**Scholars Journal of Agriculture and Veterinary Sciences**

OPEN ACCESS

Abbreviated Key Title: Sch J Agric Vet Sci
ISSN 2348–8883 (Print) | ISSN 2348–1854 (Online)
Journal homepage: https://saspublishers.com/sjavs/

# New Genome Sequencing Techniques and Bioinformatics Applied to Genetic Improvement of Plants

Tiziana Maria Sirangelo*

Life Science Department, University of Modena and Reggio Emilia, Italy

**DOI:** 10.36347/sjavs.2020.v07i08.002 | **Received:** 02.08.2020 | **Accepted:** 09.08.2020 | **Published:** 13.08.2020

***Corresponding author:** Tiziana Maria Sirangelo

**Abstract** | **Review Article**

As genome sequencing of the main plant species was performed in the last decades, agriculture started undergoing some deep changes as well. Genomics provided new tools to investigate plants genotypes and their relationship with the phenotype. Traditional methods for genetic improvement of plant species are now supported by innovative technologies, based on NGS, allowing to simply sequence genomes with times and costs significantly lower than before. Through the use of these technologies, a very large number of molecular markers, such as SSRs and SNPs, became available, generating new polymorphisms databases that made it possible to improve the genetic selection of plant species. This work, after showing a brief state of the art about genome sequencing of the main cultivated plants, discusses how their genetic improvement has changed after the introduction of the new NGS technologies. In fact, they couple very relevant molecular techniques to the traditional phenotypic selection, both in terms of markers for assisted or genomic selection and of information about genes functions influencing plants features. Undoubtedly, genomics will have a more and more relevant role in plant genetic improvement, by changing the traditional breeding to a Next Generation Breeding.

**Keywords**: Next Generation Sequencing, Bioinformatics, Plants genetic improvement, Molecular markers, Genomics.



## INTRODUCTION

Traditional methods for the genetic improvement of plant species are now being supported by innovative technologies potentially improving their efficiency and creating new characters that weren't originally present in the germplasm. Having a brief look of the biotechnologies history applied to genetic improvement of plants, it is possible to observe that since the '60s mutagenesis was widely used to generate new genetic diversity by culture in vitro, representing a starting point for the genetic transformation that took place in more recent times [1] The availability of molecular markers allowed to identify the relationship between some of them and the most relevant characters (MAS – Marked Assisted Selection) [2]. Before the introductions of molecular techniques, markers which are already evident in the early stages of the organism development and relatively simple to detect were used (e.g. the colour of the seeds was associated with their size). Different markers are currently utilized, identifiable by various techniques, including those that allow recognizing specific proteins expressed in plants, phenotypic traits and above all genetic sequences. The latter are the most used, so much that the term MAS is mainly used to indicate this type of selection.

However, MAS is only suitable for traits controlled by a small number of genes in spite of the fact that it allowed to obtain excellent results in many cases, such as highlighting the mechanisms behind disease resistance. In other cases, such as those related to drought tolerance, this method showed some limits. Therefore, it was necessary to have genome-wide markers. Genomic selection (GS) [3] is an upgraded form of MAS, a marker-assisted selection in which genetic markers cover the whole genome. If applied to plant genetic improvement the technique seems to outweigh the MAS limitations [4]. It is based on the simultaneous estimation of effects on phenotypes resulting from all loci, haploid genotypes and genetic markers. By using genome-wide markers, the method computes a genomic estimated breeding value (GEBV), to obtain a more comprehensive and reliable selection [5]. However, this strategy requires the availability of phenotypic and genotypic data sets related to the reference population, and it works by estimating some parameters for the created model to explain the differences at phenotype level.

In the light of the above, it is clear how important it is to have the genome sequences of plants and how genomics has changed the nature of agriculture technologies and in particular the plants genetic improvement methods.

**Sequencing of cultivated species genomes**

The knowledge of plants genomes is strictly linked to the advances of DNA sequencing technologies. The high throughput of modern sequencers, based on Next Generation Sequencing (NGS) methods, allowed to simply obtaining a lot of plant genomes data more cheaply than before. Genomics provided new tools and techniques making it possible to investigate the whole genome and facilitating the study of the genotype and its relationship with the phenotype [6]. A brief summary of the sequencing of some of the most important crops is shown below (Fig. 1).

