frontiers in
PLANT SCIENCE

OPINION ARTICLE
published: 30 July 2014
doi: 10.3389/fpls.2014.00367

Next generation sequencing technologies for next generation plant breeding

Soham Ray*[1] *and Pratik Satya*[2]

[1] *Crop Improvement Division, Central Rice Research Institute, Cuttack, India*
[2] *Crop Improvement Division, Central Research Institute for Jute and Allied Fibres, Kolkata, India*
**Correspondence: pscrijaf@gmail.com*

Edited by:
Diego Rubiales, Consejo Superior de Investigaciones Cientificas, Spain

Reviewed by:
Anna Maria Mastrangelo, CRA-Centro di Ricerca per la Cerealicoltura, Italy
Noel Ferro Diaz, University of Bonn, Germany
Jaime Prohens, Universitat Politècnica de València, Spain

Keywords: next generation sequencing, plant breeding, single nucleotide polymorphism, genotyping, marker discovery

As a term, "next generation plant breeding" is increasingly becoming popular in crop breeding programmes, conferences, scientific fora and social media (Schnable, 2013). Being a frontier area of crop science and business, it is gaining considerable interest among scientific community and policymakers and funds flow from entrepreneurs and research funding agencies. Plant breeding is a continuous attempt to alter genetic architecture of crop plants for efficient utilization as food, fodder, fiber, fuel or other end uses. Although the scientific concepts in plant breeding originated about 100 years ago, domestication and selection of desirable plants from prehistoric periods have contributed tremendously to ensure human food security (Gepts, 2004). During the past few decades, well supported crop improvement programmes for major crops started reaping benefits from cutting edge technologies of biological sciences, particularly in the form of molecular markers and transgenic crop development, which in combination with conventional phenotype based selection, defines the current generation plant breeding practices. Different types of molecular markers have been developed and extensively used during the last three decades for identifying linkage between genes and markers, discovering quantitative trait loci (QTLs), pyramiding desired genes and performing marker assisted foreground and background selections for introgression of desired traits (Varshney and Tuberosa, 2007). However, these markers are based mostly on electrophoretic separation of DNA fragments, which limits its detection of genetic polymorphism. In large plant breeding populations, genotyping may take up several months depending on marker system, adding more cost to genotyping. The next generation plant breeding would thus demand more efficient technologies to develop low cost, high-throughput genotyping for screening large populations within a smaller time frame.

With the availability of whole genome sequences (WGS), the perspective of identification of DNA markers has shifted from fragment based polymorphism identification to sequence based single nucleotide polymorphism (SNP) identification to expedite the marker identification process and to increase the number of informative markers. But the WGS technologies based on Sanger sequencing are time consuming, costly and provide information only on the target individual, which have limited its use in specific gene discovery. Its direct use in large breeding populations is limited by time and cost factors. The advent of next generation sequencing (NGS) technologies and powerful computational pipelines has reduced the cost of whole genome sequencing by many folds allowing discovery, sequencing and genotyping of thousands of markers in a single step (Stapley et al., 2010). NGS has emerged as a powerful tool to detect numerous DNA sequence polymorphism based markers within a short timeframe (Figure S1), growing as a powerful tool for next generation plant breeding.

The initial steps of NGS based marker development involve library construction prior to sequencing. Several targeted marker discovery techniques have been devised using NGS platforms which involve partial representation of the genome and those can be utilized even in absence of prior knowledge on WGS (**Figure 1**). Based on the approaches, partial genome representation libraries are either (i) complexity reduced representation libraries constructed by using restriction enzymes, or (ii) sequence capture libraries without involving restriction digestion. The first group includes reduced-representation libraries (Gore et al., 2009), complexity reduction of polymorphic sequences (Mammadov et al., 2010), restriction-site associated DNA sequencing (RAD-seq) (Pfender et al., 2011), sequence based polymorphic marker technology (Sahu et al., 2012), multiplexed shotgun genotyping (Andolfatto et al., 2011), and genotyping-by-sequencing (GBS) (Elshire et al., 2011). The second group includes technologies like molecular inversion probe (Porreca et al., 2007), solution hybrid selection (Gnirke et al., 2009) and microarray-based genomic selection (Albert et al., 2007). Sequence capture can also be performed for broad or specific targets in the genome such as exome sequencing (Teer and Mullikin, 2010) and sequencing of the genomic region associated with particular trait (Teer et al., 2010).

