource for mining sequences, genotypes, and phenotypes to accelerate potato breeding. The Plant Genome 7. http://dx.doi.org/10.3835/plantgenome2013.12.0042

Horner, D.S., Pavesi, G., Castrignano, T., De Meo, P.D., Liuni, S., Sammeth, M., Picardi, E., and Pesole, G. (2010). Bioinformatics approaches for genomics and post genomics applications of next-generation sequencing. Brief Bioinform. *11*, 181–197. http://dx.doi.org/10.1093/bib/bbp046

Huerta-Cepas, J., Szklarczyk, D., Forslund, K., Cook, H., Heller, D., Walter, M.C., Rattei, T., Mende, D.R., Sunagawa, S., Kuhn, M., *et al.* (2016). eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. Nucleic Acids Res. *44*, D286–293. http://dx.doi.org/10.1093/nar/gkv1248

Hughes, A.L. (2005). Gene duplication and the origin of novel proteins. Proc. Natl. Acad. Sc.i U.S.A. *102*, 8791–8792. http://dx.doi.org/10.1073/pnas.0503922102

Izzah, N.K., Lee, J., Jayakodi, M., Perumal, S., Jin, M., Park, B.-S., Ahn, K., and Yang, T.-J. (2014). Transcriptome sequencing of two parental lines of cabbage (Brassica oleracea L. var. capitata L.) and construction of an EST-based genetic map. BMC Genomics *15*, 149. http://dx.doi.org/10.1186/1471-2164-15-149

Jaillon, O., Aury, J.M., Noel, B., Policriti, A., Clepet, C., Casagrande, A., Choisne, N., Aubourg, S., Vitulo, N., Jubin, C., *et al.* (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature *449*, 463–467. http://dx.doi.org/10.1038/nature06148

Jia, J., Zhao, S., Kong, X., Li, Y., Zhao, G., He, W., Appels, R., Pfeifer, M., Tao, Y., Zhang, X., *et al.* (2013). Aegilops tauschii draft genome sequence reveals a gene repertoire for wheat adaptation. Nature *496*, 91–95. http://dx.doi.org/10.1038/nature12028

Jiao, Y., Leebens-Mack, J., Ayyampalayam, S., Bowers, J.E., McKain, M.R., McNeal, J., Rolf, M., Ruzicka, D.R., Wafula, E., Wickett, N.J., *et al.* (2012). A genome triplication associated with early diversification of the core eudicots. Genome Biol. *13*, R3. http://dx.doi.org/10.1186/gb-2012-13-1-r3

Jiao, Y., Wickett, N.J., Ayyampalayam, S., Chanderbali, A.S., Landherr, L., Ralph, P.E., Tomsho, L.P., Hu, Y., Liang, H., Soltis, P.S., *et al.* (2011). Ancestral polyploidy in seed plants and angiosperms. Nature *473*, 97–100. http://dx.doi.org/10.1038/nature09916

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

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. *28*, 27-30.

Kawahara, Y., de la Bastide, M., Hamilton, J.P., Kanamori, H., McCombie, W.R., Ouyang, S., Schwartz, D.C., Tanaka, T., Wu, J., Zhou, S., *et al.* (2013). Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data. Rice (NY) *6*, 4. http://dx.doi.org/10.1186/1939-8433-6-4

Kawahara, Y., Oono, Y., Wakimoto, H., Ogata, J., Kanamori, H., Sasaki, H., Mori, S., Matsumoto, T., and Itoh, T. (2016). TENOR: Database for comprehensive mRNA-Seq experiments in rice. Plant Cell Physiol. 57, e7.

Kersey, P.J., Allen, J.E., Armean, I., Boddu, S., and Bolt, B.J. (2016). Ensembl Genomes 2016: more genomes, more complexity. *44*, D574-580. http://dx.doi.org/10.1093/nar/gkv1209

Kim, S., Park, M., Yeom, S.-I., Kim, Y.-M., Lee, J.M., Lee, H.-A., Seo, E., Choi, J., Cheong, K., Kim, K.-T., *et al.* (2014). Genome sequence of the hot pepper provides insights into the evolution of pungency in *Capsicum* species. Nat. Genet. *46*, 270–278. http://dx.doi.org/10.1038/ng.2877

Kim, S., Plagnol, V., Hu, T.T., Toomajian, C., Clark, R.M., Ossowski, S., Ecker, J.R., Weigel, D., and Nordborg, M. (2007). Recombination and linkage disequilibrium in *Arabidopsis thaliana*. Nat. Genet. *39*, 1151–1155. http://dx.doi.org/10.1038/ng2115

Klepikova, A.V., Logacheva, M.D., Dmitriev, S.E., and Penin, A.A. (2015). RNA-seq analysis of an apical meristem time series reveals a critical point in *Arabidopsis thaliana* flower initiation. BMC Genomics *16*, 466.