***Arabidopsis thaliana***

The *Arabidopsis thaliana* genome was sequenced in 2000 [7]. The genome is relatively short, being organized in just five chromosomes and showing a total size of approximately 135-megabases. Therefore, obtaining good quality sequences made it possible to improve methods for more comprehensive analyses in all eukaryotes, detecting large sets of plant-specific gene functions and providing insights for crop improvement.

***Rice***

Rice was involved in the first whole genome sequencing project of a cereal crop. Within the Rice Genome Research Program (Japan) the rice genome was well mapped and characterized, being the smallest of the major cereal crop genomes (400 to 430 Mb), but still about 3.5 times the size of the *Arabidopsis* genome. Genomic libraries were constructed in bacterial artificial chromosomes (BACs) or P1-derived artificial chromosomes (PACs) and a shotgun approach to sequence these clones was adopted [8].

***Maize***

Maize (or *corn*) is an important experimental model plant. The observed repetitive elements prevalence in its genome (66%) was slightly higher than shown in previous genome studies (58%–63%), such as rice and Arabidopsis. The results showed that the increase in size of the genome maize depends on the number of both repetitive elements and genes [9]. Analyses based on fully sequenced BACs allowed studying full-length repeats.

***Populus***

The Populus genome is one of the major plant model systems. *Populus trichocarpa* was the first tree with a whole-genome assembly. The analysis of this genome revealed that about 8000 pairs of duplicated genes survived in the Populus genome [10].

***Vitis vinifera***

A high-quality draft of the genome sequence of grapevine (*Vitis vinifera*) was obtained from a highly homozygous genotype [11]. It was revealed that over 40 percent of the genome is composed of repetitive/transposable elements, a slightly higher proportion than that individuated in rice. The grapevine genome was sequenced through a whole-genome shotgun (WGS) strategy, by using Sanger technology on ABI 3730xl sequencers.

***Apple tree***

The apple tree genome sequencing project started in 2007 and a WGS strategy was adopted. 13 billion sequenced nucleotides were generated, through the Sanger Method and the Roche/454 Genome Sequencers [12]. These studies significantly increased the knowledge and understanding of apple tree as a fruit crop. Identification of species specific genes provided relevant clues in order to improve fruit quality and resistance to diseases [13].

***Soybean***

Soybean (*Glycine max*) is one of the most important crop plants, due to its unique nutritional profile. Its genome was sequenced through a whole-genome shotgun approach. This made it possible to predict 46,430 protein-coding genes, 70% more than Arabidopsis. The genome sequence facilitated further studies involving the identification of species specific gene traits and consequently the creation of improved soybean varieties [14].

***Tomato***

The genome of tomato (*Solanum lycopersicum L.*) [15] was sequenced and assembled by the International Tomato Genome Sequencing Consortium [16]. The Tomato Genome Sequencing Project relied on an ordered BAC approach to generate high-quality sequences, being a reference for the Solanaceae family. Whole genome shotgun sequences (Roche 454) with Sanger sequence data from BAC-ends, additional data from Solexa and SOLiD technologies were involved.

A high-quality genome sequencing of the domesticated tomato variety (*Solanum pimpinellifolium*) was carried out at 2012 [17]. It was de novo assembled using Illumina short reads, and a 739 Mb draft genome was obtained. Divergence between the wild and domesticated tomato genomes is about 0.6%, corresponding to 5.4 million single nucleotide polymorphisms (SNPs) distributed along the chromosomes.

***Potato***

The genome of another species belonging to the same family was analysed, the potato (*Solanum tuberosum L.*), and was published in 2011 [18]. It was sequenced by using a WGS approach and relying on

two NGS platforms, Illumina Genome Analyser and Roche Pyrosequencing, as well as Sanger technologies.

***Strawberry***

The wild strawberry genome was first sequenced [19] and the cultivated strawberry one was only obtained in more recent times [20]. The genome was sequenced through a combination of short- and long-read methods, including Illumina, 10X Genomics, and PacBio technologies. Genes involved in metabolic pathways of aromatic compounds and other secondary metabolites that characterize this species were proved to be particularly interesting. They are subject to genetic improvement in fruit plants belonging to the same family, just like apple and peach.