NGS technologies are already gaining widespread acceptability in the field of crop breeding. Many of the NGS based

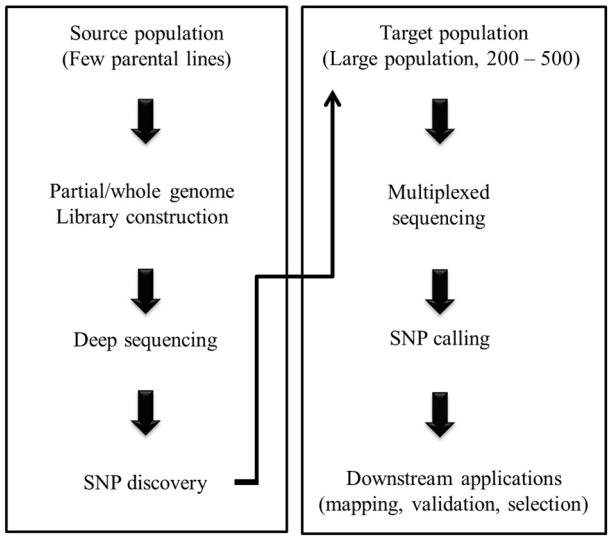

FIGURE 1 | A general outline of NGS assisted plant breeding.

marker discovery techniques allow SNP discovery and genotyping simultaneously, speeding up the whole process (**Figure 1**). Furthermore, availability of gene and transcript sequence data at a large scale in the public domain allows development of genic molecular markers or functional markers. Of the various NGS technologies RAD-seq and GBS have already been proved to be effective for next generation plant breeding (Yang et al., 2012; Glaubitz et al., 2014). RAD-seq is basically a SNP based bulked segregants analysis technique where genomic DNA is sheared with a restriction enzyme of choice followed by ligation of barcoded adapter with molecular identifier (Pfender et al., 2011; Yang et al., 2012). Next, the processed DNA sample from multiple individuals (∼20 individuals) are pooled and randomly sheared so that only a subset of generated fragments contain barcoded adapter. Another divergent adapter is ligated with the fragments for PCR. Divergent adapter ensures amplification of only those fragments containing both adapters. The resultant amplicons are sequenced using an Illumina platform. Finally, pooled samples with different identifiers are separated and SNPs are called using standard bioinformatic pipeline. This technique does not need *a priori* genome sequence information. RAD-seq tagged SNPs have been used to construct a linkage map in eggplant and to identify QTLs for anthocyanin pigmentation of the fruit (Barchi et al., 2012) and also to identify a resistance gene against anthracnose disease in lupin (Yang et al., 2012).

GBS has been used in development of high density map of 20000 SNPs in wheat and 34000 SNPs in barley (Poland et al., 2012a) and to map QTLs for spike architecture and reduced plant height in barley (Liu et al., 2014). It is a simple and highly multiplexed system which follows a modified RAD-seq based library preparation protocol for NGS that reduce sample handling, PCR and subsequent purification steps and completely excludes size fractionation of DNA using efficient barcoding technique. Unlike RAD-seq, the second adapter used in GBS is not a divergent one and hence it allows synthesis of amplicons flanked by any of the three adapter sequence combinations. Powerful bioinformatic pipelines have been established for GBS which can impute missing data utilizing available reference genome (Glaubitz et al., 2014). It allows simultaneous marker discovery and genotyping, and can be scaled up according to need.

