Trends in Biotechnology
Trends Biotechnol. September 2009 Vol. 27 No. 9, pp. 495–554 ISSN 0167-7799
CLINIC
BIOREACTORS INC.
BIOREACTORS INC.
Tissue engineering: from bioreactor to the clinic
Plasmid pharmaceuticals
Dynamic fragment-based drug discovery
Next-generation sequencing for crops breeding
Cell PRESS

Review

Next-generation sequencing technologies and their implications for crop genetics and breeding

Rajeev K. Varshney[1,2], Spurthi N. Nayak[1], Gregory D. May[3] and Scott A. Jackson[4]

[1] Centre of Excellence in Genomics (CEG), International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Patancheru 502324, A.P., India
[2] Genomics towards Gene Discovery Subprogramme, Generation Challenge Programme (GCP), c/o CIMMYT, Int APDO Postal 6-641, 06600 Mexico, DF, Mexico
[3] National Center for Genome Resources (NCGR), 2935 Rodeo Park Drive East, Santa Fe, NM 87505, USA
[4] Department of Agronomy, Purdue University, 915 W. State St., West Lafayette, IN 47907-2054, USA

Using next-generation sequencing technologies it is possible to resequence entire plant genomes or sample entire transcriptomes more efficiently and economically and in greater depth than ever before. Rather than sequencing individual genomes, we envision the sequencing of hundreds or even thousands of related genomes to sample genetic diversity within and between germplasm pools. Identification and tracking of genetic variation are now so efficient and precise that thousands of variants can be tracked within large populations. In this review, we outline some important areas such as the large-scale development of molecular markers for linkage mapping, association mapping, wide crosses and alien introgression, epigenetic modifications, transcript profiling, population genetics and *de novo* genome/organellar genome assembly for which these technologies are expected to advance crop genetics and breeding, leading to crop improvement.

Introduction

The detection and exploitation of genetic variation have always been an integral part of plant breeding. DNA-based molecular markers are useful for detecting the genetic variation available in germplasm collections and/or breeding lines. During the past two decades, many different molecular markers have been developed for most major crop species. These markers have been used extensively for the development of saturated molecular genetic and physical maps and for the identification of genes or quantitative trait loci (QTLs) controlling traits of economic importance for marker-assisted selection (MAS) [1,2]. In addition to traditional trait or QTL mapping using biparental populations, new approaches such as association mapping [3], advanced back-cross QTL analysis [4], functional genomics [5], genetical genomics [6], allele mining [1], TILLING and EcoTILLING [7] have become available in recent years. Genomics-assisted breeding is a holistic approach using different genomic strategies and tools [1]. The prediction of phenotype from genotype using different genomic tools and strategies is the basis of genomics-assisted breeding [2]. By improving the precision and efficiency of predicting phenotypes from genotypes, the development of improved cultivars with enhanced resistance or tolerance to biotic and/or abiotic stresses and higher agronomic performance can be greatly accelerated. Indeed, successful examples of genomics-assisted breeding have been demonstrated for several cereals [2,8].

Genomics-assisted breeding approaches have greatly advanced with the increasing availability of genome and transcriptome sequence data for several model plant and crop species. Complete and/or draft genome sequences have become available for several plant species such as rice [9–12], sorghum (http://www.phytozome.net/sorghum) [13,14], poplar (http://www.phytozome.net/poplar.php) [15], grape (http://www.phytozome.net/grape.php) [16], papaya [17], *Medicago* (http://www.medicago.org/genome) and soybean (http://www.phytozome.net/soybean). Whole-genome or gene-space sequencing is in progress for several other crops such as maize (http://www.maizegenome.org), wheat (http://www.wheatgenome.org), barley (http://www.public.iastate.edu/~imagefpc/IBSC%20Webpage/IBSC%20Template-home.html), tomato (http://sgn.cornell.edu/about/tomato_sequencing/) and foxtail millet (http://www.jgi.doe.gov/sequencing/why/99178.html). Complementary to genome sequencing is the widespread application of transcriptome sampling strategies, which has resulted in large collections of expressed sequence tags (ESTs) for nearly all economically important plant species (http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html). Previously, most genome and transcriptome sequencing projects used Sanger sequencing methodology [18]. However, owing to growing interest in human genome resequencing, a new generation of sequencing technologies has emerged. These next-generation sequencing (NGS) technologies are able to generate DNA sequence data inexpensively and at a rate that is several orders of magnitude faster than that of traditional technologies. Advances in sequencing technologies are driving down sequencing costs and increasing sequence capacity at an unprecedented rate, making whole-genome resequencing by individual laboratories possible [19–21]. As a result, genomics-assisted breeding should gain momentum, with

Corresponding author: Varshney, R.K. (r.k.varshney@cgiar.org).

0167-7799/$ – see front matter © 2009 Elsevier Ltd. All rights reserved. doi:10.1016/j.tibtech.2009.05.006

Glossary

Advanced backcross qualitative trait locus (AB-QTL) analysis: method for simultaneous identification and transfer of favorable QTL alleles from unadapted donor lines (e.g. land races and wild species) to elite lines for variety development. Following this strategy, QTL analysis is delayed until the second or third backcross generation and, during the development of these populations, negative selection is exercised to reduce the frequency of deleterious donor alleles. Near isogenic lines (NILs) for QTL of interest can be derived from advanced backcross populations in one or two additional generations and used to verify QTL activity and can be released as a commercial line.

Allele mining: identification of allelic variation of relevant traits within genetic resource collections such as germplasm collections, ecotypes and pathotypes.

Association mapping: also known as linkage disequilibrium (LD) mapping or association analysis, this is a population-based survey used to identify trait–marker relationships based on LD. The technique takes into account all the historic recombination events in a diverse population of individuals to generate higher resolution genetic maps and is needed to complement current map-based cloning methods.

Bacterial artificial chromosome (BAC): a DNA construct, based on a fertility plasmid, used for transforming and cloning in bacteria. BACs are typically 150–350 kbp long, but can be more than 700 kbp. They are often used to sequence the genome of an organism by amplifying its DNA as inserts, which are sequenced before being rearranged *in silico*, to give the genome sequence of the organism.

Genetic map: illustrates the order of genes/marker loci on a chromosome and their relative distances in terms of recombination frequency. Marker loci placed close to each other have a lower recombination frequency than markers placed apart from each other. Molecular markers (see below) can be used to determine the genetic distance between each other and are measured in terms of a genetic map unit or centimorgan.

Genome resequencing: sequencing of a genome for which prior sequence information is available. Owing to large multiples of coverage, resequencing facilitates identification of sequence variants, mutations, structural variations, copy number variations and rearrangements.

Marker-assisted selection (MAS): a method that uses molecular markers associated with the traits of interest to select plants at the seedling stage, thus speeding up the process of conventional plant breeding and reducing the cost involved in maintaining fields. MAS facilitates improvement of traits that cannot easily be selected using conventional breeding methods.