***Peach***

Peach, a diploid *Prunus* species, is one of the best genetically characterized cultivated plants. A complete genome assembly using Sanger WGS methods was obtained [21]. The gene density in peach genome (1.22 genes per 10 kb on average) resulted higher than that in apple (0.78 genes) but was lower than in Arabidopsis (2.29 genes). The study about peach genome, as well as other researches about *Rosaceae* clade, which compared peach DNA sequence and genomes of apple [12] and strawberry [19] provided the basis for a further understanding the peach complete genome sequence organization and the genetic changes of this fruit and its family.

***Sunflower***

The domesticated sunflower, *Helianthus annuus L.*, is an oil crop which is well adapted to climate change. A high-quality reference for the sunflower genome is reported in a study [22]. The genome, analysed through the PacBio platform, represents an important point of reference for research activities aimed to exploit genetic diversity for improving biotic and abiotic stress resistance as well as the quality of the produced oil.

***Olive tree***

Despite the economic and ecological importance of olive tree, its genome has been little characterized compared to other fruit trees. In fact, the mechanisms behind the transmission of most genes associated its growth and quality traits are still poorly understood. One of the biggest technical challenges in sequencing eukaryotic genomes is undoubtedly DNA repetitiveness [23]. The olive genome (cultivar Leccino) was sequenced applying a combination of different NGS technologies [24]. Particularly a WGS approach was adopted by using the Illumina and 454 platforms. Due to the relatively low genome coverage of the sequencing, most of the obtained contigs do not represent specific genomic loci. However, the work is still helpful to localize genes involved in agronomic traits and for MAS.

An overall list of the genomes and transcriptomes that have been sequenced is available by accessing the Gene Index Project (http://compbio.dfci.harvard.edu/tgi/plant.html) or in the NCBI Unigene database (http://www.ncbi.nlm.nih.gov/unigene).

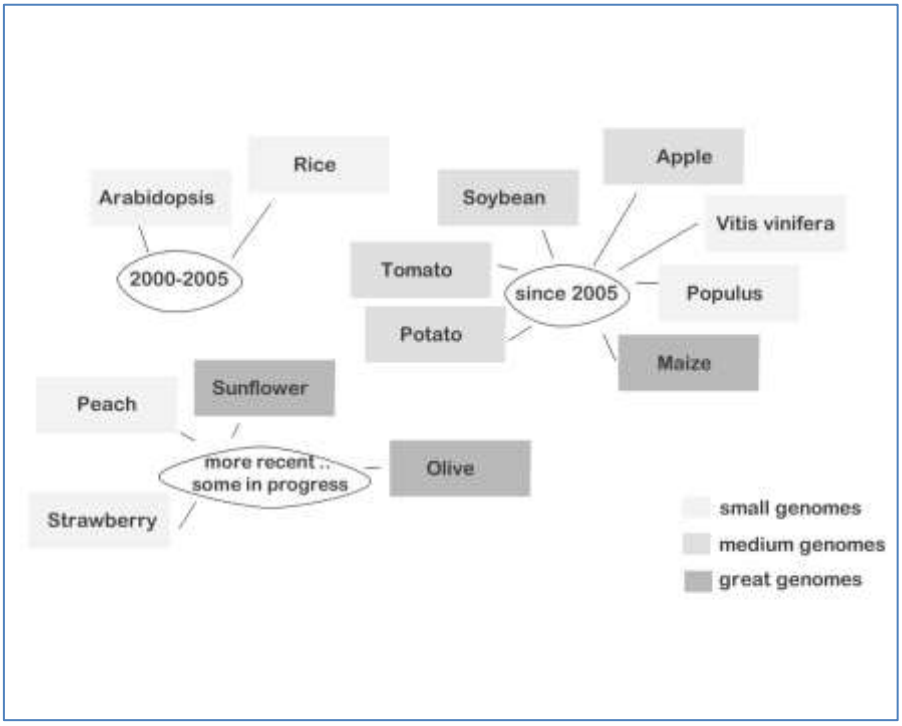


**Fig-1: The temporal order of some cultivated plants genomes sequencing**

**Changes in genetic improvement resulting from the use of the new NGS technologies**

Accurate plant genome knowledge allows to completely describing plant genetic diversity, to detail the phylogenetic relationships between species and to understand their evolutionary dynamics.

