Plant Science 242 (2016) 3–13

Contents lists available at ScienceDirect

Plant Science

journal homepage: www.elsevier.com/locate/plantsci

Next generation breeding

Delfina Barabaschi[a], Alessandro Tondelli[a], Francesca Desiderio[a], Andrea Volante[b], Patrizia Vaccino[c], Giampiero Valè[b], Luigi Cattivelli[a,*]

[a] Consiglio per la ricerca in agricoltura e l'analisi dell'economia agraria, Genomics Research Centre, Via San Protaso 302, 29017 Fiorenzuola d'Arda, Italy
[b] Consiglio per la ricerca in agricoltura e l'analisi dell'economia agraria, Rice Research Unit, SS 11 to Torino Km 2.5, 13100 Vercelli, Italy
[c] Consiglio per la ricerca in agricoltura e l'analisi dell'economia agraria, Research Unit for Cereal Selection in Continental areas, via R. Forlani, e, 26866 S. Angelo Lodigiano, Italy

ARTICLE INFO

Article history:
Received 17 April 2015
Received in revised form 10 July 2015
Accepted 11 July 2015
Available online 19 July 2015

Keywords:
Next generation sequencing
Molecular markers
Allele diversity
Genomic selection
Genome editing
Plant breeding

ABSTRACT

The genomic revolution of the past decade has greatly improved our understanding of the genetic make-up of living organisms. The sequencing of crop genomes has completely changed our vision and interpretation of genome organization and evolution. Re-sequencing allows the identification of an unlimited number of markers as well as the analysis of germplasm allelic diversity based on allele mining approaches. High throughput marker technologies coupled with advanced phenotyping platforms provide new opportunities for discovering marker-trait associations which can sustain genomic-assisted breeding. The availability of genome sequencing information is enabling genome editing (site-specific mutagenesis), to obtain gene sequences desired by breeders. This review illustrates how next generation sequencing-derived information can be used to tailor genomic tools for different breeders' needs to revolutionize crop improvement.

1. Introduction

The development of next generation sequencing (NGS) technologies has made DNA sequencing high throughput and very cost effective. Consequently, many opportunities are being opened to explore the relationships between genetic and phenotypic diversity with a resolution never reached before. Reference genome sequences have been published for many crop species [1] and many more genome sequencing projects are in progress (http://www.ncbi.nlm.nih.gov/genomes/leuks.cgi; http://plants.ensembl.org/index.html; http://phytozome.jgi.doe.gov/pz/portal.html). The sequences of crop genomes provide a useful starting point to explore genome organization and evolution and provide insight into genetic variation through partial or complete re-sequencing of different accessions [2]. Re-sequencing, leading to arrays of high-density single-nucleotide polymorphisms (SNPs), is allowing whole-genome scans to identify haplotype blocks that are significantly correlated with quantitative trait variation. The distribution of low cost sequencing technologies offers new opportunities to shape genetic diversity according to the needs of modern agriculture and, in turn, has a number of practical consequences for plant breeding: *i*) the analysis of genetic diversity can be based on genome re-sequencing; *ii*) genome wide association studies (GWAS) become an attractive approach for quantitative trait loci (QTLs) mapping in plants since broad genetic resources can be scanned for marker-trait association without any limitation of marker availability; *iii*) the great number of markers support genomic selection; and *iv*) the genome sequences allow the targeted modification of specific genes through genome editing technologies or identification of suitable mutations within mutagenized populations, resulting in the introduction of new allelic variants in the genome of cultivated varieties. Conversely, these achievements highlight new bottlenecks for breeding progress, particularly the phenotyping capacity (in terms of both precision and throughput [3]), and recombination frequency [4].

Over the last decades, plant breeding has moved from being a completely phenotyping-based process to having an increased reliance on some level of genotype-based selection [5]. This trend is

Abbreviations: CRISPR/Cas9, Clustered regularly interspaced short palindromic repeats/CRISPR-associated; GBS, genotyping by sequencing; GEBV, genomic estimated breeding values; GS, genomic selection; GWAS, genome-wide association studies; LD, linkage disequilibrium; MAGIC, multi-parent advanced generation intercross; MAS, marker-assisted selection; NAM, nested association mapping; NGS, next generation sequencing; RIL, recombinant inbreed line; QTL, quantitative trait locus; RAD, restriction-site associated DNA; SNP, single-nucleotide polymorphism; TALEN, transcription activator-like effector nucleases; TILLING, targeting induced local lesions in genomes; ZFN, zinc finger nucleases.

* Corresponding author.
E-mail address: luigi.cattivelli@entecra.it (L. Cattivelli).

http://dx.doi.org/10.1016/j.plantsci.2015.07.010
0168-9452/

expected to increase in the coming years as the NGS-based knowledge will be translated into "Next Generation Breeding". In this review, we consider current trends and future prospects for the application of genomic instruments in the improvement of plant breeding performance.