Molecular markers: a set of DNA-based markers that can detect DNA polymorphism at the level of specific loci and at the whole genome level. There are many types of molecular markers: the earliest to be developed were RFLPs (restriction fragment length polymorphisms) and others include RAPDs (random amplification of polymorphic DNAs), CAPS (cleaved amplified polymorphic sites), SSRs (simple sequence repeats) and AFLPs (amplified fragment length polymorphisms). The latest molecular markers developed include SNPs (single nucleotide polymorphisms) and SFPs (single feature polymorphisms).

Physical map: chromosome map of a species that shows the specific physical locations of its genes and/or markers on each chromosome.

Polonies: short for polymerase colonies, these are clonally identical DNA molecules attached to either a single bead or a localized region on a solid support. Polonies can be generated using techniques that include solid-phase PCR in polyacrylamide gels by bridge amplification. They are also referred to as clusters.

Quantitative trait locus (QTL): a region of DNA associated with a particular phenotypic trait. A trait can be controlled by many genes, each having only a small effect, or by a few genes with large effect. QTLs can be used to identify candidate genes underlying a trait.

Reference genome sequences: ideally referred to as genome-wide sequence data obtained by BAC-by-BAC clone sequencing and/or whole genome shotgun sequencing. These sequences give the physical framework of the genome of a particular individual. In cases for which prior genomic information is not available for the species being studied, transcript sequence data (transcript assembly) or BAC-end sequence data or genome sequences of phylogenetically related species can be considered as the reference genome for analyzing NGS data.

TILLING (and EcoTILLING): targeting-induced local lesions in genomes (TILLING) is a reverse genetic method that searches the genomes of mutagenized organisms for mutations in a chosen gene with PCR-based screening of the genes of interest. By comparing the phenotypes of isogenic genotypes differing in single sequence motifs, TILLING provides direct proof of function of both induced and natural polymorphisms without the use of transgenic modifications. A variation of this technique (EcoTILLING) can be used to determine the extent of natural variation in selected genes in crops. EcoTILLING is a cost-effective approach for haplotyping and SNP discovery.

the potential for significant improvements in the precision and efficiency for predicting phenotypes from genotypes. This review article discusses the concept and potential implications of NGS technologies for crop genetics and breeding.

NGS technologies

Sanger dideoxy sequencing [18] and its modifications [22–24] dominated the DNA sequencing field for nearly 30 years and in the past 10 years the length of Sanger sequence reads has increased from 450 bases to more than 1 kb.

The limitations of Sanger sequencing are: (i) the necessity to separate elongation products by size before scanning, requiring one capillary or gel lane per sample; and (ii) the need to produce clonal populations of DNA using *Escherichia coli*, which is labor-, robotics- and space-intensive for large-scale operations. The latter requirement could potentially be reduced by using PCR-based methods (although currently *E. coli* cloning is still used for whole-genome sequencing projects). Individual reaction costs can be reduced by performing the sequencing reactions in reduced reaction volumes [25], but the fundamental restrictions on reducing the cost of Sanger sequencing are at their technological limits.

With advances made in the fields of microfluidics, nanotechnology and informatics, alternative technologies to increase the rapidity and/or throughput of DNA sequencing have recently emerged. The term NGS is used to collectively describe technologies other than Sanger sequencing that have the potential to sequence the human genome in coming years for US$1000 [26] and such technologies are either already commercially available or in development [27,28]. Commercially available NGS technologies such as Roche/454 (http://www.454.com/), Solexa/Illumina (http://www.illumina.com/) and AB SOLiD (http://www3.appliedbiosystems.com/AB_Home/applicationstechnologies/SOLiDSystemSequencing/index.htm) have already demonstrated the potential to circumvent the limiting factors of Sanger sequencing. For example, sequencing can be multiplexed to a much greater extent by many parallel reactions at a greatly reduced cost [19]. The methodology and key features of the NGS technologies are presented in Box 1.

Currently, Roche/454, Solexa and AB SOLiD are the technologies that are predominantly used in crop genetics and breeding applications. Although Roche/454 is superior to Solexa and AB SOLiD in terms of obtaining longer sequence reads, maximum data output is higher for both Solexa and AB SOLiD [21,29]. In terms of costs per run or sequence data generation, Roche/454 is more expensive than either the Solexa or AB SOLiD technologies. All the technologies can be used in the different applications discussed in later sections of the article.

Bioinformatics tools for analyzing NGS data

Sequence reads generated from NGS technologies are shorter than traditional Sanger sequences, which makes assembly and analysis of NGS data challenging. In addition to short DNA sequence reads, these technologies can generate terabyte-sized data files with each instru-

Box 1. Key features of NGS technologies

Amplification-dependent DNA sequencing methods

Roche/454-sequencing is based on polony sequencing and pyrosequencing. The release of pyrophosphate produces light due to cleavage of oxyluciferin by luciferase [66–68]. This method has significant advantages over Sanger sequencing because it requires no electrophoresis step to separate extension products and base incorporation can be detected in real time. The precision of Roche/454 sequencing technology in handling homopolymers (short stretches of the same contiguous nucleotides) suffers in comparison with other NGS technologies.

Illumina/Solexa sequencing is similar to the Sanger-based methods because it uses terminator nucleotides incorporated by a DNA polymerase. However, Solexa terminators are reversible, allowing continuation of polymerization after fluorophore detection and deactivation. With this technology, sheared DNA fragments are immobilized on a solid surface (flow cell channel) and solid-phase amplification is performed. The DNA sequence is determined by synthesis using reversible terminator chemistry and four-channel fluorescent scanning. Unlike Roche/454 sequencing, Solexa has no problems in sequencing homopolymeric regions, but has shorter reads; however, the accuracy is comparable to or better than that of Roche/454 and the output is significantly increased. For each base position sequenced, the Solexa platform requires incorporation, imaging and cleavage of the reversible terminators, thus limiting the read length of Solexa sequences. Owing to the short reads, *de novo* genome sequencing for large plant genomes is problematic because of the difficulty of accurately assembling shorter reads. However, if a nearly identical genome or reference genome sequence is available, this can be used to assemble and/or align individual sequence reads.

AB SOLiD technology is sequencing by oligonucleotide ligation and detection (SOLiD), also known as supported oligonucleotide detection. It depends on ligation-based chemistry with di-base labeled probes and uses minimal starting material. Sequences are obtained by measuring serial ligation of an oligonucleotide to the sequencing primer by a DNA ligase enzyme.