If the reference genome sequence is available, the sequence based polymorphic marker technology is quite useful for marker discovery in targeted regions of a genome (Sahu et al., 2012). Short reads are mapped backed to the reference genome to identify putative SNPs. Assembly of multiple short reads assign confidence values to the identified SNPs. Once identified these SNPs are validated by wet lab experiments. The other technique which utilizes reference genome sequence is low coverage multiplexed shotgun genotyping where genomic DNA from multiple genotypes are pooled, sequenced and matched with reference genome with unique linked adapter. Pooling reduces sampling variation and increase efficiency of SNP identification.

The NGS technologies are pivotal to genomic selection, where performance of a target genotype can be predicted from its genomic estimated breeding value determined through statistical models derived using rigorous genotyping and phenotyping of a standard set of breeding population (Poland et al., 2012b). In addition to increasing selection efficiency in annual crop species, these methods are highly valuable for reducing duration of selection in perennial crops, where phenotypic expression of a trait may require several years. However, the complexity of plant breeding situations poses a great challenge to genomic selection, as the relationship between genotype and phenotype often depend on many macro- and micro-environmental factors. Accurate phenotyping and use of robust algorithm are thus of crucial importance to determine the genotype-phenotype relationship for application of genomic selection.

In spite of high potential, the achievements of NGS technologies have been limited to a few examples, most of which have been generated in by institutes with well-established genomic facilities. The technical expertise to extract usable information from huge sequence information presently is insufficient for large scale application of NGS technologies. The most important requirement for reaping benefits of NGS is to enable plant breeders to manage and extract information from huge genomic data. In addition, genomes with higher ploidy level, presence of homeologus sequences and more repetitive sequence poses problems for sequencing and assembly, but some of these problems may be addressed through upcoming technologies (Griffin et al., 2011; Teer et al., 2013). Successful construction of GBS map of wheat with 416,856 markers shows that the robust genetic map of polyploid crops can be constructed through NGS (Saintenac et al., 2013).

Cost of genotyping is another determining factor for adopting appropriate NGS technologies in plant breeding. Since crop breeding handles large population size, it is an expensive process itself. Choice between whole and partial genome sequencing would depend on the availability and judicious use of funds. The cost of WGS for a single genotype of three gigabase genome at 30X coverage is approximate $5000 (Hayden, 2014). Targeted sequencing approach like RAD-seq can sample 200000 SNPs in 100 individuals with same coverage depth at nearly 35-fold less cost compared to WGS of same 100 individuals (Davey et al., 2011). If the whole genome sequence is already available for the target organism the cost involved might further reduce by another 10–14 folds by using techniques like MSG or GBS. Presently, targeted sequencing seems to be more cost-effective option for large scale marker discovery, particularly in case of large and un-decoded genomes. The trend in sequencing technology development closely follows Moore's law (Wetterstrand, 2014), which indicates that the costs for WGS or NGS will reduce by several folds, and WGS may be preferred over partial genome sequencing in near future (Marroni et al., 2012). We expect that targeted sequencing approach would not be completely wiped out by the overwhelming flow of WGS; rather it would be a preferred choice for short term projects for strengthening next generation plant breeding. However, the additional associated cost for target enriched library preparation and bioinformatic analysis that precedes and succeeds the sequencing step, respectively, may not decrease as rapidly as the cost of sequencing the genome. The cost of data mining and efficiency to extract usable information may be more crucial than genotyping cost itself for application of NGS technologies in next generation plant breeding.

Apart from marker discovery, the NGS technologies are also being applied for targeted re-sequencing to identify domestication related genes by comparing the genome of crop species and their wild relatives (Henry, 2012), and also for genome wide selection studies to predict breeding value of traits, all of which have high potential to become application tools for the next generation plant breeders for development of superior cultivars. The ability to directly look into the genome sequences has revolutionized the science of plant breeding in the past few years, and NGS can serve as a worthy weapon for the next generation plant breeders to mitigate the rising demand of food, fiber and fodder in the coming decades. However, it may require some incubation period before this remarkable but complex technology can provide dividends to next generation plant breeders.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fpls.2014.00367/full