Kodama, Y., Shumway, M., and Leinonen, R. (2012). The Sequence Read Archive: explosive growth of sequencing data. Nucleic Acids Res. *40*, D54–56. http://dx.doi.org/10.1093/nar/gkr854

Kodzius, R., Kojima, M., Nishiyori, H., Nakamura, M., Fukuda, S., Tagami, M., Sasaki, D., Imamura, K., Kai, C., Harbers, M., *et al.* (2006). CAGE: cap analysis of gene expression. Nat Methods *3*, 211–222. http://dx.doi.org/10.1038/nmeth0306-211

Kristensen, D.M., Wolf, Y.I., Mushegian, A.R., and Koonin, E.V. (2011). Computational methods for Gene Orthology inference. Brief Bioinform *12*, 379–391. http://dx.doi.org/10.1093/bib/bbr030

Lam, H.M., Xu, X., Liu, X., Chen, W., Yang, G., Wong, F.L., Li, M.W., He, W., Qin, N., Wang, B., *et al.* (2010). Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. Nat. Genet. *42*, 1053–1059. http://dx.doi.org/10.1038/ng.715

Lamesch, P., Berardini, T.Z., Li, D., Swarbreck, D., Wilks, C., Sasidharan, R., Muller, R., Dreher, K., Alexander, D.L., Garcia-Hernandez, M., *et al.* (2012). The *Arabidopsis* Information Resource (TAIR): improved gene annotation and new tools. Nucleic Acids Res. *40*, D1202–1210. http://dx.doi.org/10.1093/nar/gkr1090

Leinonen, R., Akhtar, R., Birney, E., Bower, L., Cerdeno-Tarraga, A., Cheng, Y., Cleland, I., Faruque, N., Goodgame, N., Gibson, R., *et al.* (2011). The European Nucleotide Archive. Nucleic Acids Res. *39*, D28-31. http://dx.doi.org/10.1093/nar/gkq967

Ling, H.Q., Zhao, S., Liu, D., Wang, J., Sun, H., Zhang, C., Fan, H., Li, D., Dong, L., Tao, Y., *et al.* (2013). Draft genome of the wheat A-genome progenitor *Triticum urartu*. Nature *496*, 87–90. http://dx.doi.org/10.1038/nature11997

Lynch, M., and Conery, J.S. (2000). The evolutionary fate and consequences of duplicate genes. Science *290*, 1151-1155. http://dx.doi.org/10.1126/science.290.5494.1151

Ma, J.K., Drake, P.M., and Christou, P. (2003). The production of recombinant pharmaceutical proteins in plants. Nat. Rev. Genet. *4*, 794–805. http://dx.doi.org/10.1038/nrg1177

MacLean, D., Jones, J.D., and Studholme, D.J. (2009). Application of 'next-generation' sequencing technologies to microbial genetics. Nat. Rev. Microbiol. *7*, 287–296. http://dx.doi.org/10.1038/nrmicro2122

Maere, S., De Bodt, S., Raes, J., Casneuf, T., Van Montagu, M., Kuiper, M., and Van de Peer, Y. (2005). Modeling gene and genome duplications in eukaryotes. Proc. Natl. Acad. Sci. U.S.A. *102*, 5454–5459. http://dx.doi.org/10.1073/pnas.0501102102

Magadum, S., Banerjee, U., Murugan, P., Gangapur, D., and Ravikesavan, R. (2013). Gene duplication as a major force in evolution. J. Genet. *92,* 155-161.

Magi, A., Benelli, M., Gozzini, A., Girolami, F., Torricelli, F., and Brandi, M.L. (2010). Bioinformatics for Next Generation Sequencing Data. Genes *1,* 294–307. http://dx.doi.org/10.3390/genes1020294

Makita, Y., Shimada, S., Kawashima, M., Kondou-Kuriyama, T., Toyoda, T., and Matsui, M. (2015). MOROKOSHI: transcriptome database in *Sorghum bicolor*. Plant Cell Physiol. *56,* e6.