Amplification-independent (single molecule) sequencing methods

Single molecule sequencing (SMS) technology is based on sequencing a single DNA molecule, which can significantly increase the throughput [29]. Apart from the commercially available tSMS (true SMS) launched by Helicos Biosciences (http://www.helicosbio.com/), SMS development is underway at several academic laboratories and companies such as Biotage (http://www.biotage.com/), Li-COR Biosciences (http://www.licor.com/), Nanogen (http://www.nanagen.com/), Network Biosystems (http://www.networkbiosystems.com/) and Visi-Gen Biotechnologies Inc. (http://visigenbio.com/). Pacific Biosciences (http://www.pacificbiosciences.com/) has recently reported real-time sequencing [69]. It is noteworthy that all NGS technologies are constantly improving, with the goal to reduce error rates and to increase the sequence read length and read number.

ment run, greatly increasing the computer resource requirements of sequencing laboratories. Although several bioinformatics tools and algorithms are currently available (Box 2), efforts are underway to improve the accuracy of alignment of NGS data in several laboratories (e.g. [30]). Most of these technologies include software packages that accommodate limited assembly and analyses. However, because NGS technologies are particularly suited for resequencing for single nucleotide polymorphism (SNP) and variation discovery, the software available is biased toward this application. Other applications have been developed, such as the web-based cyber infrastructure

Box 2. Important bioinformatics tools for analysis of NGS data

Alignment, assembly and visualization tools

Velvet (http://www.ebi.ac.uk/~zerbino/velvet/): tool for *de novo* assembly of short and paired reads [70].

EULER (http://euler-assembler.ucsd.edu/portal/): tool to generate short-read assembly and facilitate assembly of combined reads of NGS and Sanger sequencing [71].

GMAP (http://www.gene.com/share/gmap/): program to map and align cDNA sequences to genome sequence using minimal time and memory, facilitates batch processing [72].

MOSAIK (http://bioinformatics.bc.edu/marthlab/Mosaik): tool for pairwise alignment of NGS data to reference sequences.

RMAP (http://rulai.cshl.edu/rmap/): tool to align short reads to a reference genome [30].

SHARCGS (http://sharcgs.molgen.mpg.de/): tool for *de novo* assembly of short reads [73].

SOAP (http://soap.genomics.org.cn/): program for gapped and ungapped alignment of short reads to reference sequences, facilitates single or pair-end resequencing, smRNA discovery and mRNA tag sequence mapping [74].

VCAKE (https://sourceforge.net/projects/vcake): tool for *de novo* assembly of short reads with robust error detection [75].

Zoom (http://www.bioinformaticssolutions.com/products/zoom/index.php): tool to map millions of short reads to reference genomes and carry out post-analysis [76].

EagleView (http://bioinformatics.bc.edu/marthlab/EagleView): display tool for visually inspecting the quality of genome assembly and validating polymorphism candidate sites [77].

JMP® Genomics (http://www.jmp.com/software/genomics/): tool for NGS data visualization and statistical analysis from SAS.

Sequence variant discovery tools

SNPsniffer (http://bioinformatics.bc.edu/marthlab/Polymorphism_Discovery_in_Next-Generation_Sequence_Data): tool for SNP discovery specifically designed for Roche/454 sequences.

Atlas-SNP (http://code.google.com/p/atlas-snp/): tool for SNP and indel discovery from genome resequencing using NGS technologies [78].

SeqMap (http://biogibbs.stanford.edu/~jiangh/SeqMap/): tool to map short sequences to a reference genome and detect multiple substitutions and indels [79].

ssahaSNP (http://www.sanger.ac.uk/Software/analysis/ssahaSNP/): tool to detect homozygous SNPs and indels.

Integrated tools

Alpheus™ (http://alpheus.ncgr.org/): web-based cyber infrastructure platform for pipelining, visualization and analysis of gigabase-scale NGS data and internet-accessible software for variant discovery and isoform identification [31].

MAQ (http://maq.sourceforge.net/): program for mapping and assembly of short reads. It can also report SNPs and indels using a simple assembly visualizer (Maqview) [80].

Next*GENe*™ (http://www.softgenetics.com/NextGENe.html): software to analyze NGS data for *de novo* assembly, SNP and indel detection and transcriptome analysis.

SeqMan genome analyzer (http://www.dnastar.com/products/SMGA.php): software with capacity to align NGS and Sanger data and detect SNPs; also facilitates visualization.

CLCbio Genomics Workbench (http://www.clcbio.com): tool for *de novo* and reference assembly of Sanger and NGS sequence data, SNP detection and browsing.

PanGEA (http://www.kofler.or.at/Bioinformatics/PanGEA/index.html): tool to map NGS data to whole genomes, with SNP detection and display capabilities [81].

