Esposito *et al. Chem. Biol. Technol. Agric. (2016) 3:9*
DOI 10.1186/s40538-016-0054-8

Chemical and Biological Technologies in Agriculture

REVIEW

Open Access

Bioinformatics for agriculture in the Next-Generation sequencing era

Alfonso Esposito, Chiara Colantuono, Valentino Ruggieri and Maria Luisa Chiusano*

Abstract: The key role of bioinformatics is acquiring striking importance in the era of outstanding advances in omics technologies for its fundamental support in describing the multifaceted aspects of biological functionalities. The manifold omics efforts flourishing worldwide are also contributing fundamental novelties in many aspects of agricultural sciences and, as a consequence, bioinformatics is acquiring a crucial role also in these research fields. Indeed, the transformation of natural environment for improvement of goods from animal, plants, and microbial worlds for human nutrition and health requires the comprehension of the molecular mechanisms influencing the structure and the function of the individuals, the populations, and the communities. The expanding knowledge about the molecules and the mechanisms associated with specific phenotypic traits and specific responses to biotic or abiotic stresses, complemented with the predictive power of bioinformatics, has an impact on agriculture practices and favors innovative methods in diagnostics, monitoring, and traceability, improving human benefits at lower costs, thus supporting sustainability. We here describe main bioinformatics approaches in the era of Next-Generation Sequencing for its impact in genomics, transcriptomics, and metagenomics efforts, describing their role in agriculture sciences. We aim to introduce common aspects, open questions and perspectives in this cutting-edge field of research.

Keywords: NGS, Agriculture bioinformatics, Omics technologies, Biological challenges

Introduction

Bioinformatics emerged from the initial requirement of suitable informatics for biological data organization, management, and distribution [1], but soon it revealed also fundamental in providing tools for data analysis, interpretation, and modeling. Moreover, thanks to bioinformatics, it was possible to analyze and understand structure and function not only of single bio-molecules but also of larger molecular collections, derived from the so-called omics experimental approaches. These efforts permit to depict different aspects (genomics, transcriptomics, proteomics, metabolomics, etc.) of the biomolecular organization of complex biological systems, from cells to ecosystems. The fast spreading of omics techniques, with its growing power and more accessible costs, drastically increased the amount of molecular data collections from different levels of organization of an organism or an environmental sample. This favored a holistic view on systems organization and functionality, further challenging bioinformatics with data size and the need of integrative efforts [2, 3]. The recent introduction of Next-Generation Sequencing (NGS) technologies (Table 1) further revolutionized the sequencing of nucleic acids contributing to a new era in omics approaches. Indeed, on one hand these technologies introduced an incredible efficiency in terms of experimental execution time and a deeper resolution. On the other hand, they stimulated an unexpected interest from scientists due to the higher affordability in terms of experimental procedures and economical requirements. We here introduce the novelties that the advent of NGS technologies contributed in agriculture, overviewing the main bioinformatics strategies and challenges, as well as perspectives in the field.

Review

1-Single and multi species genomics for agriculture

Genomics, transcriptomics, proteomics, and metabolomics may contribute to the comprehension of the organization and the functionality of biological systems,

*Correspondence: chiusano@unina.it
Department of Agricultural Sciences, University of Naples Federico II, 80055 Portici (Na), Italy

Table 1 Main features of the most used NGS technologies in omic studies

Technology	Read length	Yield (Reads per Run)	Reference
Roche 454	700	~700 thousand	http://454.com/products/gs-flx-system/
Illumina HiSeq	300	~300 billion	http://www.illumina.com/systems/hiseq-3000-4000/specifications.html
SOLiD	100	~200 billion	https://www.thermofisher.com/it/en/home/life-science/sequencing/next-generation-sequencing/solid-next-generation-sequencing/solid-next-generation-sequencing-systems-reagents-accessories.html
Ion Torrent	200	~60 billion	http://www.thermofisher.com/it/en/home/life-science/sequencing/next-generation-sequencing/ion-torrent-next-generation-sequencing-workflow/ion-torrent-next-generation-sequencing-run-sequence/ion-proton-system-for-next-generation-sequencing/ion-proton-system-specifications.html
PacBio RS II	14,000	~47 thousand	http://files.pacb.com/pdf/PacBio_RS_II_Brochure.pdf

with the possibility to also trace molecular variability during development, in different conditions, such as physiological, pathological, or influenced by environmental changes [4].

Samples for omics studies can derive from one or several individuals of a species (a population), or from multiple species (a community) [5–7]. The differences among these approaches consist mainly in the objectives of the specific studies.

In single individual approaches, the organization and the functionality of specific cells, tissues, or organs (e.g., roots, fruit, rumen) are investigated, mainly to identify factors influencing emergent properties, like the quality and the shape. This also paves the way to the characterization of even more complex traits (e.g., yield, resistance to stresses, diseases) or processes (e.g., fruit ripening, growth efficiency, senescence) [4, 8].

The description of the molecular components in a population of the same species aims to the understanding of evolutionary processes influencing genetic variability. This can also widely contribute to dissect complex quantitative traits by identifying novel and superior alleles [9, 10] or to assess the impact of genetic variation on patterns of gene expression and on phenotypic plasticity in response to environmental changes [11].

The study of the collective genetic pool deriving from communities is termed "metagenomics" [12], a term dating back to 1998. The community can derive from environmental samples, such as soil [13], seawater [14], or other [15], but also part of individuals, such as gut or roots [16, 17]. Metagenomics usually aims to describe the prokaryotic component of the community, but may be also useful to trace the different eukaryotes existing in a specific environment [18].

Nucleic acid sequencing always contributed the majority of the data in all the approaches here summarized. This is why the recent introduction of NGS technologies impressively impacted the productivity and the advancement in these research fields.

2-Impact of NGS in agriculture

The multifaceted scientific topics in agriculture sciences may be consistently supported by NGS omics for single individuals, populations, or communities [19–21].

The sequencing of whole genomes from several species permits to define their organization and provides the starting point for understanding their functionality [22–25], therefore favoring human agriculture practice. Efforts addressed to the achievement of an appropriate knowledge of associated molecular information, such as the one arising from transcriptome and proteome sequencing, are also essential to better depict the gene content of a genome and its main functionalities. These efforts indeed led to major advancements in all biological sciences [4] and in agriculture as well [8, 26]. Moreover, the elucidation of the complexity of genes and their networking is also fundamental for being eventually translated into breeding practice for crops or livestock, contributing to their health, resistance, and productivity. Indeed, the contribution of genomics to agriculture spans the identification and the manipulation of genes linked to specific phenotypic traits [27] as well as genomics breeding by marker-assisted selection of variants [28, 29]. The so-called "agricultural genomics" (or agri-genomics), indeed, aims to find innovative solutions through the study of crops or livestock genomes, achieving information for protection [30, 31] and sustainable productivity for food industry, but also for alternative aspects like energy production or design of pharmaceuticals [32–35].