Mardis, E.R. (2008a). The impact of next-generation sequencing technology on genetics. Trends Genet. *24,* 133–141. http://dx.doi.org/10.1016/j.tig.2007.12.007

Mardis, E.R. (2008b). Next-generation DNA sequencing methods. Annu. Rev. Genomics Hum. Genet. *9,* 387–402. http://dx.doi.org/10.1146/annurev.genom.9.081307.164359

Mardis, E.R. (2009). New strategies and emerging technologies for massively parallel sequencing: applications in medical research. Genome Med *1,* 40. http://dx.doi.org/10.1186/gm40

Mayer, K.F., Waugh, R., Brown, J.W., Schulman, A., Langridge, P., Platzer, M., Fincher, G.B., Muehlbauer, G.J., Sato, K., Close, T.J., *et al.* (2012). A physical, genetic and functional sequence assembly of the barley genome. Nature *491,* 711–716. http://dx.doi.org/10.1038/nature11543

Moniz de Sa, M., and Drouin, G. (1996). Phylogeny and substitution rates of angiosperm actin genes. Mol. Biol. Evol. *13,* 1198–1212.

Moretto, M., Sonego, P., Dierckxsens, N., Brilli, M., Bianco, L., Ledezma-Tejeida, D., Gama-Castro, S., Galardini, M., Romualdi, C., Laukens, K., *et al.* (2015). COLOMBOS v3.0: leveraging gene expression compendia for cross-species analyses. Nucleic Acids Res. *44,* D620–D623.

Morozova, O., and Marra, M.A. (2008a). Applications of next-generation sequencing technologies in functional genomics. Genomics *92,* 255–264. http://dx.doi.org/10.1016/j.ygeno.2008.07.001

Morozova, O., and Marra, M.A. (2008b). From cytogenetics to next-generation sequencing technologies: advances in the detection of genome rearrangements in tumors. Biochem Cell Biol *86,* 81-91. http://dx.doi.org/10.1139/o08-003

Morrissy, A.S., Morin, R.D., Delaney, A., Zeng, T., McDonald, H., Jones, S., Zhao, Y., Hirst, M., and Marra, M.A. (2009). Next-generation tag sequencing for cancer gene expression profiling. Genome Res. *19,* 1825–1835. http://dx.doi.org/10.1101/gr.094482.109

Mukherjee, G., Abeygunawardena, N., Parkinson, H., Contrino, S., Durinck, S., Farne, A., Holloway, E., Lilja, P., Moreau, Y., Oezcimen, A., *et al.* (2005). Plant-Based Microarray Data at the European Bioinformatics Institute. Introducing AtMIAMExpress, a Submission Tool for *Arabidopsis* Gene Expression Data to ArrayExpress. Plant Physiol *139,* 632-636. http://dx.doi.org/10.1104/pp.105.063156

Nakano, M., Nobuta, K., Vemaraju, K., Tej, S.S., Skogen, J.W., and Meyers, B.C. (2006). Plant MPSS databases: signature-based transcriptional resources for analyses of mRNA and small RNA. Nucleic Acids Res. *34,* D731–D735.

Nawrocki, E.P., Burge, S.W., Bateman, A., Daub, J., Eberhardt, R.Y., Eddy, S.R., Floden, E.W., and Gardner, P.P. (2015). Rfam 12.0: updates to the RNA families database. Nucleic Acids Res. *43,* D130–137. http://dx.doi.org/10.1093/nar/gku1063

Nystedt, B., Street, N.R., Wetterbom, A., Zuccolo, A., Lin, Y.C., Scofield, D.G., Vezzi, F., Delhomme, N., Giacomello, S., Alexeyenko, A., *et al.* (2013). The Norway spruce genome sequence and conifer genome evolution. Nature *497,* 579-584. http://dx.doi.org/10.1038/nature12211

O'Brien, K.P., Remm, M., and Sonnhammer, E.L. (2005). Inparanoid: a comprehensive database of eukaryotic orthologs. Nucleic Acids Res. *33,* D476-480. http://dx.doi.org/10.1093/nar/gki107

O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., *et al.* (2015). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. http://dx.doi.org/10.1093/nar/gkv1189 http://www.ncbi.nlm.nih.gov/refseq

O'Rourke, J.A., Fu, F., Bucciarelli, B., Yang, S.S., Samac, D.A., Lamb, J.F.S., Monteros, M.J., Graham, M.A., Gronwald, J.W., Krom, N., *et al.* (2015). The *Medicago sativa* gene index 1.2: a web-accessible gene expression atlas for investigating expression differences between *Medicago sativa* subspecies. BMC Genomics *16,* 1–17.

O'Rourke, J.A., Iniguez, L.P., Fu, F., Bucciarelli, B., Miller, S.S., Jackson, S.A., McClean, P.E., Li, J., Dai, X., Zhao, P.X., *et al.* (2014). An RNA-Seq based gene expression atlas of the common bean. BMC Genomics *15,* 866.