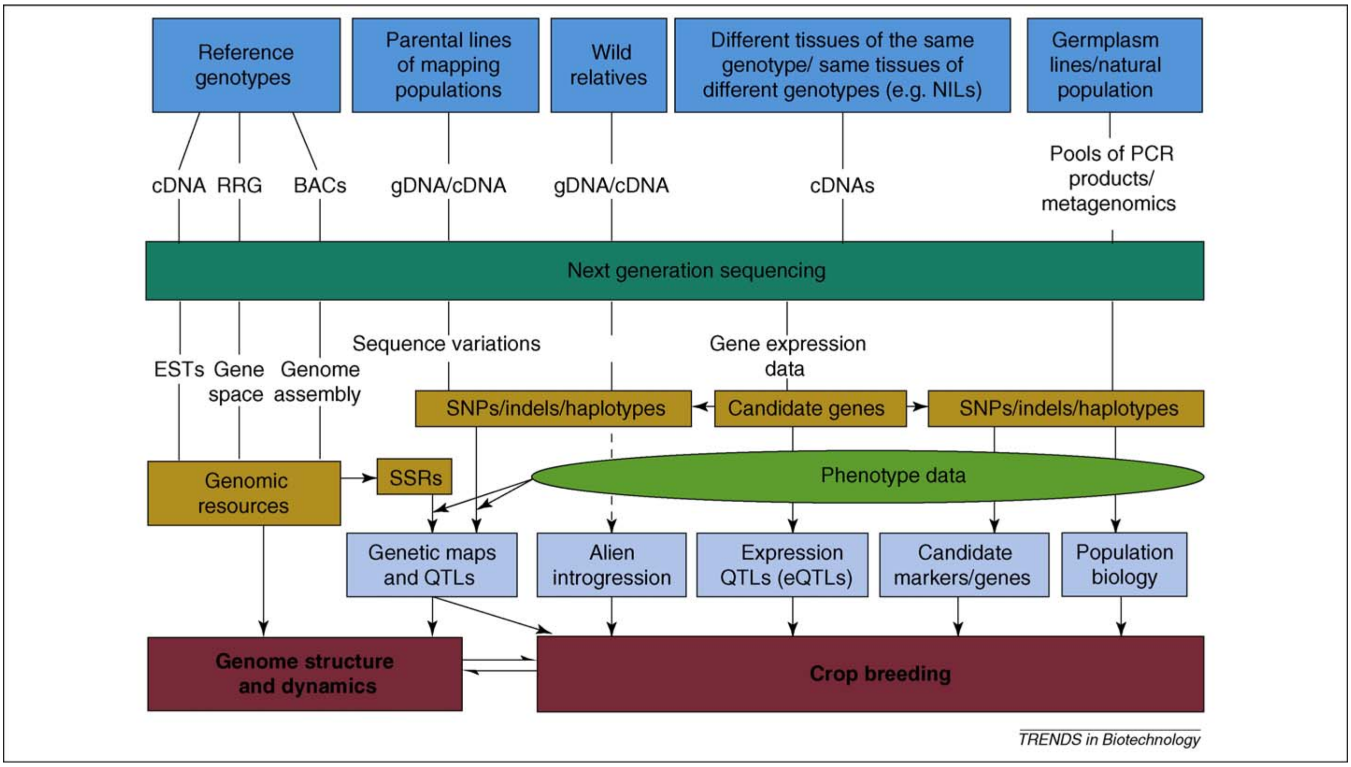

Figure 1. Overview of NGS applications in crop genetics and breeding. NGS technologies have several potential applications in crop genetics and breeding, including the generation of genomic resources, marker development and QTL mapping, wide crosses and alien gene introgression, expression analysis, association genetics and population biology, as shown here. For instance, sequencing of genomic DNA including bacterial artificial chromosomes (BACs), reduced representation of genome (RRG) or cDNA from the reference genotypes using NGS technologies can provide genomic resources such as ESTs, gene space and genome assembly. These resources have a direct impact on understanding the genome architecture for crop genetics. Another application of NGS is in parental genotyping of mapping populations or of wild relatives, which can accelerate the development of molecular markers, e.g. simple sequence repeat (SSR) and single nucleotide polymorphism (SNP) markers. These markers can be used to construct genetic maps, to identify QTLs and to monitor alien genome introgression in the case of wide crosses. These QTL-associated markers for a trait of interest can then be used in selecting progenies carrying favorable alleles via marker-assisted selection (MAS). To develop the functional or perfect gene-based marker, NGS of cDNAs of contrasting genotypes for the trait of interest can be used to identify candidate genes involved in or associated with the trait. The expression mapping of these candidate genes, together with phenotyping of the segregating populations developed from the contrasting genotypes, will provide expression QTLs (eQTLs) and markers associated with these eQTLs should thus serve as the perfect markers for MAS in crop breeding. Another important application of NGS is in association genetics or population biology, where either genomes or pools of PCR products of thousands of candidate genes can be sequenced in hundreds of individuals using barcodes. The sequence data obtained could then be used to identify SNPs or haplotypes across genes or genomes for use in association genetics and/or population biology.

platform Alpheus (http://alpheus.ncgr.org/) [31], which is useful for pipelining, visualization and analysis of giga-base-scale sequence data for identification of SNPs and expression analysis. However, there is still a need for the development of improved bioinformatics tools, pipelines or platforms to facilitate sequence analysis of NGS data in an efficient, reliable and user-friendly manner.

Applications of NGS technologies

NGS technologies have already been used for a variety of applications, such as developing SNP-based markers in a number of plant species both where a reference genome is available (*Arabidopsis* [32] and *Medicago* [33]) and where it is not (maize [34] and *Eucalyptus* [35]). Where reference genome sequences are not available, NGS technologies can be used for draft sequencing via other methods, including pools of bacterial artificial chromosomes (BACs) clones, that can facilitate quick genome assembly, as shown for barley [36]. Interestingly, NGS technologies are proving useful for rapid and efficient development of genomic resources for minor or so-called orphan crop species [37]. NGS technologies are also fast becoming the method of choice for gene expression analysis, particularly for species for which reference genome sequences are already available [32,33]. Efforts are also underway to use NGS technologies for association mapping, wide crosses and alien introgression, epigenetic modifications and population biology. An overview of NGS applications that are relevant to crop genetics and breeding is shown in Figure 1 and some important applications are detailed in the following sections.

Genome variation and molecular markers for marker-assisted selection

Finding and exploiting the DNA sequence variation within a genome is of utmost importance for crop genetics and breeding. Genetic variation can be assayed using a variety of molecular markers. Once molecular markers have been linked to a trait of interest, these markers can be used to select desired lines from a large-scale population through marker-assisted selection (MAS), which saves both costs and time. Furthermore, the availability of gene and transcript sequence data in the public domain [38] has made it possible to develop molecular markers from genes, which have been designated genic molecular markers (GMMs) [39] or functional markers [40]. The development and

Table 1. Applications of NGS technologies in plant genetics and breeding

Species	Details	Refs
Arabidopsis	Among 541 852 ESTs generated through pyrosequencing, 16 000 were novel. This study suggested that two runs were sufficient to detect 90% of all transcripts and found ~9687 novel ESTs. Gene expression studies (digital northerns) obtained from sequence analysis were comparable with earlier studies on microarrays	[32]
Arabidopsis	Solexa sequencing of natural variants of three *Arabidopsis* accessions yielded 120 million–173 million reads that were aligned to a *Arabidopsis* reference genome sequence. Solexa sequence analysis yielded 823 325 unique SNPs.	[42]
Barley	574 Mbp of Solexa sequences were generated and used to generate a mathematically defined repeat index to identify and mark repetitive regions and putative gene spaces	[82]
Chickpea	Transcriptome assembly derived from Solexa tags of root tissues of a drought-tolerant (ICC 4958) and a drought-sensitive (ICC 1882) genotype yielded 5.2 and 3.6 million sequence reads, respectively and ~500 SNPs could be identified (http://www.intl-pag.org/16/abstracts/PAG16_P05f_385.html). This study demonstrated the usefulness of NGS for less-characterized species	
Eucalyptus	Assembly of 148 Mbp of Roche/454 ESTs obtained for multiple genotypes was aligned and 23 742 SNPs were found in uncharacterized, less-studied species such as *Eucalyptus*	[35]
Maize	Roche/454 sequencing generated 261 000 ESTs from shoot apical meristem, of which 30% were novel; ~400 unique ESTs were also identified, for which 27 genes were validated using RT-PCR	[83]
Maize	Transcriptomes of shoot apical meristem from two inbred lines, B73 (260,000 ESTs) and Mo 17 (280,000 ESTs): >7000 SNPs found, 85% of which were successfully validated by Sanger sequencing	[34]
Medicago	Generated 292 465 ESTs comprising 184 599 unique sequences; ~20% novel sequences and 400 SSRs could be identified	[33]
Pinus	Sequencing of the plastome of *Pinus* assemblies to estimate ~88–94% of the complete chloroplast genome	[61]
Wheat	Roche/454 ESTs for two hexaploid wheat lines generated an assembly of 11 700 and 8700 contigs, which were compared with sequences for ancestors of polyploid wheat; 2500 contig assemblies were assigned to one of the homeologous wheat genomes and ~1000 SNPs were found (http://www.intl-pag.org/17/abstracts/P03e_PAGXVII_144.html). The study demonstrates that NGS could be utilized for SNP discovery in polyploidy crops	

application of such GMMs is gaining momentum because their discovery is inexpensive and putative functions can often be deduced by homology searches. Because these markers represent functional units, they are useful for assaying functional diversity in natural populations or germplasm collections and are valuable anchor markers for comparative mapping, evolutionary studies and for MAS [1,8].