REFERENCES

Albert, T. J., Molla, M. N., Muzny, D. M., Nazareth, L., Wheeler, D., Song, X., et al. (2007). Direct selection of human genomic loci by microarray hybridization. *Nat. Methods* 4, 903–905. doi: 10.1038/nmeth1111

Andolfatto, P., Davison, D., Erezyilmaz, D., Hu, T. T., Mast, J., Sunayama-Morita, T., et al. (2011). Multiplexed shotgun genotyping for rapid and efficient genetic mapping. *Genome Res.* 21, 610–617. doi: 10.1101/gr.115402.110

Barchi, L., Lanteri, S., Portis, E., Valè, G., Volante, A., Pulcini, L., et al. (2012). A RAD Tag derived marker based eggplant linkage map and the location of QTLs determining anthocyanin pigmentation. *PLoS ONE* 7:e43740. doi: 10.1371/journal.pone.0043740

Davey, J. W., Hohenlohe, P. A., Etter, P. D., Boone, J. Q., Catchen, JM, and Blaxter, M. L. (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat. Rev. Genet.* 12, 499–510. doi: 10.1038/nrg3012

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., et al. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE* 6:e19379. doi: 10.1371/journal.pone.0019379

Gepts, P. (2004). Crop domestication as a long term selection experiment. *Plant Breed. Rev.* 24, 1–44. doi: 10.1002/9780470650288.ch1

Glaubitz, J. C., Casstevens, T. M., Lu, F., Harriman, J., Elshire, R. J., Sun, Q., et al. (2014). TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. *PLoS ONE* 9:e90346. doi: 10.1371/journal.pone.0090346

Gnirke, A., Melnikov, A., Maguire, J., Rogov, P., LeProust, E. M., Brockman, W., et al. (2009). Solution hybrid selection with ultra-long oligonucleotides formassively parallel targeted sequencing. *Nat. Biotechnol.* 27, 182–189. doi: 10.1038/nbt.1523

Gore, M. A., Chia, J. M., Elshire, R. J., Sun, Q., Ersoz, E. S., Hurwitz, B. L., et al. (2009) A first-generation haplotype map of maize. *Science* 326, 1115–1117. doi: 10.1126/science.1177837

Griffin, P. C., Robin C., and Hoffmann, A. A. (2011). A next-generation sequencing method for overcoming the multiple gene copy problem in polyploid phylogenetics, applied to *Poa* grasses. *BMC Biol.* 9:19. doi: 10.1186/1741-7007-9-19

Hayden, E. C. (2014). The $1000 genome. *Nature* 507, 294–295. doi: 10.1038/507294a

Henry, R. J. (2012). Next-generation sequencing for understanding and accelerating crop domestication. *Brief. Funct. Genomics* 11, 51–56. doi: 10.1093/bfgp/elr032

Liu, H., Bayer, M., Druka, A., Russell, J. R., Hackett, C. A., Poland, J., et al. (2014). An evaluation of genotyping by sequencing (GBS) to map the *Breviaristatum-e (ari-e)* locus in cultivated barley. *BMC Genomics* 15:104. doi: 10.1186/1471-2164-15-104

Mammadov, J. A., Chen, W., Ren, R., Pai, R., Marchione, W., Yalçin, F., et al. (2010). Development of highly polymorphic SNP markers from the complexity reduced portion of maize (*Zea mays* L.) genome for use in marker-assisted breeding. *Theor. Appl. Genet.* 121, 577–588. doi: 10.1007/s00122-010-1331-8

Marroni, F., Pinosio, S., and Morgante, M. (2012). The quest for rare variants: pooled multiplexed next generation sequencing in plants. *Front. Plant Sci.*3:133. doi: 10.3389/fpls.2012.00133

Pfender, W. F., Saha, M. C., Johnson, E. A., and Slabaugh, M. B. (2011). Mapping with RAD (restriction-site associated DNA) markers to rapidly identify QTL for stem rust resistance in *Lolium perenne*. *Theor. Appl. Genet.* 122, 1467–1480. doi: 10.1007/s00122-011-1546-3

