Mol Breeding (2010) 25:553–570
DOI 10.1007/s11032-009-9357-9

REVIEW

Utilization of next-generation sequencing platforms in plant genomics and genetic variant discovery

Stéphane Deschamps · Matthew A. Campbell

Received: 14 August 2009 / Accepted: 23 November 2009 / Published online: 5 December 2009

Abstract Genome-wide variant detection within a species is the primary initial step towards linking genotypic variation and phenotypes. The conversion of these genetic variants (the most prevalent of these being single-nucleotide polymorphisms or SNPs) into genetic markers is particularly important in agronomically valuable crop species to allow for cost-effective marker-assisted selection strategies, whole-genome fingerprinting, association studies, map-based gene cloning and population-based analyses. Towards these goals, an increasing number of large-scale genetic variant discovery initiatives are being undertaken in conjunction with next-generation sequencing platforms, allowing for drastically quicker and cheaper variant discovery, and leading towards a far more comprehensive view of the genome or transcriptome. This review will summarize the current status of these initiatives and will discuss the expanding role of next-generation sequencing technologies in facilitating crop improvement.

S. Deschamps (✉)
DuPont Crop Genetics, Experimental Station,
PO Box 80353, Wilmington, DE 19880, USA
e-mail: Stephane.Deschamps@cgr.dupont.com

M. A. Campbell
Pioneer Hi-Bred International, 7300 NW 62nd Ave,
Johnston, IA 50131, USA

Keywords Plant genomics · SNP discovery · Genome · Transcriptome · Resequencing · Next generation sequencing

Introduction

The ability to obtain important information on genes, underlying genetic variation, and gene function originates with the ability to sequence both the genomic and transcribed DNA of an organism. This sequencing can be used to generate a reference genome assembly, to query the transcribed fraction of the genome in order to assist with the annotation of the reference sequence, or to obtain genetic variation among individuals within a population through the comparison of sequences at a given locus or loci. The advent of next-generation sequencing platforms has drastically increased the speed at which DNA sequence can be acquired while reducing the costs by several orders of magnitude. These next-generation platforms generate shorter reads with lower quality when compared to the Sanger platform. This reduction in read length and quality necessitates the development of bioinformatics tools to assist in either the mapping of these shorter reads to reference sequences or *de novo* assembly. In order to compensate for the lower quality, redundancy of reads at a given nucleotide position (coverage) is employed to discern sequencing errors from true genetic variation.

This review will focus on the current and future new sequencing platforms as well as how two of these platforms currently are being utilized in genetic variant discovery among agronomically important species within the Plant Kingdom.

Sequencing systems

Since the advent of the modern genomics era in the early 1990's, the majority of prokaryotic and eukaryotic genome sequencing projects has relied almost exclusively on instruments carrying out semi-automated versions of the Sanger biochemistry (Sanger et al. 1977; Hunkapiller et al. 1991). The resulting genomic assemblies (The Arabidopsis Genome Initiative 2000; Lander et al. 2001; Goff et al. 2002; Yu et al. 2002; IRGSP 2005; Tuskan et al. 2006; Jaillon et al. 2007; Paterson et al. 2009) are based upon strategies deriving from shotgun *de novo* sequencing (Roe 2004), where many copies of randomly fragmented DNA cloned into a high-copy number plasmid, generated after transformation into *Escherichia coli*, are end-sequenced in reactions consisting of successive rounds of template denaturation, primer annealing and primer extension, and then assembled *de novo* to recreate larger sequences, or "contigs". Another shotgun application of the Sanger biochemistry is the sequencing of cDNA fragments derived from transcripts and cloned into high-copy number plasmids. The resulting sequences (classified as expressed sequenced tags, or "ESTs") can be used to generate a "snapshot" of the expression of genes within a tissue, or pool of tissues, at a given time or under given conditions (Clifton and Mitreva 2009; Parkinson and Blaxter 2009). In addition to genome assembly and expression studies, another major application of the Sanger-based biochemistry is the discovery of genetic variation among a set of individuals in a population. An original method of discovery employs the digestion of genomic DNA with a restriction endonuclease followed by size-selection on an agarose gel, shotgun sequencing and alignment of homologous sequences from each individual to reveal sequence variation (Altshuler et al. 2000). A more targeted approach has used separate PCR amplifications at a locus of interest among a set of individuals followed by the separate end-sequencing of the individual amplicons and the comparison of sequence assemblies to determine the presence of genetic variation (Rostoks et al. 2005; Choi et al. 2007). Another strategy is to screen ESTs or EST-derived transcript assemblies resulting from Sanger sequencing data to obtain genetic variation within exons among a selected number of genotypes (Buetow et al. 1999; Picoult-Newberg et al. 1999; Sachidanandam et al. 2001; Useche et al. 2001; Lijavetzky et al. 2007; Duran et al. 2009). Significant improvements, such as the introduction of fluorescent dye terminators of different colors (Smith et al. 1986; Prober et al. 1987) and the replacement of the original slab gel electrophoresis with capillary gel electrophoresis (Luckey et al. 1990; Swerdlow and Gesteland 1990), have allowed the Sanger biochemistry to achieve higher throughputs. Additionally, reductions in volumes for the reactions can be employed to reduce reagent costs (Smailus et al. 2006). These successive improvements have led to current read lengths of up to $\sim$1,000 bps at a per-base error rate as low as 0.001% (Ewing and Green 1998). The current cost of Sanger sequencing now is on the order of $0.01 per base.