Trait mapping and the use of markers, at least for selected traits, have become routine for major crop species, including wheat, maize, rice and soybean. However, for the majority of crop species, particularly less-studied crops such as pearl millet, rye, pigeonpea, cowpea and chickpea, sufficient molecular markers are not available for trait mapping and MAS. NGS methods for developing molecular markers for MAS in crop breeding can be effectively used in two scenarios: (i) in major crop species for which genome, gene space and/or transcriptome sequence data already exist and (ii) in less-characterized species with no or limited genome resources [41], as discussed below.

Resequencing in well-characterized species In species for which genome or EST sequence data are available, genotypes of interest to breeders, such as parental genotypes of mapping populations, can be sequenced by NGS technologies and genome-wide markers can be discovered using NGS sequence data, either from cDNA populations or from genomic DNA of different genotypes (obtained from entire genomes or a reduced representative genome). The sequence data generated can then be aligned to a reference genome (genome or transcriptome assembly) so that variants between genotypes can be identified either on a genome-wide scale or by comparison to the reference genotype. For instance, generation of 15- to 25-fold Solexa sequence data for the *Arabidopsis* reference accession (Col-0) and two divergent accessions (Bur-0 and Tsu-1) and subsequent sequence alignment led to the identification of 823 325 unique SNPs and 79 961 unique 1–3-bp insertion/deletion polymorphisms (indels) [42]. In cases for which complete genome sequence data are not available, alignment of shorter reads (obtained by Solexa or AB SOLiD) with partial genome or transcriptome assembly is challenging. However, several bioinformatics tools and pipelines have recently been developed to address this issue (Box 2). Some examples of the use of NGS for marker discovery for constructing genetic maps and trait mapping for MAS are given in Table 1.

De novo sequencing of crop species without reference sequences Although NGS technologies are ideal for resequencing, *de novo* sequencing can also be undertaken using these sequencing technologies. Generation of a whole genome sequence assembly by alignment of small sequence fragments without the availability of a reference genome is tedious, if not impossible, at present. However, more than one genotype can be used to generate sequence data using NGS technologies and alignment of these data can be facilitated by: (i) genome or transcriptome sequence data for model or major crop species closely related to the species; or (ii) whole transcriptome or reduced representative genome sequence data for the species of interest, generated using Roche/454 sequence technology. Aligning sequence data for more than two genotypes from a single species using one of the above approaches provides confidence in alignments of short sequences and the detection of sequence variants. This approach has also been used for marker discovery in some crop species. For instance, using Roche/454 sequencing, 443 969 ESTs for chickpea and 495 286 ESTs for pigeonpea have been generated at International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), India in collaboration with the J. Craig Venter Institute (JCVI), USA. Furthermore, a collaborative project among ICRISAT, NCGR and UC-Davis (USA) has been instigated to align Solexa sequence data generated for parental genotypes of mapping populations of chickpea and pigeonpea with the transcript assemblies defined based on above-mentioned-Roche/454 tags to identify SNPs between parental genotypes of mapping populations. The SNPs could then be used to develop markers in these marker-deficient crops for trait mapping for MAS. Several other examples of marker discovery in other plant species are presented in Table 1.

Association mapping using natural populations

Association mapping uses one of two approaches, candidate gene sequencing (CGS) or whole genome scanning (WGS) of natural populations [43]. Population surveys for haplotypes identified based on either CGS or WGS can take advantage of past recombination events to identify trait–marker relationships on the basis of linkage disequilibrium (LD). NGS technology has the potential, although not yet demonstrated in the form of publications, to accelerate both CGS- and WGS-based association mapping approaches [44].

In general, CGS-based approaches involve Sanger sequencing of PCR amplicons for selected candidate genes across hundreds of genotypes of the natural population, which is time-intensive and expensive. NGS approaches, in particular Solexa approach, offer the possibility to sequence pools of PCR amplicons for a larger number of candidate genes generated for several hundred genotypes of the natural population with the help of barcodes. Thus, in a single Solexa run, sequence data (SNPs and haplotypes) will be available for a larger number of candidate genes in the natural population within a short time and at considerably lower cost compared to Sanger sequencing. By contrast, WGS approaches require the screening of natural populations with a large set of genome-wide markers, which is not possible in many crop species. However, as mentioned above, NGS technologies can facilitate the rapid development of genome-wide markers [34,42] that could be subsequently used for WGS approaches to association mapping [44].

Wide crosses and alien introgression

The use of genes from wild crop relatives to improve crop performance is well established, in particular for crops that have a narrow genetic base [45]. The use of NGS for wild germplasm is anticipated to have a profound affect because additional molecular markers could be rapidly developed on a genome-wide scale and help to target more narrowly defined genome regions to trace introgression and selection cycles. Sequence-based analysis of the genomes of related germplasm would also reveal information regarding patterns of LD and genome structure, which might make it possible to determine the efficiency of a genome segment introgression. For example, if a genome segment is within a region of high LD, it is less likely to be broken up. Thus, such a genomic drag might be relevant for genes with deleterious effects that are carried together with the gene of interest.

In allopolyploid crops, such as *Brassica*, cotton, tobacco and wheat, SNP identification is challenging because SNPs occurring between genomes have to be discriminated from those present within a genome. For instance, in the paleopolyploid soybean, NGS was successfully used to locate SNPs between several accessions and cultivars and the sequenced reference genome (http://acs.confex.com/crops/2008am/webprogram/Paper45068.html, P.B. Cregan, personal communication). For SNP identification in polyploid crop species, the use of NGS on low-complexity DNA, e.g. restriction-digested DNA samples or cDNA, should be the preferred approach. Although there are bioinformatics analysis issues associated with the analysis of NGS data from polyploid crops, it should be possible to distinguish between duplicated genes (paralogs) in NGS data as opposed to Sanger sequence data. For species without a reference genome, this task is more difficult but not insurmountable because ESTs can be used in the first instance as a reference for SNPs by assigning NGS reads to specific paralogs computationally.