Page, G.P., and Coulibaly, I. (2008). Bioinformatic Tools for Inferring Functional Information from Plant Microarray Data: Tools for the First Steps. Int J Plant Genomics *2008,* 147563. http://dx.doi.org/10.1155/2008/147563

Paterson, A.H., Bowers, J.E., Bruggmann, R., Dubchak, I., Grimwood, J., Gundlach, H., Haberer, G., Hellsten, U., Mitros, T., Poliakov, A., *et al.* (2009). The Sorghum bicolor genome and the diversification of grasses. Nature *457,* 551–556. http://dx.doi.org/10.1038/nature07723

Pearce, S., Vazquez-Gross, H., Herin, S.Y., Hane, D., Wang, Y., Gu, Y.Q., and Dubcovsky, J. (2015). WheatExp: an RNA-seq expression database for polyploid wheat. BMC Plant Biol. *15,* 1–8.

Porreca, G.J., Shendure, J., and Church, G.M. (2006). Polony DNA sequencing. Curr. Protocols Mol. Biol., 7.8.1–7.8.22. http://dx.doi.org/10.1002/0471142727.mb0708s76

Proost, S., Van Bel, M., Vaneechoutte, D., Van de Peer, Y., Inze, D., Mueller-Roeber, B., and Vandepoele, K. (2015). PLAZA 3.0: an access point for plant comparative genomics. Nucleic Acids Res. *43,* D974–981. http://dx.doi.org/10.1093/nar/gku986

Rouard, M., Guignon, V., Aluome, C., Laporte, M.A., Droc, G., Walde, C., Zmasek, C.M., Perin, C., and Conte, M.G. (2011). GreenPhylDB v2.0: comparative and functional genomics in plants. Nucleic Acids Res. *39,* D1095-1102. http://dx.doi.org/10.1093/nar/gkq811

Sanger, F., Nicklen, S., and Coulson, A.R. (1977). DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. U.S.A. *74,* 5463–5467.

Schatz, M.C., Witkowski, J., and McCombie, W.R. (2012). Current challenges in *de novo* plant genome sequencing and assembly. Genome Biol *13*, 243. http://dx.doi.org/10.1186/gb4015

Schmid, M., Davison, T.S., Henz, S.R., Pape, U.J., Demar, M., Vingron, M., Scholkopf, B., Weigel, D., and Lohmann, J.U. (2005). A gene expression map of *Arabidopsis thaliana* development. Nat. Genet. *37*, 501-506. http://dx.doi.org/10.1038/ng1543

Schnable, P.S., Ware, D., Fulton, R.S., Stein, J.C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T.A., *et al.* (2009). The B73 maize genome: complexity, diversity, and dynamics. Science *326*, 1112–1115. http://dx.doi.org/10.1126/science.1178534

Schuster, S.C. (2008). Next-generation sequencing transforms today's biology. Nat. Methods *5*, 16–18. http://dx.doi.org/10.1038/nmeth1156

Severin, A.J., Woody, J.L., Bolon, Y.-T., Joseph, B., Diers, B.W., Farmer, A.D., Muehlbauer, G.J., Nelson, R.T., Grant, D., Specht, J.E., *et al.* (2010). RNA-Seq Atlas of Glycine max: A guide to the soybean transcriptome. BMC Plant Biol. *10*, 160.

Sierro, N., Battey, J.N.D., Ouadi, S., Bakaher, N., Bovet, L., Willig, A., Goepfert, S., Peitsch, M.C., and Ivanov, N.V. (2014). The tobacco genome sequence and its comparison with those of tomato and potato. Nat. Commun. 5. http://dx.doi.org/10.1038/ncomms4833

Tang, H., and Zhao, Z. (2015). Bioinformatics drives the applications of next-generation sequencing in translational biomedical research. Methods *79–80*, 1–2. http://dx.doi.org/10.1016/j.ymeth.2015.04.035

Tello-Ruiz, M.K., Stein, J., Wei, S., Preece, J., Olson, A., Naithani, S., Amarasinghe, V., Dharmawardhana, P., Jiao, Y., Mulvaney, J., *et al.* (2016). Gramene 2016: comparative plant genomics and pathway resources. Nucleic Acids Res. *44*, D1133–1140. http://dx.doi.org/10.1093/nar/gkv1179

The *Arabidopsis* Genome Initiative (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature *408*, 796-815. http://dx.doi.org/10.1038/35048692

The Tomato Genome Consortium (2012). The tomato genome sequence provides insights into fleshy fruit evolution. Nature *485*, 635–641. http://dx.doi.org/10.1038/nature11119

Trachana, K., Forslund, K., Larsson, T., Powell, S., Doerks, T., von Mering, C., and Bork, P. (2014). A phylogeny-based benchmarking test for orthology inference reveals the limitations of function-based validation. PLOS ONE *9*, e111122. http://dx.doi.org/10.1371/journal.pone.0111122

Tringe, S., and Coleman-Derr, D. (2014). Building the crops of tomorrow: advantages of symbiont-based approaches to improving abiotic stress tolerance (Ernest Orlando Lawrence Berkeley National Laboratory, Berkeley, CA (US)).