Entirely new strategies for sequencing (now dubbed "next-generation sequencing") have emerged over the past 5 years (Shendure and Ji 2008), reducing the cost of sequencing by over two orders of magnitude, and enabling, in a matter of days, sequencing throughputs that otherwise would have taken months to generate with a classic Sanger sequencing approach (Table 1). Even though the complexity and timelines necessary to generate the libraries vary markedly, existing next-generation sequencing strategies all follow a similar pattern for library preparation that can be summarized in three major steps: (1) random shearing of DNA, either via nebulization or sonication; (2) ligation of universal adapters at both ends of the sheared DNA fragments and (3) immobilization and amplification of the adapter-flanked fragments to generate clustered amplicons to serve as templates for the sequencing reactions (Shendure and Ji 2008). These new platforms also share the fact that they parallelize in a single run from hundreds of thousands to hundreds of millions of sequencing reactions. Through alternating cycles of base incorporation and image capture, these platforms produce short DNA sequences which range in size from 25 to 500 bases and generate from several hundreds of millions to several billions bases per runs.

Table 1 Comparison of next-generation sequencing technologies

Sequencing platform	Sequencing chemistry	Run time[a]	Read length (bp)	Reads per run (million)	Throughput per run (Gbp)
Roche 454 FLX	Sequencing by synthesis	10 h	400–500	~1	0.4–0.5
Illumina GAIIx	Sequencing by synthesis	5.5 days	100	160	16
ABI SOLiD	Sequencing by ligation	6–7 days	50	500	25
Helicos HeliScope	Sequencing by synthesis	8 days	25–55	600–800	21–28
Polonator	Sequencing by ligation	80 h[b]	28 (2 × 14)[b]	300–400	9[b]

[a] Not including fragment library construction

[b] Paired-end read sequencing

The raw data generated by the instrument generally are contained within terabyte-sized folders that require complex computer storage and analytical power to be translated into DNA sequences. Some next-generation sequencing systems offer real-time basecalling abilities where raw data are translated into sequences in a real-time fashion, before the DNA sample is fully processed and the instrument completes the last cycle. The resulting sequences and accompanying files are smaller gigabyte-sized text-encoded files that can be transferred and stored on a local desktop. The raw data are automatically deleted from the system after being processed, thus alleviating the need for large storage capacity required to maintain the images of each cycle. All next-generation sequencing systems also provide computational tools for basecalling and post-processing data analysis, such as alignment to a reference sequence and variant detection. Examples of such tools are given in a later chapter.

While next-generation sequencing platforms offer much higher throughput with greatly reduced costs, one liability is that the error rates are on average 10 times greater than that of Sanger sequencing (Shendure and Ji 2008). However, the massive throughput typically provides a very high base coverage which can be used to screen for sequencing errors and separate them from true variations (Hillier et al. 2008). Thus genetic variant discovery with these next-generation sequencing platforms currently has a trade-off between coverage needs for desired rates of true SNP discovery with a given capacity versus the total number of unique individuals being screened (Pushkarev et al. 2009). All current next-generation sequencing platforms (often labeled as "second-generation" sequencing platforms, due to the emergence of "third-generation" sequencing platforms, described in a later chapter) offer "paired-end" read capability (i.e. both ends of a given DNA fragments are sequenced), which, depending on the physical distance between the read pairs, can become useful in applications such as *de novo* sequencing or metagenomics.