Expression and nucleotide polymorphisms in transcriptomes

Sequence data from RNA samples can be used to detect new RNA species or to measure levels of gene expression and thus to determine the transcriptional state of different cells or tissues [46,47]. Previous studies on high-throughput analysis of the transcriptome relied on microarray analysis and/or serial analysis of gene expression (SAGE), whereas NGS technologies are now used routinely for transcript profiling (Table 1). Unlike microarrays, NGS technologies are not limited to sequenced genomes because they generate tags independently of knowledge of gene annotation, but have the disadvantage that they require extensive sequencing and a reference genome to determine gene identity. Indeed, in several model species such as *Arabidopsis* [32], *Helicobacter* [48], salmon [49] and *Caenorhabditis elegans* [50], NGS was used to demonstrate that deep coverage sequencing with an unbiased representation of transcripts, capturing several rare transcripts, is important for gene discovery and gene expression analysis [51,52].

Population genetics and evolutionary biology

Using NGS technologies, DNA from whole populations can be sequenced rather than just from individuals, thus helping to further our understanding of population genetics. This is commonly referred to as metagenomics [53], a field that is rapidly expanding because of the falling costs of DNA sequencing. Indeed, identifying a species within a given population by its highly conserved sequences, such as single-stranded rRNA, initiated the era of metagenomics [54]. Metagenomics has been successful in furthering our understanding of microbial populations and the community structure and composition of varied environmental conditions, including deep seas [55,56], soil [57] and deep mines [58]. NGS technologies can enhance the power of metagenomic sequencing approaches to resolve rare species [56,59]. It is anticipated that genome resequencing using NGS technologies for species with reference genome sequences will revolutionize the study of population-level plant diversity [19]. One example of this approach in plants is the 1001 genomes *Arabidopsis* project (http://1001genomes.org/index.html), in which sequencing of 1001 accessions of the model plant *Arabidopsis* is being undertaken using Solexa, Roche/454 and AB SOLiD technologies. The resulting information is expected to provide genome-wide LD structures and haplotype data that might have broad implications for evolutionary sciences and plant breeding.

Organellar and genome-wide assembly

In the past, the genomes of organelles, such as chloroplasts and mitochondria, were sequenced using Sanger technology to study cytoplasmic inheritance. For instance, male sterility genes, which are important for hybrid crops, are present in mitochondria and therefore sequence analysis of the mitochondrial genome could help to improve hybrid crop production. NGS technologies have increased the availability of organellar genomes, such as for mitochondria [60] and chloroplasts [61], with further increases anticipated in the near future.

The use of NGS technologies for *de novo* assembly of whole genomes has been much anticipated [62]. However, assembly of whole genomes of plant species from sequences generated by NGS technologies is difficult because most plant or crop genomes are large and full of repetitive DNA sequences. The short reads inherent to NGS technologies cannot be assembled using current informatics technology because the repetitive sequences present are longer than the reads and thus many or most reads cannot be unambiguously assigned, resulting in very short sequence scaffolds. Even for relatively simple genomes, such as bacteria and *Arabidopsis*, NGS has not resulted in complete chromosomal or even chromosomal-arm scaffolds. For instance, to test the efficacy of NGS for BAC sequencing in barley (a large and complex genome species), Wicker *et al.* [36] compared Roche/454 sequencing with Sanger sequencing for four BAC clones. They found that although Roche/454 sequencing covered all gene-containing regions efficiently, the method exhibited problems in the sequencing of repetitive DNA sequences. Thus, a combination of approaches is being considered to sequence crop genomes using NGS technologies to capitalize on cost savings. One approach is to reduce the complexity of the genome by sequencing BAC clones either from a pre-assembled physical map or in the absence of a physical map. The idea is that BAC clones, each of ~100–150 kb, would be easier to assemble individually than an entire genome. Cost-saving measures such as BAC pooling, barcoding of clones and others are being tested. Another approach is to combine NGS technologies with some Sanger sequencing. Using paired ends of larger insert clones sequenced via Sanger, the rest of the genome could theoretically be filled in by producing a massive amount of NGS sequence data. This approach remains largely untested at present, but considerable research efforts are underway in several species, including pigeonpea and wheat, among others.

Epigenetic modifications

Epigenetics is the study of heritable gene regulation. Epigenetic changes do not involve the DNA sequence itself, but modification by DNA methylation or post-translational modification of histone tails, which are known to play a key role in gene expression and in plant development under stress. The DNA–protein interactions that underlie this type of regulation of gene expression are frequently determined by chromatin immunoprecipitation (ChIP). The most prominent genomic approaches for analyzing epigenetic changes have used ChIP followed by microarray hybridization, the so-called ChIP-chip. More recently, NGS technologies have replaced ChIP-chip with so-called ChIP sequencing, which entails conventional ChIP followed by direct sequencing. This method offers superior data compared with ChIP-chip, with less noise and higher resolution. ChIP sequencing is already well established for human genome analysis [63], but only a few reports are available for plant systems. For instance, in *Arabidopsis* the cytosine methylome (methylC-seq), transcriptome (mRNA-seq), and small RNA transcriptome (smRNA-seq) were directly sequenced using Solexa technology, which led to the generation of highly integrated epigenome maps for wild-type *Arabidopsis* and for mutants defective in either DNA methyltransferase or demethylase activity. Moreover, previously undetected DNA methylations could be identified at the single nucleotide level. Deep sequencing of smRNAs also showed perturbation of smRNA biogenesis upon loss of CpG DNA, thereby establishing a potential link between epigenetics and smRNA regulation [64].

Prospects for crop improvement

As evident from the above examples, NGS can have significant implications for crop genetics and breeding. The development of large-scale genomic resources, including transcript and sequence data, molecular markers and genetic and physical maps, is significant, in addition to other potential applications. Transcriptome and genome sequencing (both resequencing and *de novo*) using NGS technology is increasing for crop plants. The use of NGS technologies has already led to a quantum leap in the amount of genomic data available for crops for which not many genomic resources were previously available, such as chickpea and pigeonpea [37]. Moreover, the availability of large numbers of genetic markers developed through NGS technologies is facilitating trait mapping and making marker-assisted breeding more feasible. For instance, large-scale development of molecular markers using NGS can facilitate linkage mapping and WGS-based association genetics that are of practical use for MAS in marker-deficient crops. Metagenomics approaches and the sequencing of pooled amplicons generated for a large number of candidate genes across large populations offer possibilities to better understand population biology and to study genome-wide association genetics. Another important application of NGS is in gene expression studies, for which NGS has the potential to replace microarray experiments in the near future; in contrast to other gene expression approaches such as microarray and real-time PCR, NGS technologies can provide insights into the spatial and temporal control of gene expression owing to their ability to identify all RNA transcripts produced at a specific time [65].