2. Genome sequencing and sequence-based markers

Molecular markers have been available for more than 25 years, nevertheless the advent of NGS represented a breakthrough in this field. Before NGS, a typical linkage map was based on few hundreds markers. In the age of NGS, thousands of markers can be easily included in any map, including in species with little a priori genome information available. With NGS technologies the DNA marker identification has shifted from fragment-based (RFLPs, AFLPs, microsatellites) to sequence-based polymorphisms (SNPs). Uniplex or multiplex SNP genotyping platforms that combine a variety of chemistries, detection methods, and reaction formats are available. Uniplex SNP genotyping platforms are more suitable for applications requiring small to moderate numbers of SNPs for a large number of samples. TaqMan™ (http://www.appliedbiosystems.com) and competitive allele-specific PCR (KASP™, http://www.lgcgenomics.com) are among the most popular techniques on the market. Multiplexed SNP analysis can be run on middle throughput platforms with a capacity of a few hundred SNPs per run (e.g. Illumina BeadXpress, Fluidigm EP1) or with high-throughput array-based technologies capable of generating between a few thousand to over one million SNPs per run (e.g. Illumina BeadArray™, Affymetrix GeneChip™ technology). The advent of high-density SNP arrays coupled with powerful computational pipelines has allowed the fast and easy scoring of large set of markers across many genotypes. Medium or high density arrays are available for many crop species, e.g. grapevine [6], maize [7], tomato [8], peach [9], soybean [10], barley [11], rice [12], wheat [13] and apple [14].

Nevertheless, the production of a high-quality array requires a substantial investment of resources, and the SNP panel, optimized for the population used to develop it, might be biased toward particular panels of germplasm. To circumvent these limitations, NGS technologies offer the possibility of shifting from array-based genotyping assays with pre-defined SNP panels to the direct sequencing of the populations of interest [15], producing a genome-wide and unbiased set of markers. These techniques employ a reduced genome representation achieved through restriction enzyme digestion and subsequent adaptor-mediated PCR amplification, and require no a priori knowledge of the SNPs being interrogated, making them useful for genetic analysis in species where no reference sequence is available. Among them, restriction-site associated DNA sequencing (RAD) [16] and genotyping by sequencing (GBS) [17] have been adopted in plants [18–22]. Furthermore, a strategy based on low coverage genome sequencing of all genotypes from a segregating population (POPSEQ) was recently employed for the development of high density genetic maps. POPSEQ was used to explore the organization of the gene space within the large, complex and highly repetitive barley genome [23], and contributed to the assembling of the hexaploid wheat genome [24].

3. Mining plant diversity: from genotype to phenotype

Many valuable genes and alleles are stored in seed bank collections, hidden in cultivars, landraces, mutagenized populations and wild species. The identification of these genes requires both genome information and phenotyping capacities. With the advent of NGS technologies a different dimension to the exploration of plant diversity arose. Extensive insights into plant genome composition and organization have been gained from the genome sequencing and new findings on plant origin and evolution (genome duplication, ancestral re-arrangements and polyploidization events) have been revealed [2]. An in silico paleogenomic study based on a deep comparison of monocot and eudicot genomes, allowed the reconstruction of ancestral protochromosome segments and a description of the evolutionary dynamics leading to the present-day genomes, their genome organization and regulation [25].

Since a single reference genome is not enough to represent the diversity within a species [26], the re-sequencing of different cultivars, landraces and wild accessions assumes an important role to reveal domestication events [27], identify gene diversification and variations [28] and explain heterosis mechanisms [29]. A most striking example is the "3,000 Rice Genomes Project", an initiative dedicated to the re-sequencing of 3000 rice accessions selected to represent the genetic and functional rice diversity available worldwide [30]. With re-sequencing information of many accessions, a strategy can be applied to search for allele diversity at candidate gene loci for which a clear association with specific phenotypic traits is known. This allele mining strategy can help trace the evolution of alleles, identify new useful haplotypes and guide the development of allele-specific markers for use in marker-assisted selection (MAS). For instance, following an allele mining approach several resistance gene homologues and functional resistance genes have been isolated in potato [31], wheat [32], rice [33,34], and barley [35].

When re-sequencing is applied to TILLING populations (TILLING-by-Sequencing), it allows a fast genome-wide identification of mutations [36,37]. Enhanced opportunities for functional genomics come from a genome-wide discovery pipeline of induced mutations based on multiplexed (10- to 30- fold) exome capture and sequencing. This method called MAPS (mutations and polymorphisms surveyor), was used to identify about 18,000 mutations in 72 independent M2 rice lines, of which >2,600 appeared to be detrimental to gene function [38].

Beside the description of allele diversity, genome re-sequencing coupled with de novo assembly of the sequences not matching the reference genome offers the possibility of harnessing the gene repertoire from wild relatives of crops leading to the description of their pan-genomes. Pan genome refers to the full complement of genes in a group of individuals (e.g. species) and consists of a core genome containing DNA sequences shared by all the genotypes and of a dispensable genome composed of partially shared genomic features (i.e. present in only some genotypes) [39]. For instance, the re-sequencing of seven accessions of *Glycine soja* led to the identification in the dispensable genome of many genes that had structural variant involved in the adaptation to the environment (R-genes, flowering time-related genes, genes involved in oil and fatty acid content), and which were therefore potentially useful in crop breeding [40].

Since the observed phenotypic diversity is not completely explained by variation at the genomic level, the establishment of the pan-genome should also be supported by the analysis of differences in gene expression and regulation. In maize the whole-seedling transcriptome variations among 503 diverse inbred lines have been described by RNA-sequencing. Besides the pan-genome, the analysis of the maize pan-transcriptome helps to explain some components of phenotypic variation (e.g. heterosis) in terms of variation in abundance of transcripts and their alternatively spliced forms in inbred lines [41]. In addition to gene regulation, epigenetic regulation represents another crucial factor to be considered for advanced crop breeding, especially for physiological phenotypes, having a fundamental role in regulating gene expression in response to developmental and environment changes [42].