Plant, soil, and livestock microbiome also play a key role in agriculture since it determines plant fitness [36, 37], soil biogeochemical properties [38], and affects both yield and quality traits [39, 40]. However, little knowledge is available for microbes and the communities in which they are included. As an example, it is acknowledged that soil is one of the biggest carbon reservoirs on earth, and prokaryotes constitute an important amount of the soil biomass [41]. However, culture-independent studies in the last three decades showed that, although sequencing

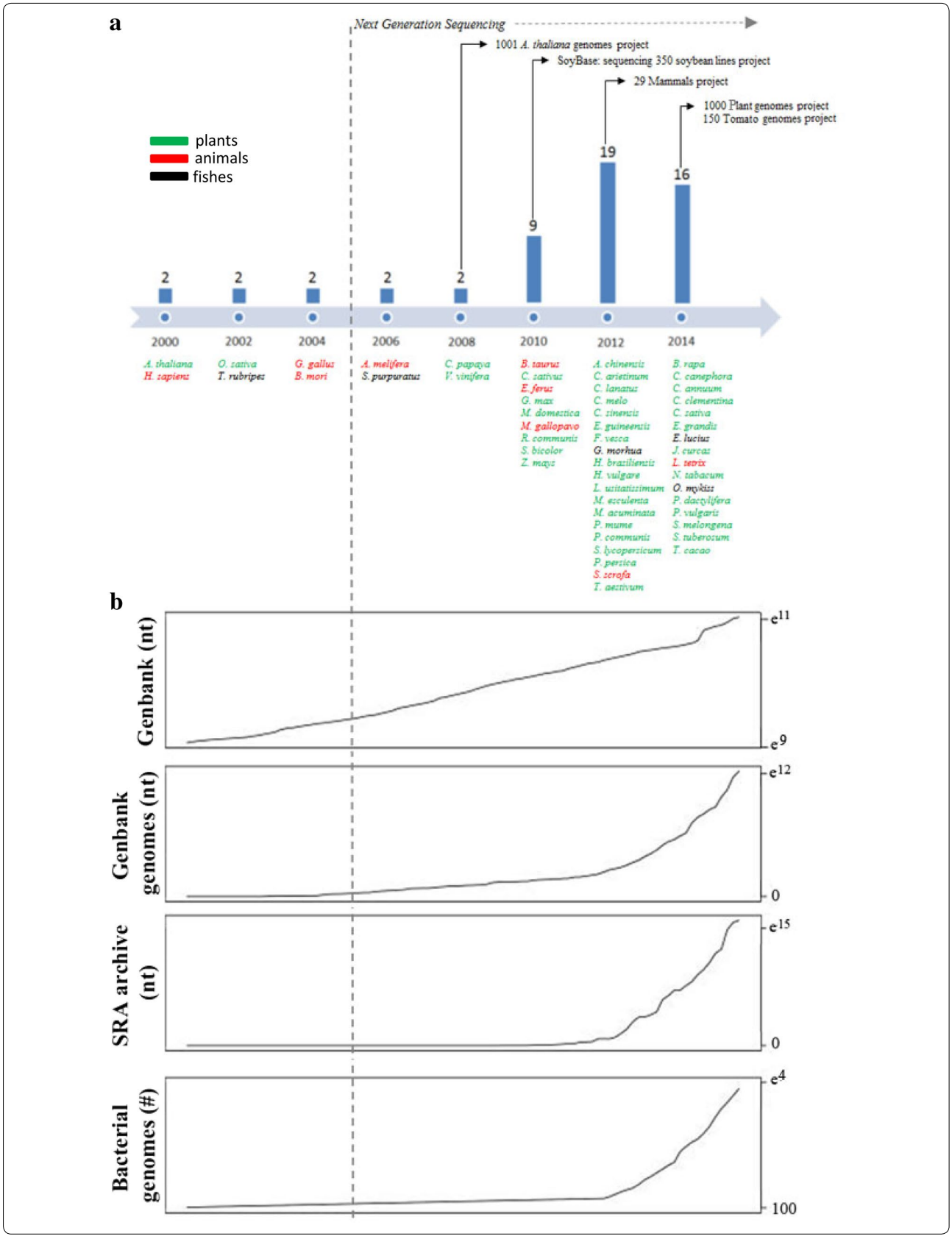
strategies are fast evolving, the great majority of bacterial species is still unknown [42, 43]. Therefore most of the methods used for profiling microbial communities and describe their main functional features are now adopting whole DNA extraction and the use of NGS on the entire sample, with the objective of sequencing and characterizing DNA fragments of all the species included, i.e., the metagenome. The application of metagenomics in agriculture also proved to be appropriate for depicting the complex patterns of interactions occurring among microorganisms in soil [44] and in plant rhizosphere [45], as well as in specific tissues or organs [6, 46, 47]. Metagenomics recently revealed to be useful to trace the shift in taxonomic composition and functional redundancy of microbial communities in rhizosphere and in soil in connection to environmental changes associated to fertilization [48] and agricultural management [49, 50]. Metagenomics studies can also help deciphering the role of soil bacteria in plant nutrition [51, 52] or in the cycle of the elements [53]. Further applications can lead to the discovery of new genes, bio-products, plant growth promoting microorganisms consortia, useful for understanding relevant aspects such as response to stresses [36] or dysbiosis [54–56].

3-The revolution in the omics technologies and the impact on bioinformatics

The introduction of omics approaches strongly impacted bioinformatics in data collection, organization [57–59], integration, and in the implementation of suitable data mining tools [2, 60]. The support of efficient bioinformatics favored the introduction of the so-called high-throughput technologies, paving the way to the flourishing of genome sequencing efforts of key model species, such as *Homo sapiens* and *Arabidopsis thaliana* (Fig. 1). The same technologies were then exploited to further push forward the genome sequencing of other model and non-model species, many of which of agriculture interest. These efforts were also preceded or accompanied by transcriptome sequencing efforts using different technologies [61–64], in support of gene prediction [65], but also for depicting transcriptional processes and define cell functionality in physiological, pathological, or stress conditions. These approaches also required the design of appropriate resources to distribute the data [66] and/or dedicated collections of processed results [67–70] to all the interested scientific community, enhancing the need for suitable pipelines for moving from raw to value added information and integrative data mining [71, 72] (Fig. 2, Table 2).

NGS strikingly contributed to expand the number of genomes currently completely sequenced (Fig. 1), as well as to the establishment of novel ambitious efforts, for instance, those focused on multi-genome sequencing [24, 25, 73] or those aiming to define global metagenomes from different environmental samples to define reference collections [74, 75]. These technologies are also exploited for the production of alternative, related collections, namely from transcriptomics, epigenomics, and metagenomics projects. The unexpected amount of raw data the new technologies are providing requires also dedicated storage for centralized data maintenance, currently solved by the SRA system [76]. Worthy to note is the size reached by the SRA archive in a short time span when compared to the entire nucleotide collection currently available (Fig. 1). NGS data size represents a big challenge for bioinformatics. Indeed, main computational tasks are today focused on the optimization and adaptation of typical methodologies in bioinformatics to the magnitude of NGS collections.