Although the initial aim of NGS technologies was resequencing, they are currently being used to explore *de novo* genome sequencing in several crop species, including wheat, pigeonpea and common bean. If the ongoing revolution in NGS technologies can reduce the cost for resequencing the genome to only a few hundred US dollars, genome sequencing/resequencing will not be limited to model plant and major crop species and could be extended to parental and progeny lines of mapping populations and of germplasm lines currently present in different germplasm repositories. On one hand, genome-wide sequence data should greatly facilitate our understanding of complex phenomena, such as heterosis and epigenetics, which have implications for crop genetics and breeding. On the other hand, these genomics data will also enable breeders to visualize which fragment of a chromosome is derived from which parent in the progeny line, thereby identifying clear crossover events occurring in every progeny line and placing markers on genetic and physical maps without ambiguity. Eventually, this will help in introducing specific chromosome regions from one cultivar to another. Therefore, it can be anticipated that NGS technologies will be particularly useful for developing and confirming introgression lines for a trait of interest. In addition to facilitating genomics-assisted breeding, NGS can also accelerate the development of transformation technologies for crops because it will become easier to modify genes with the increasing availability of genomic data. Although large-scale NGS data analysis remains a challenge at present, significant progress is being made in improving existing tools and in developing new approaches for this task.

In summary, we envisage an exponential increase in the use of NGS technologies, not only for major crop species, but also for so-called orphan crops. The results of these efforts will have a profound impact on crop breeding.

Acknowledgements

We thank the Generation Challenge Programme (GCP) (R.K.V., G.D.M.), the National Science Foundation (DBI 0501877 and 0227414 to S.A.J.), the Pigeonpea Genomics Initiative of the Indian Council of Agricultural Research (ICAR) under the umbrella of the Indo–US Agricultural Knowledge Initiative (AKI) (R.K.V.), the Department of Biotechnology (R.K.V.) and the Council of Scientific and Industrial Research (for a fellowship to S.N.N.) of the Government of India for funding research in our laboratories. Thanks are also due to B. Jayashree, V. Thakur and the anonymous reviewers for their useful suggestions on earlier versions of the manuscript.

References

1 Varshney, R.K. *et al.* (2005) Genomics-assisted breeding for crop improvement. *Trends Plant Sci.* 10, 621–630

2 Varshney, R.K. *et al.* (2006) Advances in cereal genomics and applications in crop breeding. *Trends Biotechnol.* 24, 490–499

3 Ersoz, E.S. *et al.* (2007) Applications of linkage disequilibrium and association mapping in crop plants. In *Genomics Assisted Crop Improvement: Genomics Approaches and Platforms* (Varshney, R.K. and Tuberosa, R.T., eds), pp. 97–120, Springer

4 Tanksley, S.D. and Nelson, J.C. (1996) Advanced backcross QTL analysis: a method for the simultaneous discovery and transfer of valuable QTLs from unadapted germplasm into elite breeding lines. *Theor. Appl. Genet.* 92, 191–203

5 Schena, M. (1998) Microarrays: biotechnology's discovery platform for functional genomics. *Trends Biotechnol.* 16, 301–306

6 Jansen, R.C. and Nap, J.P. (2001) Genetical genomics**:** the added value from segregation. *Trends Genet.* 17, 388–391

7 Till, B.J. *et al.* (2007) TILLING and EcoTILLING for crop improvement. In *Genomic Assisted Crop Improvement: Genomics Approaches and Platforms* (Varshney, R.K. and Tuberosa, R., eds), pp. 333–349, Springer

8 Varshney, R.K. *et al.* (2007) Application of genomics for molecular breeding of wheat and barley. *Adv. Genet.* 58, 122–155

9 Barry, G. (2001) The use of the Monsanto draft rice genome sequence in research. *Plant Physiol.* 125, 1164–1165

10 Yu, J. *et al.* (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science* 296, 79–92

11 Goff, S.A. *et al.* (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science* 296, 92–100

12 International Rice Genome Sequencing Project (2005) The map-based sequence of the rice genome. Nature 436, 793–800

13 Bedell, J.A. *et al.* (2005) Sorghum genome sequencing by methylation filtration. *PLoS Biol.* 3, 103–115

14 Paterson, A.H. *et al.* (2009) The *Sorghum bicolor* genome and the diversification of grasses. *Nature* 457, 551–556

15 Tuskan, G.A. *et al.* (2006) The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* 313, 1596–1604

16 Jaillon, O. *et al.* (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* 449, 463–467

17 Ming, R. *et al.* (2008) The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). *Nature* 452, 991–996

18 Sanger, F. *et al.* (1977) DNA sequencing with chain- terminating inhibitors. *Proc. Natl. Acad. Sci. U. S. A.* 74, 5463–5467

19 Hudson, M. (2008) Sequencing breakthroughs for genomic ecology and evolutionary biology. *Mol. Ecol. Resour.* 8, 3–17

20 Mardis, E.R. (2008) The impact of next-generation sequencing technology on genetics. *Trends Genet.* 24, 133–141

21 Gupta, P.K. (2008) Ultrafast and low-cost DNA sequencing methods for applied genomics research. *Proc. Natl. Acad. Sci. India* 78, 91–102

22 Prober, J.M. *et al.* (1987) A system for rapid DNA sequencing with fluorescent chain terminating dideoxynucleotides. *Science* 238, 336–341

23 Smith, L.M. *et al.* (1986) Fluorescence detection in automated DNA sequence analysis. *Nature* 321, 674–679

24 Madabhushi, R.S. (1998) Separation of 4-color DNA sequencing extension products in noncovalently coated capillaries using low viscosity polymer solutions. *Electrophoresis* 19, 224–230

25 Smailus, D.E. *et al.* (2005) Simple, robust methods for high-throughput nanoliter-scale DNA sequencing. *Genome Res.* 15, 1447–1450

26 Service, R.F. (2006) The race for the $1000 genome. *Science* 311, 1544–1546

27 Shendure, J. *et al.* (2004) Advanced sequencing technologies: Methods and goals. *Nat. Genet.* 5, 335–344

28 Kling, J. (2005) The search for a sequencing thoroughbred. *Nat. Biotechnol.* 23, 333–1335

29 Gupta, P.K. (2008) Single-molecule DNA sequencing technologies for future genomics research. *Trends Biotechnol.* 26, 602–611

30 Smith, A.D. *et al.* (2008) Using quality scores and longer reads improves accuracy of Solexa read mapping. *BMC Bioinformatics* 9, 128

31 Miller, N.A. *et al.* (2008) Management of high-throughput DNA sequencing projects: *Alpheus J. Comput. Sci. Syst. Biol.* 1, 132–148

32 Weber, A.P.M. *et al.* (2007) Sampling the *Arabidopsis* transcriptome with massively parallel pyrosequencing. *Plant Physiol.* 144, 32–42

33 Cheung, F. *et al.* (2006) Sequencing *Medicago truncatula* expressed sequenced tags using 454 life sciences technology. *BMC Genomics* 7, 272–282

34 Barbazuk, W.B. *et al.* (2007) SNP discovery via 454 transcriptome sequencing. *Plant J.* 51, 910–918

35 Novaes, E. *et al.* (2008) High-throughput gene and SNP discovery in *Eucalyptus grandis*, an uncharacterized genome. *BMC Genomics* 9, 312