Sequencing allows defining plants genomes size, the organization and redundancy of the expressed sequences, clarifying in many cases the role of repeated elements, mainly transposons and retrotransposons that are very frequent in cultivated plants. At the same time, it makes it possible to understand the genetic basis of plants domestication, to fully describe the alleles of a given locus and to identify insertions/deletions and mutations [25].

Completely sequenced collections of plants genomes allow studying the biodiversity not just by analysing the phenotype. In fact, when genes related to given characters are known, it is possible to find their allelic variants and to estimate the impact of the most significant ones on the phenotype, as well as to individuate specific markers to be used in assisted selection. For instance, genes associated with resistance to pathogens were detected in many plant species, allowing their genetic improvement [26].

Studies based on the comparison among genomes also allow to analyse in depth cases of polyploidy in plants and to establish that the size of genomes of cultivated plants depends on the number of repeated elements and on ploidy level [27].

By using NGS technologies several thousands of molecular markers, such as SSRs (Simple Sequence Repeats) and SNPs (Single Nucleotide Polymorphisms), may be identified in a short time and at a little cost [28]. These high-throughput methods, not requiring a previous knowledge of polymorphisms, are also used to analyse species for which no reference genome is available. The availability of a very large number of molecular markers and the ability to search for associations between these markers and the loci responsible for plants characters generated a great amount of genetic data that, if properly used, allows improving the genetic selection of these species.

Resequencing techniques can also be applied to plants deriving from mutagenesis in order to identify mutations occurring within specific genes, and plants showing the desired mutations are made available to interested researchers for further selection [29].

By adopting the MAS approach it is possible to individuate a relatively small number of genes showing a relevant effect on a given phenotype and associated to a limited set of loci. GS also allows evaluating the total effect on this phenotype, caused by many more loci, each of them having a smaller impact on it. This is a crucial, because the success of many new varieties does not depend only on a few loci with a high phenotypic effect, but on the role of a well selected, specific loci combination. Markers distributed on the whole genome are currently used to calculate genomic indices and estimate the effect of each marker and of their combinations on a given phenotype. GS was adopted to analyse several characters of different plant species, such as barley and rice, and it was also proved valid in cases of complex genomes, such as maize [30].

Just one year ago, the genetic map of durum wheat was completed [31]. This result was the key to further investigate the drought-resistant wheat varieties, with higher yields and richer in nutrients.

Very recently, it has been demonstrated that it is not always true that among the varieties of the same plant species genetic differences are minimal. In fact, the analysis of available sequencing data among different cultivars of the same species revealed the existence of genomes characterized by a changing number of genes that differ from each other even for long DNA stretches [32]. This study demonstrated the incredible plasticity of the genome of cultivated plants in general and barley in particular, the species used as a model for the study of more complicated cereals such as wheat. It highlights how genetic diversity within the same plant species is not just the result of mutations in individual genes, but also of frequent deletion or duplication events. These results will provide new insights on genetic diversity and, consequently, on genetic improvement.

## CONCLUSIONS

Today, breeding plants still combine traditional methods with knowledge based on molecular markers to obtain quality improvement of the crop [33]. But the development of NGS technologies and the relatively low costs allowed DNA sequencing methods to become available for all researchers working on plants genetic improvement. Re-sequencing of genomes is very useful to discover markers resulting from high-throughput genotyping platforms, like SSRs and SNPs, and to generate high-density genetic maps.

Genome knowledge facilitates plant genetic improvement, adding to the traditional phenotypic selection very relevant molecular techniques, both in terms of markers for assisted selection or genomic selection and of information about genes functions influencing plants characteristics. NGS data and bioinformatics analyses allow to discover new genes, regulatory DNA sequences, and makes large collections of molecular markers available. Online resources also allow to access to data about allelic variability for genes affecting characters of agronomic interest.

The extensive use of molecular markers is now moving the selection from a phenotypic to a new genotype-based approach. Furthermore, the knowledge resulting from genome sequencing together with genome editing techniques will allow broadening the genetic diversity for selection purposes.