In Table 2, main methodologies and resources exploited for NGS data analyses are summarized. Data are usually delivered by sequencing centers in the form of raw, fragmented sequences, to be pre-processed, i.e., cleaned, from additional fragments due to the specific technology employed (Table 1), such as vectors, adaptors, barcodes, or other contaminations (Table 2). Structure definition from fragmented data usually requires an assembly step to reconstruct the most reliable original molecules, such as longer genomic sequences or transcripts. The assembly is based on sequence alignments driven by identical regions shared by the fragments (Table 2). The assembly procedure may include already available reference sequences (guided approaches), as in the case of transcript assemblies based on a genome reference, or they are based on reference-free methodologies (de novo) [77]. This procedure is a key step for many different applications, indeed high-quality longer backbones are useful to properly proceed towards successive steps in data processing, which are mainly structure and functional assignments and predictions (Fig. 2). In genomics, transcriptomics, population genetics, and metagenomics, these are widespread fundamental tasks solved by different computational methodologies (Table 2), though based on similar strategies (Fig. 2), mainly ab initio approaches or similarity-based ones. Ab initio-based algorithms exploit complex probabilistic models to detect expected features (genes, motives, propensities) as defined on the basis of training datasets that support the proper identification of specific features. Similarity-based methods, on the other hand, are the principal and more frequently used approaches in bioinformatics since they permit identifications, predictions, structure, and functional assignments. They rely on sequence or tridimensional similarities exploiting a typical concept in biology, which considers similarity in structure as a


a
Next Generation Sequencing
1001 A. thaliana genomes project
SoyBase: sequencing 350 soybean lines project
29 Mammals project
1000 Plant genomes project
150 Tomato genomes project
plants
animals
fishes
2
2
2
2
2
9
19
16
2000
2002
2004
2006
2008
2010
2012
2014
A. thaliana
H. sapiens
O. sativa
T. rubripes
G. gallus
B. mori
A. melifera
S. purpuratus
C. papaya
V. vinifera
B. taurus
C. sativus
E. ferus
G. max
M. domestica
M. gallopavo
R. communis
S. bicolor
Z. mays
A. chinensis
C. arietinum
C. lanatus
C. melo
C. sinensis
E. guineensis
F. vesca
G. morhua
H. brasiliensis
H. vulgare
L. usitatissimum
M. esculenta
M. acuminata
P. mume
P. communis
S. lycopersicum
P. persica
S. scrofa
T. aestivum
B. rapa
C. canephora
C. annuum
C. clementina
C. sativa
E. grandis
E. lucius
J. curcas
L. sativa
N. tabacum
O. mykiss
P. dactylifera
P. vulgaris
S. melongena
S. tuberosum
T. cacao
b
Genbank (nt)
e11
e9
Genbank genomes (nt)
e12
0
SRA archive (nt)
e15
0
Bacterial genomes (#)
e4
100

(See figure on previous page.)
Fig. 1 (**a**) Timeline from 2000 to 2014 indicating the release of the completely sequenced genomes for some of the major plants (*green*), animals (*red*), and fishes (*black*) of interest in agriculture. The *dashed line* indicates the start of Next-Generation Sequencing (NGS) era. The start of major massive sequencing projects are also indicated: the 1001 *A. thaliana* genomes project (http://1001genomes.org/), the SoyBase project (sequencing of 350 soybean lines) (http://www.soybase.org/), the 29 Mammals genomes project (https://www.broadinstitute.org/scientific-community/science/projects/mammals-models/29-mammals-project), the 1000 Plant genomes project (https://sites.google.com/a/ualberta.ca/onekp/), and the 150 Tomato genomes ReSequencing project (http://www.tomatogenome.net/). (**b**) *Graphs* indicate the growth as number of nucleotides (nt) of GenBank (all entries), GenBank genomes (only genome sequencing efforts), and SRA archive (all entries) and the number (#) of bacterial genomes (https://gold.jgi.doe.gov/) released in the same timeline

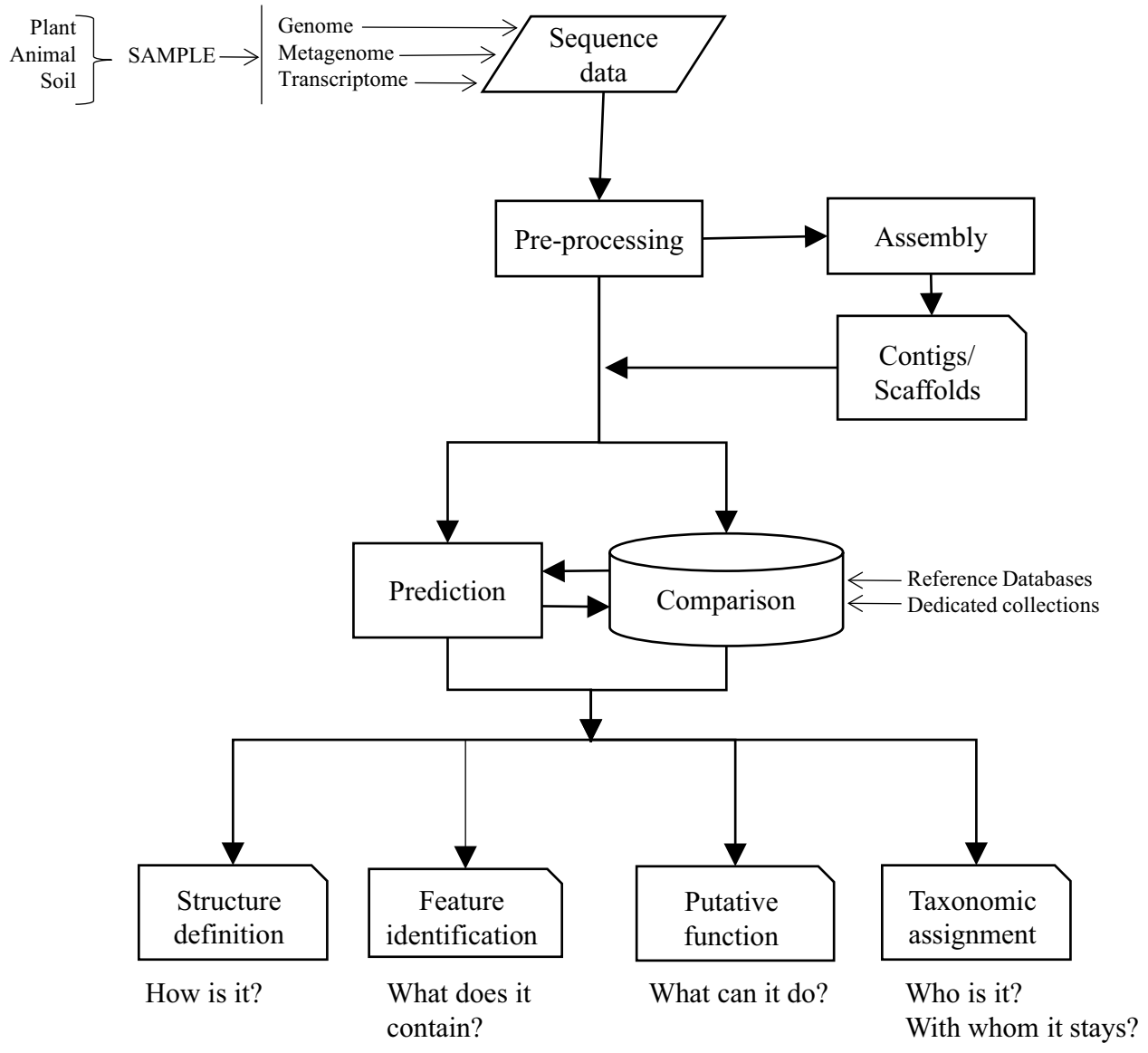


Fig. 2 General description of a standard workflow in omics data analyses. Starting from the biological sample of interest and sequencing by different approaches data can be processed by assembly into contigs/scaffolds or directly undergo prediction analyses and/or comparison with reference databases. The analyses support different kinds of investigations such as those required for (i) structure definition, (ii) feature identification (e.g., genes, regulatory and repeat regions), (iii) putative functionality, and (iv) taxonomic assignment