Poland, J. A., Brown, J. P., Sorells, M. E. and Jannick, J. L. (2012a). Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. *PLoS ONE* 7:e32253. doi: 10.1371/journal.pone.0032253

Poland, J., Endelman, J., Dawson, J., Rutkoski, J., Wu, S., Manes, Y., et al. (2012b). Genomic selection in wheat breeding using genotyping-by-sequencing. *Plant Genome* 5, 103–113. doi: 10.3835/plantgenome2012.06.0006

Porreca, G. J., Zhang, K., Li, J. B., Xie, B., Austin, D., Vassallo, S. L., et al. (2007). Multiplex amplification of large sets of human exons. *Nat. Methods* 4, 931–936. doi: 10.1038/nmeth1110

Sahu, B. B., Sumit, R., Srivastava, S. K., and Bhattacharya, M. K. (2012) Sequence based polymorphic (SBP) marker technology for targeted genomic regions: its application in generating a molecular map of the *Arabidopsis thaliana* genome. *BMC Genomics* 13:20. doi: 10.1186/1471-2164-13-20

Saintenac C., Jiang D., Wang S., and Akhunov E. (2013). Sequence-based mapping of the polyploid wheat genome. *G3* 3, 1105–1114. doi: 10.1534/g3.113.005819

Schnable, P. S. (2013). *Next Generation Phenotyping and Breeding*. Available online at: http://schnablelab.plantgenomics.iastate.edu/docs/resources/media/Schnable-UMN-3-25-13.pdf (Accessed May 5, 2013).

Stapley, J., Reger, J., Feulner, P. G. D., Smadja, C., Galindo, J., Ekblom, R., et al. (2010). Adaptation genomics: the next generation. *Trends Ecol. Evol.* 25, 705–712. doi: 10.1016/j.tree.2010.09.002

Teer, J. K., Bonnycastle, L. L., Chines, P. S., Hansen, N. F., Aoyama, N., Swift, A. J., et al. (2010). Systematic comparison of three genomic enrichment methods for massively parallel DNA sequencing. *Genome Res.* 20, 1420–1431. doi: 10.1101/gr.106716.110

Teer, J. K., Johnston, J. J., Anzick, S. L., Pineda, M., Stone, G., NISC Comparative Sequencing Program, et al. (2013). Massively-parallel sequencing of genes on a single chromosome: a comparison of solution hybrid selection and flow sorting. *BMC Genomics* 14:253. doi: 10.1186/1471-2164-14-253

Teer, J. K., and Mullikin, J. C. (2010). Exome sequencing: the sweet spot before whole genomes. *Hum. Mol. Genet.* 19, R145–R151. doi: 10.1093/hmg/ddq333

Varshney, R. K., Tuberosa, R. (eds.). (2007). *Genomic Assisted Crop Improvement: Genomics Approaches and Platforms.* New York, NY: Springer.

Wetterstrand, K. A. (2014). *DNA Sequencing costs: Data From the NHGRI Genome Sequencing Program (GSP)*. available online at: www.genome.gov/sequencingcosts (Accessed June 6, 2014).

Yang, H., Tao, Y., Zheng, Z., Li, C., Sweetingham, M., and Howieson, J. (2012). Application of next-generation sequencing for rapid marker development in molecular plant breeding: a case study on anthracnose disease resistance in *Lupinus angustifolius* L. *BMC Genomics* 13:318. doi: 10.1186/1471-2164-13-318

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 20 May 2014; accepted: 09 July 2014; published online: 30 July 2014.

Citation: Ray S and Satya P (2014) Next generation sequencing technologies for next generation plant breeding. Front. Plant Sci. ***5****:367. doi: 10.3389/fpls.2014.00367*

This article was submitted to Crop Science and Horticulture, a section of the journal Frontiers in Plant Science.