The analysis of plants genomes allows individuating DNA polymorphisms that can be used for authentication and for traceability at different levels of agriculture productions [34]. Genomics is changing agriculture also about a renewed sustainability at the service of farmers as well as of consumers. The environment monitoring based on DNA analysis is revolutionizing this field, with the possibility to describe agricultural biodiversity in a new innovative way [35].

In conclusion, the GS approach has great potential for genetic improvement in the plant field, even though it requires high bioinformatics skills in order to correctly interpret data coming from genome sequencing [36]. Genetic improvement is and will be an activity characterized by a highly technological content and genomic will have an increasingly relevant role in this area, by changing the traditional breeding to a new Next Generation Breeding [37].

## REFERENCES


1. Pérez-de-Castro AM, Vilanova S, Cañizares J, Pascual L, Blanca JM, Díez MJ, Prohens J and Picó B. Application of Genomic Tools in Plant Breeding. Curr Genomics. 2012; 13(3): 179–195.
2. Galbiati M, Gentile A, La Malfa S, Tonelli C. Sustainable biotechnology - Science and Innovation in agriculture to face the challenges of food safety and environmental sustainability - Edagricole - First Edition. 2017.
3. Goddard M. Genomic selection: prediction of accuracy and maximisation of long term response. Genetica. 2009; 136: 245-257.
4. Jannink JL, Lorenz AJ, Iwata H. Genomic selection in plant breeding: from theory to practice. Brief. Funct. Genomics. 2010; 9:166–177.
5. Wang X, Xu Y, Hu Z, Xu C. Genomic selection methods for crop improvement: Current status and prospects. The Crop Journal. 2018; 6(4): 330-340.
6. Tester M, Langridge P. Breeding technologies to increase crop production in a changing world. Science. 2010; 327:818–822.
7. The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature. 2000; 408: 796–815.
8. Eckardt NA. Sequencing the Rice Genome. Plant Cell. 2000; 12(11): 2011–2018.
9. Haberer G, Young S, Bharti AK, Gundlach H, Raymond C, Fuks G, Butler E, Wing RA, Rounsley S, Birren B, Nusbaum C. Structure and architecture of the maize genome. Plant physiology. 2005 Dec 1;139(4):1612-24.
10. Tuskan GA, DiFazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A, Schein J. The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). science. 2006 Sep 15;313(5793):1596-604.
11. Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C, Vezzi A. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. nature. 2007 Sep;449(7161):463.
12. Velasco R, Zharkikh A, Affourtit J, Dhingra A, Cestaro A, Kalyanaraman A, Fontana P, Bhatnagar SK, Troggio M, Pruss D, Salvi S. The genome of the domesticated apple (Malus× domestica Borkh.). Nature genetics. 2010 Oct;42(10):833-9.
13. Xu K. The Apple Genome: A Delicious Promise. New york fruit quarterly. 2010; 18(4).
14. Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J, Xu D. Genome sequence of the palaeopolyploid soybean. nature. 2010 Jan;463(7278):178-83.
15. Peralta IE, Spooner D M & Knapp S. Taxonomy of tomatoes: a revision of wild tomatoes (Solanum section Lycopersicon) and their outgroup relatives in sections Juglandifolia and Lycopersicoides. Syst. Bot. Monogr. 2008; 84: 1–186.
16. Mueller LA, Klein L. A Snapshot of the Emerging Tomato Genome Sequence. The Plant Genome. 2009; 2: 78-92.
17. Sato S, Tabata S, Hirakawa H. The tomato genome sequence provides insights into fleshy fruit evolution. Nature. 2012; 485: 635–641.
18. Xu X, Pan S, Cheng S. Genome sequence and analysis of the tuber crop potato. Nature. 2011; 475: 189–195.
19. Shulaev V, Sargent D, Crowhurst R. The genome of woodland strawberry (Fragaria vesca). Nat Genet. 2011; 43: 109–116.
20. Edger PP, Poorten TJ, VanBuren R. Origin and evolution of the octoploid strawberry genome. Nat Genet. 2019; 51:541–547.
21. Verde I, Abbott AG, Scalabrin S, Jung S, Shu S, Marroni F, Zhebentyayeva T, Dettori MT, Grimwood J, Cattonaro F, Zuccolo A. The high-quality draft genome of peach (Prunus persica) identifies unique patterns of genetic diversity, domestication and genome evolution. Nature genetics. 2013 May;45(5):487-94.
22. Badouin H, Gouzy J, Grassa CJ, Murat F, Staton SE, Cottret L, Lelandais-Brière C, Owens GL, Carrère S, Mayjonade B, Legrand L. The sunflower genome provides insights into oil metabolism, flowering and Asterid evolution. Nature. 2017 Jun;546(7656):148-52.
23. Alkan C, Coe BP & Eichler EE. Genome Structural Variation Discovery and Genotyping. Nature Reviews Genetic. 2011; 12(5): 363-376.
24. Muleo R, Morgante M, Velasco R, Cavallini A, Perrotta G and Baldoni L. Olive Tree Genomic.