Table 2 Most used open-source software and reference databases in genomic, transcriptomic, and metagenomic studies

Category	Task	Name	Aims and Scope	Usage	Reference
Software and pipelines	Reads pre-processing	FastQC	Quality check and report of NGS data	GM	http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
		cutadapt	Adapter trimming algorithm	GM	[95]
		FASTX-toolkit	Toolset for manipulation of sequence data and format conversion	GM	http://hannonlab.cshl.edu/fastx_toolkit/index.html/
	Assembly	(META) VELVET/OASES	De novo genomic/transcriptomic assembly based on the de Brujin graph	GM	[96, 97]
		SOAP DE NOVO	De novo short-read assembler based on the de Brujin graph	G	[98]
		TRINITY	De novo assembly of RNA-seq data	G	[99]
	Gene prediction/annotation	Ensembl genome annotation	Gene annotation pipeline	G	http://www.ensembl.org/info/genome/genebuild/genome_annotation.html/
		Infernal	RNA secondary structure prediction based on reference multiple sequence alignments	G	[100]
		(Meta) Genemark	Gene prediction with unsupervised and semi-supervised training	GM	[101]
		(Meta) Genomethreader	Gene prediction by similarity with cDNA/EST and/or protein sequences	GM	[102]
		NCBI genome annotation	Genome annotation pipeline released by NCBI	G	http://www.ncbi.nlm.nih.gov/books/NBK169439/
		tRNAscan-SE	tRNA gene prediction	G	[103]
		Repeat masker	Similarity-based detection of DNA interspersed repeats and low complexity sequences	G	http://www.repeatmasker.org/
	Mapping	Star	RNA-seq to genome aligner	G	[104]
		Tophat/cufflinks	RNA-seq to genome aligner and quantification tools	G	[105]
	Marker-based metagenome	Mothur	Tools and software for 16S data clustering, classification, and ecological inference	M	[106]
		Qiime	Customizable pipeline for marker-gene-based metagenomics	M	[107]
		RDPipeline	RDP-based web interface for bacterial and fungal ribosomal marker gene analysis	M	[108]
	Mixed	Galaxy	Web-based platform of general purposes	GM	[109]
		transPLANT	e-infrastructure for exploring genomic data from crop and model plants	G	http://www.transplantdb.eu/

Table 2 continued

Category	Task	Name	Aims and Scope	Usage	Reference
	Shotgun metagenome	Megan	Stand-alone blast output parser and mining tool for phylogenetic and functional assignment based on the lowest common ancestor algorithm	M	[110]
		Metamos	Customizable pipeline for shotgun data assembly and analysis	M	[111]
		(Mg-)Rast	Fully automated online server for analyses of shotgun data	G[a]M	[112]
	Population genomic	Metabel	Software for meta-analysis of genome-wide SNP association	G	[113]
		Metal	Tool for mining variation data and perform association studies	G	[114]
		Plink	Tools for managing genomic variation data	GM	[115]
		SVS	Genomic and phenotypic data analysis and visualization	G	http://www.goldenhelix.com
		Tassel	Tools and pipelines for genome variation studies	G	[116]
		VcfTools	Tools for genome comparisons and mining plant variation data	GM	[117]
Reference Databases	General	Genomes online database	Metadata repository for genome and metagenome sequencing projects	GM	https://gold.jgi.doe.gov/
		JGI Phytozome	Plant Comparative Genomics at the Joint Genome Institute	G	http://phytozome.jgi.doe.gov/pz/portal.html
		INSDC	DDBJ, EMBL-EBI, and NCBI, common repository	GM	http://www.insdc.org/
		PLANTGDB	Unified plant genomic database	G	http://pgdbj.jp/
	Taxonomic annotation	RDP/Silva/Greengenes	Repositories of ribosomal RNA genes	GM	[118–120]
	Functional annotation	KEGG	Integrated resources for functional annotation of genes	GM	[121]
		COG	Clusters of ortholog groups	GM	[122]
		SEED	Integrated resources for functional annotated microbial genes	G[a]M	[123]
		RFAM	RNA families collection	G	[124]
		DFAM	Repetitive DNA elements collection	G	[125]
		UNIPROT	Database of functional annotated protein sequences	G	http://www.uniprot.org/

G use in genomics and transcriptomics, *M* use in metagenomics

[a] Dedicated to microbial genomes

possible indication of similarity in role. Comparison with reference (nucleotide or amino acid general databases) or more specific collections, such as those from genomes, gene families, transcriptomes, or repeats, is fundamental to transfer information from already annotated molecules to newly defined ones. Beyond supporting structure and functional assignments by detection of common features, similarity searches also support identification of peculiarities and provide hints for evolutionary investigations [78].

The availability of an increasing number of reference genomes associated with a decreasing sequencing cost per base enabled also the analysis of genome variations based on single nucleotide polymorphisms (SNP) discovery. These studies intend to identify the variability between genomic sequences from individual genomes. By comparing the sequenced genomes, a catalog of mutations from individuals is obtained, usually defined as SNPs and/or insertion-deletion (INDELs), but also as larger rearrangements (e.g., copy number (CNV) and presence absence (PAV) variations, translocations). These features can be associated with specific phenotypes of interest. Currently, millions of polymorphisms have been discovered in plants, such as *A. thaliana* [79], rice [80, 81], soybean [82], tomato [73, 83], maize [84, 85], and in animals [86]. These resources are essential in breeding challenges for species of agriculture interest. Indeed, the possibility to exploit data from larger collections of individuals strongly increases the potential to identify more alleles useful for improved a sustainable productions, providing solutions for growing demand for better food, in a climate changing world.

As introduced, metagenomics approaches aim at the identification of the species in a sample and to the definition of their relationships. While bioinformatics approaches for what concerns pre-processing and assembly steps are similar (Fig. 2), downstream analyses depend on the peculiarities of the implemented strategy. The taxonomic composition of the microbiome can be profiled using the marker-based approach, i.e., a PCR-amplification with universal primers of a taxon-specific gene, followed by the extensive sequencing of the amplicon by the preferred platform (Table 1). Sequences derived from such studies are usually compared with dedicated databases representing high-quality full-length reference tags. For example, in the case of bacteria the choice falls mainly on the *16S* gene from the ribosomal operon and the most used reference databases for comparisons and identification are listed in Table 2. Pipelines have been also implemented to aid non-experts users in a correct parsing of the metagenome-derived NGS data (Table 2). Phylogenetic relationships, obtained by sequence similarity, can be used for ecological inference using dedicated pipelines [87]. However, although widespread, the marker-based approach falls short in predicting the functionality and the activity of the microbial community. Indeed, the methodology suffers the typical PCR biases, such as (i) the misincorporation of nucleotides (which would lead to the overestimation of sequence diversity); (ii) the differential amplification of the same gene from different organisms (true for example in the case of *16S* genes whose number of copies in the genome varies among taxa [88]); and (iii) the formation of chimeric artifacts. Moreover, markers can have limits in taxonomical assignments, mainly because of lack of consistent genome information from all possible species, affecting the specificity of the identification of the components [43]. This leads to the wide use of the operative taxonomic units (OTUs) for distinguishing all the different components in a sample, since they represent groups of highly similar sequences [89]. The "shotgun" approach is alternative to the marker-based one. It consists in the high-throughput sequencing of a pool of DNA fragments that may encompass various genomic loci from all taxa represented in the sample (prokaryotic, eukaryotic, and viral genomes). Unlike target-based approaches, the shotgun technique provides more details on the genomic structure of the community, offering a wider description of its potential functionality [13, 14, 50, 55]. Data from whole metagenome shotgun consist of short DNA reads that can be assembled to obtain coding sequences or genomic contigs. Coding sequences can be identified through the comparison with specific databases (Table 2). The assembly should be carefully evaluated because most of the assemblers were developed for genome assembly and are not designed to deal with the heterogeneity of metagenomic datasets. In alternative, raw reads can be also used for direct assignment and annotation, though their short length may limit the exhaustivity of the results. Ultimately, other limits of the shotgun method are (i) the initial amount of extracted DNA for library production should be rather high (>10 ng); and (ii) in case of large and complex communities, or communities where one or few species dominate over the others, the coverage of the entire components may be limited. Indeed, the likelihood that the species poorly represented will be covered by sufficient reads that will also permit the assembly of representative contigs is rather low. Examples of main software dedicated to metagenome analyses are reported in Table 2. Ultimately, other limits of the shotgun method are (i) the initial amount of extracted DNA for library production is rather high (>10 ng); and (ii) in case of large and complex communities, or communities where one or few species dominate over the others, the likelihood that more than one read will cover a single gene is rather low; therefore, little information will be obtained about the species with