36 Wicker, T. *et al.* (2006) 454 sequencing put to the test using the complex genome of barley. *BMC Genomics* 7, 275

37 Varshney, R.K. *et al.* (2009) Orphan legume crops enter the genomics era! *Curr. Opin. Plant Biol.* 12, 202–210

38 Rudd, S. (2003) Expressed sequence tags: alternative or complement to whole genome sequence. *Trends Plant Sci.* 8, 321–329

39 Varshney, R.K. *et al.* (2007) Genetic molecular markers in plants: development and applications. In *Genomic Assisted Crop Improvement: Genomics Approaches and Platforms* (Varshney, R.K. and Tuberosa, R., eds), pp. 13–30, Springer

40 Andersen, J.R. and Lubberstedt, T. (2003) Functional markers in plants. *Trends Plant Sci.* 8, 554–560

41 Vera, J.C. *et al.* (2008) Rapid transcriptome characterization for nonmodel organism using 454 pyrosequencing. *Mol. Ecol.* 17, 1636–1647

42 Ossowski, S. *et al.* (2008) Sequencing of natural strains of *Arabidopsis thaliana* with short reads. *Genome Res.* 12, 2024–2033

43 Rafalski, A. (2002) Novel genetic mapping tools in plants: SNPs and LD-based approaches. *Plant Sci.* 162, 329–333

44 Nordborg, M. and Weigel, D. (2008) Next-generation genetics in plants. *Nature* 456, 720–723

45 Hajjar, R. and Hodgkin, T. (2007) The use of wild relatives in crop improvement: a survey of developments over the last 20 years. *Euphytica* 156, 1–13

46 Velculescu, V.E. *et al.* (1995) Serial analysis of gene expression. *Science* 270, 484–487

47 Bentley, D.R. (2006) Whole-genome re-sequencing. *Curr. Opin. Genet. Dev.* 16, 545–552

48 Dohm, J.C. *et al.* (2008) Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res.* 36, 105

49 Quinn, N.L. *et al.* (2008) Assessing the feasibility of GS FLX pyrosequencing for sequencing the Atlantic salmon genome. *BMC Genomics* 9, 404

50 Shin, H. *et al.* (2008) Transcriptome analysis for *Caenorhabditis elegans* based on novel expressed sequence tags. *BMC Biol.* 6, 30

51 Barski, A. *et al.* (2007) High-resolution profiling of histone methylations in the human genome. *Cell* 129, 823–837

52 Johnson, D.S. *et al.* (2007) Genome-wide mapping of *in vivo* protein–DNA interactions. *Science* 316, 1497–1502

53 Handelsman, J. (2004) Metagenomics: application of genomics to uncultured microorganisms. *Microbiol. Mol. Biol. Rev.* 68, 669–685

54 Huse, S.M. *et al.* (2007) Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol.* 8, R143

55 Venter, J.C. *et al.* (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science* 304, 66–74

56 Sogin, M.L. *et al.* (2006) Microbial diversity in the deep sea and the under-explored "rare biosphere". *Proc. Natl. Acad. Sci. U. S. A.* 103, 12115–12120

57 Leininger, S. *et al.* (2006) Archaea predominate among ammonia-oxidizing prokaryotes in soils. *Nature* 442, 806–809

58 Edwards, R.A. *et al.* (2006) Using pyrosequencing to shed light on deep mine microbial ecology. *BMC Genomics* 7, 57

59 Hall, N. (2007) Advanced sequencing technologies and their wider impact in microbiology. *J. Exp. Biol.* 209, 1518–1525

60 Jex, A.R. *et al.* (2008) Using 454 technology for long-PCR based sequencing of the complete mitochondrial genome from single *Haemonchus contortus* (Nematoda). *BMC Genomics* 11, 9–11

61 Cronn, R. *et al.* (2008) Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology. *Nucleic Acids Res.* 36, e122

62 Paux, E. *et al.* (2008) A physical map of the 1-gigabase bread wheat chromosome 3b. *Science* 322, 101–104

63 Mardis, E.R. (2007) ChIP-seq: welcome to the new frontier. *Nat. Methods* 4, 613–614

64 Lister, R. *et al.* (2008) Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. *Cell* 133, 1–14

65 Ronaghi, M. *et al.* (1996) Real-time DNA sequencing using detection of pyrophosphate release. *Anal. Biochem.* 242, 84–89

66 Ronaghi, M. *et al.* (1998) A sequencing method based on real-time pyrophosphate. *Science* 281, 363–365

67 Ronaghi, M. (2001) Pyrosequencing sheds light on DNA sequencing. *Genome Res.* 11, 3–11

68 Blow, N. (2009) Transcriptomics: the digital generation. *Nature* 458, 239–242

69 Eid, J. *et al.* (2009) Real-time DNA sequencing from single polymerase molecules. *Science* 323, 133–138

70 Zerbino, D.R. and Velvet, E.B. (2008) Algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res.* 18, 821–829

71 Chaisson, M.J. and Pevzner, P.A. (2008) Short read fragment assembly of bacterial genomes. *Genome Res.* 18, 324–330

72 Wu, T.D. and Watanabe, C.K. (2005) GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* 21, 1859–1875

73 Dohm, J.C.C. *et al.* (2007) SHARCGS, a fast and highly accurate short-read assembly algorithm for *de novo* genomic sequencing. *Genome Res.* 17, 1697–1706

74 Li, R. *et al.* (2008) SOAP: short oligonucleotide alignment program. *Bioinformatics* 24, 713–714

75 Jeck, W.R. *et al.* (2007) Extending assembly of short DNA sequences to handle error. *Bioinformatics* 23, 2942–2944

76 Lin, H. *et al.* (2008) ZOOM! Zillions of oligos mapped. *Bioinformatics* 24, 2431–2437

77 Huang, W. and Marth, G.T. (2008) EagleView: a genome assembly viewer for next-generation sequencing technologies. *Genome Res.* 18, 1538–1543

78 Wheeler, D.A. *et al.* (2008) The complete genome of an individual by massively parallel DNA sequencing. *Nature* 452, 872–876

79 Jiang, H. and Wong, W.H. (2008) SeqMap: mapping massive amount of oligonucleotides to the genome. *Bioinformatics* 24, 2395–2396

80 Li, H. *et al.* (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* 18, 1851–1858

81 Kofler, R. *et al.* (2009) PanGEA: Identification of allele specific gene expression using the 454 technology. *BMC Bioinformatics* 10, 143

82 Wicker, T. *et al.* (2008) Low-pass shotgun sequencing of the barley genome facilitates rapid identification of genes, conserved non-coding sequences and novel repeats. *BMC Genomics* 9, 518

83 Emrich, S.J. *et al.* (2007) Gene discovery and annotation using LCM-454 transcriptome sequencing. *Genome Res.* 17, 69–73