Chapter 7 in Olive Germplasm – The Olive Cultivation, Table Olive and Olive Oil Industry in Italy. Intech. 2012.
25. Xu X, Liu X, Ge S, Jensen JD, Hu F, Li X, Dong Y, Gutenkunst RN, Fang L, Huang L, Li J. Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. Nature biotechnology. 2012 Jan;30(1):105-11.
26. Wang D, Guo C, Huang J, Yang S, Tian D, Zhang X. Allele-mining of rice blast resistance genes at AC134922 locus. Biochemical and Biophysical Research Communications. 2014; 446(4):1085-1090.
27. Wendel JF, Jackson SA, Meyers BC, Wing RA. Evolution of plant genome architecture. Genome biology. 2016 Dec 1;17(1):37.
28. Voss-Fels K, Snowdon RJ. Understanding and utilizing crop genome diversity via high-resolution genotyping. Plant Biotechnol. J. 2016; 14 1086–1094.
29. Krasileva KV, Vasquez-Gross HA, Howell T, Bailey P, Paraiso F, Clissold L, Simmonds J, Ramirez-Gonzalez RH, Wang X, Borrill P, Fosker C. Uncovering hidden variation in polyploid wheat. Proceedings of the National Academy of Sciences. 2017 Feb 7;114(6):E913-21.
30. Cattivelli L, Valè G. Next generation Breeding: genomic knowledge revolutionize the genetic improvement in Sustainable Biotechnology. Science and Innovation in agriculture to face the challenges of food safety and environmental sustainability - Edagricole - First Edition. 2017.
31. Maccaferri M, Harris NS, Twardziok SO, Pasam RK, Gundlach H, Spannagl M, Ormanbekova D, Lux T, Prade V, Milner S, Himmelbach A, Mascher M, Bagnaresi P, Faccioli P, Cozzi P, Lauria M, Lazzari B, Stella A, Manconi A, Gnocchi M, Moscatelli M, Avni R, Deek J, Biyiklioglu S, Frascaroli E. Durum wheat genome highlights past domestication signatures and future improvement targets. Nature Genetics. 2019; 51: 885–895.
32. Bretani G, Rossini L, Ferrandi C, Russell J, Waugh R, Kilian B, Bagnaresi P, Cattivelli L, Fricano A. Segmental duplications are hot spots of copy number variants affecting barley gene content. The Plant Journal. 2020.
33. Rasmussen SK. Molecular Genetics, Genomics, and Biotechnology in Crop Plant Breeding. Agronomy. Editorial. 2020.
34. Fontanesi L. Meat Authenticity and Traceability in Lawrie's Meat Science, 8th Edition, Editor: Toldrá F. – Woodhead Publishing, Elsevier, Oxford, UK. 2017. 585-633.
35. Utzeri VJ, Schiavo G, Ribani A, Tinarelli S, Bertolini F, Bovo S, Fontanesi L. Entomological signatures in honey: an environmental DNA metabarcoding approach can disclose information on plant-sucking insects in agricultural and forest landscapes. Scientific Reports. 2018; 8:9996.
36. Horner DS, Pavesi G, Castrignanò T, Meo PDOD, Liuni S, Sammeth M, Picardi E, Presole G. Bioinformatics approaches for genomics and post genomics applications of next-generation sequencing. Brief. Bioinform. 2009; 2:181–197
37. Barabaschi D, Tondelli A, Desiderio F, Volante A, Vaccino P, Valè G, Cattivelli L. Next generation breeding. Plant Sci. 2016.