low abundance, and assembly would probably result in short contigs (if any). As introduced, reference data collections are fundamental since the beginning of bioinformatics. Data sharing, by general reference or specialized databases, is precious to the majority of the bioinformatics strategies here presented. They not only support fundamental analyses for straightforward characterization of the investigated molecules, but are also essential to offer results to the whole scientific community. To this aim, the effort of setting up comprehensive collections, suitably representing the metadata derived by their mining and by the integration of different resources, demands major efforts in the NGS bioinformatics era. Indeed, NGS technologies attracted an unexpected interest from the scientific world for their accessibility and their resolution power, further challenging the stabilization of resources and data integration. As an example, the fast sequencing of complete genomes from different species, such as those from different genotypes or cultivars, faces the bottleneck caused by the need of suitable data analyses and curation. On the other hand, the fast production of collections from parallel efforts makes the update of novel release hard, though the presence of reference databases, favoring the flourishing of community specific collections, often misaligned with reference ones, affecting also the quality of the results.

Conclusions

Bioinformatics is the exclusive approach capable of exploiting and sharing the large amount of omics data the different technologies may provide. Suitable computational methods and appropriate resources are fundamental for detecting value added biological information providing novel insights into the organization of biological systems. The identification of structure and functional properties of the molecular data in a specific process allows the in-depth understanding of systems organization and behavior, supporting the design of reliable and representative models and paving the way to the comprehension of emergent properties that only holistic approaches can offer.

However, despite the introduction of highly processive experimental technologies and of innovative computational approaches in support of the molecular characterizations, only 10 % of the genome organization and associated functionalities have been today understood and an even lower percentage of metagenomics datasets have been confidently annotated [90]. This confirms that though at quite 70 years from the discovery of the DNA structure and the beginning of bioinformatics, we are still at the very early stage of the genomics era, and surely quite far from achieving the ambitious goal of the in silico simulation of complex living organisms as well as ecosystem relationships. Indeed, these efforts still demand for extensive and suitable studies of genomes, transcriptomes, and metagenome data for proper links with sample organization and functionality, considering single species analyses and community approaches.

Despite these limits, the NGS bioinformatics era is revolutionizing the experimental design in molecular biology, strikingly contributing in increasing scientific knowledge while impacting relevant applications in many different aspects of agriculture. Data from disparate research fields, such as breeding, microbiology, and environmental sciences, are favoring a common exploitation and advances in molecular knowledge from massive efforts, with bioinformatics as driving methodology for its power and multifaceted capabilities. Organizing, detecting, integrating data information content, and data sharing are contributing to multidisciplinary interactions, expanding resources and spreading common methodologies. This revolutionize agriculture practice and production, offering knowledge and tools for improved product quality and ameliorated strategies of protection against environmental stress, diseases, and parasites [40, 91]. The different applications here overviewed, beyond providing relevant scientific knowledge based on their specificities, are also fundamental for translational approaches providing contributions with technological innovation, novel products, predictive and monitoring approaches [92], also supporting innovative applications for crop and livestock management [93, 94].

The increase of omics-based studies needs education in the associated technologies and in bioinformatics for appropriate experimental design and analyses, and for properly conveying experimental and computational efforts towards an in-depth knowledge and appropriate modeling of the biological systems [126].

Authors' contributions

AE introduced the metagenomics section and contributed to the organization of the manuscript; CC and VR introduced the genomics and transcriptomics topics; MLC planned, organized and supervised the entire effort. All authors contributed to the writing of the manuscript. All authors read and approved the final version.

Acknowledgements

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

Competing interests

The authors declare that they have no competing interests.

Received: 26 November 2015 Accepted: 23 February 2016
Published online: 02 April 2016