UniProt Consortium (2015). UniProt: a hub for protein information. Nucleic Acids Res. *43*, D204–212. http://dx.doi.org/10.1093/nar/gku989

Velasco, R., Zharkikh, A., Troggio, M., Cartwright, D.A., Cestaro, A., Pruss, D., Pindo, M., FitzGerald, L.M., Vezzulli, S., Reid, J., *et al.* (2007). A High Quality Draft Consensus Sequence of the Genome of a Heterozygous Grapevine Variety. PLOS ONE *2*, e1326. http://dx.doi.org/10.1371/journal.pone.0001326

Velculescu, V.E., Zhang, L., Vogelstein, B., and Kinzler, K.W. (1995). Serial analysis of gene expression. Science *270*, 484–487. http://dx.doi.org/10.1126/science.270.5235.484

Vitulo, N., Forcato, C., Carpinelli, E.C., Telatin, A., Campagna, D., D'Angelo, M., Zimbello, R., Corso, M., Vannozzi, A., Bonghi, C., *et al.* (2014). A deep survey of alternative splicing in grape reveals changes in the splicing machinery related to tissue, stress condition and genotype. BMC Plant Biol *14*, 99. http://dx.doi.org/10.1186/1471-2229-14-99

Wang, G., Liu, Y., Zhu, D., Klau, G.W., and Feng, W. (2015). Bioinformatics Methods and Biological Interpretation for Next-Generation Sequencing Data. Biomed Res Int *2015*, 690873. http://dx.doi.org/10.1155/2015/690873

Wang, J.-P.Z., Lindsay, B.G., Cui, L., Wall, P.K., Marion, J., Zhang, J., and dePamphilis, C.W. (2005). Gene capture prediction and overlap estimation in EST sequencing from one or multiple libraries. BMC Bioinformatics *6*, 300. http://dx.doi.org/10.1186/1471-2105-6-300

Wang, X., Wang, H., Wang, J., Sun, R., Wu, J., Liu, S., Bai, Y., Mun, J.H., Bancroft, I., Cheng, F., *et al.* (2011). The genome of the mesopolyploid crop species Brassica rapa. Nat. Genet. *43*, 1035–1039. http://dx.doi.org/10.1038/ng.919

Weigel, D., and Mott, R. (2009). The 1001 genomes project for *Arabidopsis thaliana*. Genome Biol *10*, 107. http://dx.doi.org/10.1186/gb-2009-10-5-107

Wilson, S.A., and Roberts, S.C. (2014). Metabolic engineering approaches for production of biochemicals in food and medicinal plants. Curr. Opin. Biotechnol. *26*, 174–182. http://dx.doi.org/10.1016/j.copbio.2014.01.006

Wolfe, K.H. (2001). Yesterday's polyploids and the mystery of diploidization. Nat Rev Genet *2*, 333–341. http://dx.doi.org/10.1038/35072009

Xu, C., Jiao, C., Zheng, Y., Sun, H., Liu, W., Cai, X., Wang, X., Liu, S., Xu, Y., Mou, B., *et al.* (2015). *De novo* and comparative transcriptome analysis of cultivated and wild spinach. Sci Rep *5*, 17706.

Xu, X., Pan, S., Cheng, S., Zhang, B., Mu, D., Ni, P., Zhang, G., Yang, S., Li, R., Wang, J., *et al.* (2011). Genome sequence and analysis of the tuber crop potato. Nature *475*, 189–195. http://dx.doi.org/10.1038/nature10158

Yang, M.Q., Athey, B.D., Arabnia, H.R., Sung, A.H., Liu, Q., Yang, J.Y., Mao, J., and Deng, Y. (2009). High-throughput next-generation sequencing technologies foster new cutting-edge computing techniques in bioinformatics. BMC Genomics *10 Suppl 1*, I1. http://dx.doi.org/10.1186/1471-2164-10-s1-i1

Yuan, J.S., Tiller, K.H., Al-Ahmad, H., Stewart, N.R., and Stewart, C.N., Jr. (2008). Plants to power: bioenergy to fuel the future. Trends Plant Sci *13*, 421–429. http://dx.doi.org/10.1016/j.tplants.2008.06.001