References

1. Dayhoff MO. Atlas of protein sequence and structure. Silver Spring: National Biomedical Research Foundation; 1965.
2. Chiusano ML, D'Agostino N, Traini A, Licciardello C, Raimondo E, Aversano M, Frusciante L, Monti L. ISOL@: an Italian SOLAnaceae genomics resource. BMC Bioinform. 2008;9(Suppl 2):S7.
3. Bostan H, Chiusano ML. NexGenEx-Tom: a gene expression platform to investigate the functionalities of the tomato genome. BMC Plant Biol. 2015;15(1):48.
4. Barh D, Zambare V, Azevedo V, Omics. Applications in biomedical, agricultural, and environmental sciences. Boca Raton: CRC Press. 2013.
5. Grice EA, Segre JA. The human microbiome: our second genome. Annu Rev Genomics Hum Genet. 2012;13:151.
6. Knief C. Analysis of plant microbe interactions in the era of next generation sequencing technologies. Front Plant Sci. 2014;5.
7. Wang J, McLenachan PA, Biggs PJ, Winder LH, Schoenfeld BI, Narayan VV, Phiri BJ, Lockhart PJ. Environmental bio-monitoring with high-throughput sequencing. Brief Bioinform. 2013;14(5):575–88.
8. Van Emon J. Omics revolution in agricultural research. J Agri Food Chem. 2015.
9. Zhu B, Pennack JA, McQuilton P, Forero MG, Mizuguchi K, Sutcliffe B, Gu CJ, Fenton JC, Hidalgo A. Drosophila neurotrophins reveal a common mechanism for nervous system formation. PLoS Biol. 2008;6(11):e284.
10. Semagn K, Bjørnstad Å, Xu Y. The genetic dissection of quantitative traits in crops. Electron J Biotechnol. 2010;13(5):16–7.
11. Ekblom R, Wolf JBW. A field guide to whole-genome sequencing, assembly and annotation. Evol Appl. 2014;7(9):1026–42.
12. Handelsman J, Rondon MR, Brady SF, Clardy J, Goodman RM. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. Chem Biol. 1998;5(10):R245–9.
13. Howe AC, Jansson JK, Malfatti SA, Tringe SG, Tiedje JM, Brown CT. Tackling soil diversity with the assembly of large, complex metagenomes. Proc Natl Acad Sci. 2014;111(13):4904–9.
14. Kodzius R, Gojobori T. Marine metagenomics as a source for bioprospecting. Mar Genom. 2015.
15. Esposito A, Ahmed E, Ciccazzo S, Sikorski J, Overmann J, Holmström SJ, Brusetti L. Comparison of rock varnish bacterial communities with surrounding non-varnished rock surfaces: taxon-specific analysis and morphological description. Microb Ecol 2015; 1–10.
16. Consortium HMP. Structure, function and diversity of the healthy human microbiome. Nature. 2012;486(7402):207–14.
17. Ofek-Lalzar M, Sela N, Goldman-Voronov M, Green SJ, Hadar Y, Minz D. Niche and host-associated functional signatures of the root surface microbiome. Nat Commun. 2014; 5.
18. Epp LS, Boessenkool S, Bellemain EP, Haile J, Esposito A, Riaz T, Erseus C, Gusarov VI, Edwards ME, Johnsen A. New environmental metabarcodes for analysing soil DNA: potential for studying past and present ecosystems. Mol Ecol. 2012;21(8):1821–33.
19. Benkeblia N. Sustainable agriculture and new biotechnologies. Boca Raton: CRC Press; 2011.
20. Siol M, Wright SI, Barrett SC. The population genomics of plant adaptation. New Phytol. 2010;188(2):313–32.
21. Barabaschi D, Tondelli A, Desiderio F, Volante A, Vaccino P, Valè G, Cattivelli L. Next generation breeding. Plant Sci. 2015.
22. Morrell PL, Buckler ES, Ross-Ibarra J. Crop genomics: advances and applications. Nat Rev Genet. 2012;13(2):85–96.
23. Ellegren H. Genome sequencing and population genomics in non-model organisms. Trends Ecol Evol. 2014;29(1):51–63.
24. Weigel D, Mott R. The 1001 genomes project for *Arabidopsis thaliana*. Genome Biol. 2009;10(5):107.
25. Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, Washietl S, Kheradpour P, Ernst J, Jordan G, Mauceli E, Ward LD, Lowe CB, Holloway AK, Clamp M, Gnerre S, Alfoldi J, Beal K, Chang J, Clawson H, Cuff J, Di Palma F, Fitzgerald S, Flicek P, Guttman M, Hubisz MJ, Jaffe DB, Jungreis I, Kent WJ, Kostka D, Lara M, et al. A high-resolution map of human evolutionary constraint using 29 mammals. Nature. 2011;478(7370):476–82.
26. Chiusano M, D'Agostino N, Barone A, Carputo D, Frusciante L. Genome analysis of species of agricultural interest, In: Advances in modeling agricultural systems. Berlin: Springer; 2009. p. 385–402.
27. Zhang Z, Ober U, Erbe M, Zhang H, Gao N, He J, Li J, Simianer H. Improving the accuracy of whole genome prediction for complex traits using the results of genome wide association studies. PLoS One. 2014;9(3):e93017.
28. Organization EPS. European plant science: a field of opportunities. J Exp Bot. 2005;56(417):1699–709.
29. Iovene M, Barone A, Frusciante L, Monti L, Carputo D. Selection for aneuploid potato hybrids combining a low wild genome content and resistance traits from Solanum commersonii. Theor Appl Genet. 2004;109(6):1139–46.
30. van der Vlugt R, Minafra A, Olmos A, Ravnikar M, Wetzel T, Varveri C, Massart S. Application of next generation sequencing for study and diagnosis of plant viral diseases in agriculture. 2015.
31. Van Borm S, Belák S, Freimanis G, Fusaro A, Granberg F, Höper D, King DP, Monne I, Orton R, Rosseel T. Next-generation sequencing in veterinary medicine: how can the massive amount of information arising from high-throughput technologies improve diagnosis, control, and management of infectious diseases? In: Veterinary infection biology: molecular diagnostics and high-throughput strategies. Berlin: Springer; 2015. p. 415–36.
32. Blanchfield J. Genetically modified food crops and their contribution to human nutrition and food quality. J Food Science. 2004, 69(1):CRH28-CRH30.
33. Yuan JS, Tiller KH, Al-Ahmad H, Stewart NR, Stewart CN Jr. Plants to power: bioenergy to fuel the future. Trends Plant Sci. 2008;13(8):421–9.
34. Ma JKC, Drake PMW, Christou P. The production of recombinant pharmaceutical proteins in plants. Nat Rev Genet. 2003;4(10):794–805.
35. Wilson SA, Roberts SC. Metabolic engineering approaches for production of biochemicals in food and medicinal plants. Curr Opin Biotechnol. 2014;26:174–82.
36. Timmusk S, El-Daim IAA, Copolovici L, Tanilas T, Kännaste A, Behers L, Nevo E, Seisenbaeva G, Stenström E, Niinemets Ü. Drought-tolerance of wheat improved by rhizosphere bacteria from harsh environments: enhanced biomass production and reduced emissions of stress volatiles. 2014.
37. Haney CH, Samuel BS, Bush J, Ausubel FM. Associations with rhizosphere bacteria can confer an adaptive advantage to plants. Nat Plants 2015.
38. Acosta-Martínez V, Cotton J, Gardner T, Moore-Kucera J, Zak J, Wester D, Cox S. Predominant bacterial and fungal assemblages in agricultural soils during a record drought/heat wave and linkages to enzyme activities of biogeochemical cycling. Applied Soil Ecology. 2014;84:69–82.
39. Babu AN, Jogaiah S. Ito S-i, Nagaraj AK, Tran L-SP. Improvement of growth, fruit weight and early blight disease protection of tomato plants by rhizosphere bacteria is correlated with their beneficial traits and induced biosynthesis of antioxidant peroxidase and polyphenol oxidase. Plant Sci. 2015;231:62–73.
40. Deusch S, Tilocca B, Camarinha-Silva A, Seifert J. News in livestock research—use of Omics-technologies to study the microbiota in the gastrointestinal tract of farm animals. Computational and Structural Biotechnology Journal. 2015;13:55–63.
41. Whitman WB, Coleman DC, Wiebe WJ. Prokaryotes: the unseen majority. Proc Natl Acad Sci. 1998;95(12):6578–83.
42. Staley JT, Konopka A. Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. Ann Rev Microbiol. 1985;39(1):321–46.
43. Rappé MS, Giovannoni SJ. The uncultured microbial majority. Ann Rev Microbiol. 2003;57(1):369–94.
44. Carbonetto B, Rascovan N, Álvarez R, Mentaberry A, Vázquez MP. Structure, composition and metagenomic profile of soil microbiomes associated to agricultural land use and tillage systems in Argentine Pampas. 2014.
45. Mendes LW, Kuramae EE, Navarrete AA, van Veen JA, Tsai SM. Taxonomical and functional microbial community selection in soybean rhizosphere. The ISME journal. 2014;8(8):1577–87.
46. Fouts DE, Szpakowski S, Purushe J, Torralba M, Waterman RC, MacNeil MD, Alexander LJ, Nelson KE. Next generation sequencing to define prokaryotic and fungal diversity in the bovine rumen. 2012.

47. Rastogi G, Coaker GL, Leveau JH. New insights into the structure and function of phyllosphere microbiota through high-throughput molecular approaches. FEMS Microbiol Lett. 2013;348(1):1–10.
48. Pan Y, Cassman N, de Hollander M, Mendes LW, Korevaar H, Geerts RH, van Veen JA, Kuramae EE. Impact of long-term N, P, K, and NPK fertilization on the composition and potential functions of the bacterial community in grassland soil. FEMS Microbiol Ecol. 2014;90(1):195–205.
49. Bevivino A, Paganin P, Bacci G, Florio A, Pellicer MS, Papaleo MC, Mengoni A, Ledda L, Fani R, Benedetti A. Soil Bacterial community response to differences in agricultural management along with seasonal changes in a mediterranean region. 2014.
50. Souza RC, Hungria M, Cantão ME, Vasconcelos ATR, Nogueira MA, Vicente VA. Metagenomic analysis reveals microbial functional redundancies and specificities in a soil under different tillage and crop-management regimes. Appl Soil Ecol. 2015;86:106–12.
51. Lavecchia A, Curci M, Jangid K, Whitman WB, Ricciuti P, Pascazio S, Crecchio C. Microbial 16S gene-based composition of a sorghum cropped rhizosphere soil under different fertilization managements. Biol Fertil Soils. 2015;51(6):661–72.
52. Pii Y, Borruso L, Brusetti L, Crecchio C, Cesco S, Mimmo T. The interaction between iron nutrition, plant species and soil type shapes the rhizosphere microbiome. Plant Physiol Biochem. 2016;99:39–48.
53. Stempfhuber B, Richter-Heitmann T, Regan KM, Kölbl A, Kaul P, Marhan S, Sikorski J, Overmann J, Friedrich MW, Kandeler E. Spatial interaction of archaeal ammonia-oxidizers and nitrite-oxidizing bacteria in an unfertilized grassland soil. Front Microbiol. 2015;6:1567.
54. Vayssier-Taussat M, Albina E, Citti C, Cosson J-F, Jacques M-A, Lebrun M-H, Le Loir Y, Ogliastro M, Petit M-A, Roumagnac P. Shifting the paradigm from pathogens to pathobiome: new concepts in the light of meta-omics. Front Cell Infect Microbiol. 2014; 4.
55. Handelsman J. Metagenomics: application of genomics to uncultured microorganisms. Microbiol Mol Biol Rev. 2004;68(4):669–85.
56. Urano K, Kurihara Y, Seki M, Shinozaki K. 'Omics' analyses of regulatory networks in plant abiotic stress responses. Curr Opin Plant Biol. 2010;13(2):132–8.
57. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Rapp BA, Wheeler DL. GenBank. Nucleic Acids Res. 2000;28(1):15–8.
58. Pruitt KD, Tatusova T, Maglott DR. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 2007;35(suppl 1):D61–5.
59. Hubbard T, Barker D, Birney E, Cameron G, Chen Y, Clark L, Cox T, Cuff J, Curwen V, Down T. The Ensembl genome database project. Nucleic Acids Res. 2002;30(1):38–41.
60. Dong Q, Schlueter SD, Brendel V. PlantGDB, plant genome database and analysis tools. Nucleic Acids Res. 2004;32(suppl 1):D354–9.
61. Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B, Moreno RF, et al. Complementary DNA sequencing: expressed sequence tags and human genome project. Science. 1991;252(5013):1651–6.
62. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. Serial analysis of gene expression. Science. 1995;270(5235):484–7.
63. Kodzius R, Kojima M, Nishiyori H, Nakamura M, Fukuda S, Tagami M, Sasaki D, Imamura K, Kai C, Harbers M. CAGE: cap analysis of gene expression. Nat Methods. 2006;3(3):211–22.
64. Brenner S, Johnson M, Bridgham J, Golda G, Lloyd DH, Johnson D, Luo S, McCurdy S, Foy M, Ewan M. Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. Nat Biotechnol. 2000;18(6):630–4.
65. Iquebal M, Jaiswal S, Mukhopadhyay C, Sarkar C, Rai A, Kumar D: Applications of Bioinformatics in Plant and Agriculture in *PlantOmics: The Omics of Plant Science*. 2015, Springer:755-89.
66. Boguski MS, Lowe TM, Tolstoshev CM. dbEST–database for "expressed sequence tags". Nat Genet. 1993;4(4):332–3.
67. Pontius JU, Wagner L, Schuler GD. 21. UniGene: A Unified View of the Transcriptome. The NCBI Handbook. Bethesda, MD: National Library of Medicine (US), NCBI 2003.
68. Pertea G, Huang X, Liang F, Antonescu V, Sultana R, Karamycheva S, Lee Y, White J, Cheung F, Parvizi B. TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. Bioinformatics. 2003;19(5):651–2.
69. D'Agostino N, Traini A, Frusciante L, Chiusano ML. SolEST database: a. BMC Plant Biol. 2009;9(1):142.
70. Mita K, Morimyo M, Okano K, Koike Y, Nohata J, Kawasaki H, Kadono-Okuda K, Yamamoto K, Suzuki MG, Shimada T. The construction of an EST database for Bombyx mori and its application. Proc Natl Acad Sci. 2003;100(24):14121–6.
71. Christoffels A, van Gelder A, Greyling G, Miller R, Hide T, Hide W. STACK: Sequence Tag Alignment and Consensus Knowledgebase. Nucleic Acids Res. 2001;29(1):234–8.
72. D'Agostino N, Aversano M, Chiusano ML. ParPEST: a pipeline for EST data analysis based on parallel computing. BMC Bioinform. 2005;6(Suppl 4):S9.
73. Aflitos S, Schijlen E, Jong H, Ridder D, Smit S, Finkers R, Wang J, Zhang G, Li N, Mao L. Exploring genetic variation in the tomato (Solanum section Lycopersicon) clade by whole-genome sequencing. Plant J. 2014;80(1):136–48.
74. Vogel TM, Simonet P, Jansson JK, Hirsch PR, Tiedje JM, van Elsas JD, Bailey MJ, Nalin R, Philippot L. TerraGenome: a consortium for the sequencing of a soil metagenome. Nat Rev Microbiol. 2009;7(4):252.
75. Gilbert JA, Meyer F, Antonopoulos D, Balaji P, Brown CT, Brown CT, Desai N, Eisen JA, Evers D, Field D. Meeting report: the terabase metagenomics workshop and the vision of an Earth microbiome project. Stand Genomic Sci. 2010;3(3):243.
76. Leinonen R, Sugawara H, Shumway M. The sequence read archive. Nucleic Acids Res. 2010;gkq1019.
77. Martin JA, Wang Z. Next-generation transcriptome assembly. Nat Rev Genet. 2011;12(10):671–82.
78. Mathé C, Sagot MF, Schiex T, Rouze P. Current methods of gene prediction, their strengths and weaknesses. Nucleic Acids Res. 2002;30(19):4103–17.
79. Gan X, Stegle O, Behr J, Steffen JG, Drewe P, Hildebrand KL, Lyngsoe R, Schultheiss SJ, Osborne EJ, Sreedharan VT. Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*. Nature. 2011;477(7365):419–23.
80. Xu X, Liu X, Ge S, Jensen JD, Hu F, Li X, Dong Y, Gutenkunst RN, Fang L, Huang L. Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. Nat Biotechnol. 2012;30(1):105–11.
81. Subbaiyan GK, Waters DL, Katiyar SK, Sadananda AR, Vaddadi S, Henry RJ. Genome-wide DNA polymorphisms in elite indica rice inbreds discovered by whole-genome sequencing. Plant Biotechnol J. 2012;10(6):623–34.
82. Lam H-M, Xu X, Liu X, Chen W, Yang G, Wong F-L, Li M-W, He W, Qin N, Wang B. Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. Nat Genet. 2010;42(12):1053–9.
83. Aflitos SA, Sanchez-Perez G, Ridder D, Fransz P, Schranz ME, Jong H, Peters SA. Introgression browser: high-throughput whole-genome SNP visualization. Plant J. 2015;82(1):174–82.
84. Lai J, Li R, Xu X, Jin W, Xu M, Zhao H, Xiang Z, Song W, Ying K, Zhang M. Genome-wide patterns of genetic variation among elite maize inbred lines. Nat Genet. 2010;42(11):1027–30.
85. Hufford MB, Xu X, Van Heerwaarden J, Pyhäjärvi T, Chia J-M, Cartwright RA, Elshire RJ, Glaubitz JC, Guill KE, Kaeppler SM. Comparative population genomics of maize domestication and improvement. Nat Genet. 2012;44(7):808–11.
86. Romiguier J, Gayral P, Ballenghien M, Bernard A, Cahais V, Chenuil A, Chiari Y, Dernat R, Duret L, Faivre N. Comparative population genomics in animals uncovers the determinants of genetic diversity. Nature. 2014;515(7526):261–3.
87. Lozupone C, Knight R. UniFrac: a new phylogenetic method for comparing microbial communities. Appl Environ Microbiol. 2005;71(12):8228–35.
88. Rastogi R, Wu M, DasGupta I, Fox GE. Visualization of ribosomal RNA operon copy number distribution. BMC Microbiol. 2009;9(1):208.
89. Schloss PD, Westcott SL. Assessing and improving methods used in operational taxonomic unit-based approaches for 16S rRNA gene sequence analysis. Appl Environ Microbiol. 2011;77(10):3219–26.
90. Gilbert JA, Field D, Swift P, Thomas S, Cummings D, Temperton B, Weynberg K, Huse S, Hughes M, Joint I. The taxonomic and functional

diversity of microbes at a temperate coastal site: a 'multi-omic' study of seasonal and diel temporal variation. PLoS ONE. 2010;5(11):e15545.
91. Coleman-Derr D, Tringe SG. Building the crops of tomorrow: advantages of symbiont-based approaches to improving abiotic stress tolerance. Front Microbiol. 2014;5(283):1–6.
92. Sloan SS, Lebeis SL. Exercising influence: distinct biotic interactions shape root microbiomes. Curr Opin Plant Biol. 2015;26:32–6.
93. Sachdev DP, Cameotra SS. Biosurfactants in agriculture. Appl Microbiol Biotechnol. 2013;97(3):1005–16.
94. Rolli E, Marasco R, Vigani G, Ettoumi B, Mapelli F, Deangelis ML, Gandolfi C, Casati E, Previtali F, Gerbino R. Improved plant resistance to drought is promoted by the root-associated microbiome as a water stress-dependent trait. Environ Microbiol. 2015;17(2):316–31.
95. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J. 2011;17(1):10–2.
96. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 2008;18(5):821–9.
97. Schulz MH, Zerbino DR, Vingron M, Birney E. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics. 2012;28(8):1086–92.
98. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience. 2012;1(1):18.
99. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q. Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. Nat Biotechnol. 2011;29(7):644.
100. Nawrocki EP, Kolbe DL, Eddy SR. Infernal 1.0: inference of RNA alignments. Bioinformatics. 2009;25(10):1335–7.
101. Lomsadze A, Ter-Hovhannisyan V, Chernoff YO, Borodovsky M. Gene identification in novel eukaryotic genomes by self-training algorithm. Nucleic Acids Res. 2005;33(20):6494–506.
102. Gremme G, Brendel V, Sparks ME, Kurtz S. Engineering a software tool for gene structure prediction in higher organisms. Inf Softw Technol. 2005;47(15):965–78.
103. Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 1997;25(5):0955–64.
104. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15–21.
105. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc. 2012;7(3):562–78.
106. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol. 2009;75(23):7537–41.
107. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Pena AG, Goodrich JK, Gordon JI. QIIME allows analysis of high-throughput community sequencing data. Nat Methods. 2010;7(5):335–6.
108. Cole JR, Wang Q, Cardenas E, Fish J, Chai B, Farris RJ, Kulam-Syed-Mohideen A, McGarrell DM, Marsh T, Garrity GM. The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. Nucleic Acids Res. 2009;37(suppl 1):D141–5.
109. Goecks J, Nekrutenko A, Taylor J. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol. 2010;11(8):R86.
110. Huson DH, Auch AF, Qi J, Schuster SC. MEGAN analysis of metagenomic data. Genome Res. 2007;17(3):377–86.
111. Treangen TJ, Koren S, Sommer DD, Liu B, Astrovskaya I, Ondov B, Darling AE, Phillippy AM, Pop M. MetAMOS: a modular and open source metagenomic assembly and analysis pipeline. Genome Biol. 2013;14(1):R2.
112. Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M, Paczian T, Rodriguez A, Stevens R, Wilke A. The metagenomics RAST server–a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinform. 2008;9(1):386.
113. Aulchenko YS, Ripke S, Isaacs A, Van Duijn CM. GenABEL: an R library for genome-wide association analysis. Bioinformatics. 2007;23(10):1294–6.
114. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics. 2010;26(17):2190–1.
115. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, De Bakker PI, Daly MJ. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81(3):559–75.
116. Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES. TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics. 2007;23(19):2633–5.
117. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST. The variant call format and VCFtools. Bioinformatics. 2011;27(15):2156–8.
118. Wang Q, Garrity GM, Tiedje JM, Cole JR. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl Environ Microbiol. 2007;73(16):5261–7.
119. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. 2012:gks1219.
120. McDonald D, Price MN, Goodrich J, Nawrocki EP, DeSantis TZ, Probst A, Andersen GL, Knight R, Hugenholtz P. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. ISME J. 2012;6(3):610–8.
121. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28(1):27–30.
122. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN. The COG database: an updated version includes eukaryotes. BMC Bioinform. 2003;4(1):41.
123. Overbeek R, Begley T, Butler RM, Choudhuri JV, Chuang HY, Cohoon M, de Crecy-Lagard V, Diaz N, Disz T, Edwards R, Fonstein M, Frank ED, Gerdes S, Glass EM, Goesmann A, Hanson A, Iwata-Reuyl D, Jensen R, Jamshidi N, Krause L, Kubal M, Larsen N, Linke B, McHardy AC, Meyer F, Neuweger H, Olsen G, Olson R, Osterman A, Portnoy V, et al. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Res. 2005;33(17):5691–702.
124. Burge SW, Daub J, Eberhardt R, Tate J, Barquist L, Nawrocki EP, Eddy SR, Gardner PP, Bateman A. Rfam 11.0: 10 years of RNA families. Nucleic Acids Res. 2012:gks1005.
125. Wheeler TJ, Clements J, Eddy SR, Hubley R, Jones TA, Jurka J, Smit AF, Finn RD. Dfam: a database of repetitive DNA based on profile hidden Markov models. Nucleic Acids. 2013;41:D70–82 **(Database issue)**.
126. Chiusano ML. On the multifaceted aspects of bioinformatics in the next generation era: the run that must keep the quality. J Next Gener Seq Appl 2015. doi:10.4172/jngsa.1000e106.