As current second-generation sequencing platforms generate reads whose lengths are shorter than typical read lengths achieved by Sanger sequencing, second-generation sequencing technologies have been primarily applied in contexts that could benefit from such short read lengths (Barski et al. 2007; Fahlgren et al. 2007; Johnson et al. 2007; Kasschau et al. 2007; Cokus et al. 2008; Lister et al. 2008; Lu et al. 2008; Mortazavi et al. 2008; Nobuta et al. 2008; Sultan et al. 2008; Sunkar et al. 2008; Wilhelm et al. 2008). An important application of second-generation sequencing technologies has been the identification of sequencing variants, by aligning short DNA or cDNA reads to the sequence of a DNA molecule from another individual, another strain or another species (Barbazuk et al. 2007; Hillier et al. 2008; Ossowski et al. 2008; Van Tassel et al. 2008; Wang et al. 2008; Ahn et al. 2009; Amaral et al. 2009; Eck et al. 2009; Gore et al. 2009; Huang et al. 2009; Ramos et al. 2009; Trick et al. 2009). Given their relatively short length and higher error rates when compared with Sanger sequencing, the presence of multiple overlapping reads for a confident base assignment and to confirm a given mutation or polymorphism is paramount to the success of that approach. This review focuses on the use of next-generation sequencing reads for sequencing variant discovery in plants.

454 GS FLX

The first next-generation sequencing instrument to become commercially available was the 454 GS20 sequencing system from Roche (Basel, Switzerland)

(Margulies et al. 2005). The GS20 generates over ~200,000 100 bps sequences (~20 million bps), in just over 4 h, while the latest 454 sequencing system, the GS FLX Titanium series, now is capable of generating over ~1,000,000 450 bp sequences in a 10 h run. The ability to generate ~450 bps reads represents the key advantage of the 454 sequencing system over other second-generation sequencing systems that may offer much larger throughputs but contained within smaller reads. Because of the importance of read length in *de novo* assembly projects, the 454 sequencing system often is the second-generation sequencing system of choice for sequencing large genomes (Velasco et al. 2007; Wheeler et al. 2008). However, the lower sequencing throughput of the 454 sequencing system means that the overall cost per base is higher than for other second-generation sequencers. While *de novo* assemblies of complex regions generated with 454 sequences can provide large amounts of biologically relevant information, the length of 454 reads often precludes from generating assemblies that can be as informative as assemblies generated with the Sanger biochemistry. When compared to the Sanger-based assembly of a similar genomic region, 454 assemblies often lead to copies of repeats being pooled into false "consensus" contigs and the creation of shorter contigs in low-copy regions (Wicker et al. 2006). As a result, sequencing information from different sources (including Sanger-based sequences and 454 sequences) often are mixed together to create high-quality *de novo* genomic or transcriptomic assemblies (Cheung et al. 2008; Diguistini et al. 2009).

The 454 sequencing systems use a sequencing-by-synthesis approach, in which a multitude of adapter-flanked DNA fragments are captured at the surface of specifically designed beads and amplified via emulsion PCR (Dressman et al. 2003). After breaking the emulsion, the clonally enriched beads are loaded onto an array of picoliter-scale wells for sequencing. Sequencing is performed via pyrosequencing (Ronaghi 2001) and involves the sequential release of single nucleotide types in a fixed order. The addition of one nucleotide to the extending strand results in the release of pyrophosphate and the immediate generation of a chemiluminescent signal, via ATP sulfurylase and luciferase, that is captured by a charge-coupled device (CCD) camera, and whose intensity is used to determine the number of nucleotides added after each incorporation event. A major drawback of this system relates to homopolymer tracts, which is a contiguous stretch of sequence of the same nucleotide, where more than one nucleotide can be incorporated per cycle. For these homopolymer tracts, the length in bases of the homopolymer must be inferred from the signal intensity which becomes less sensitive as the length of the homopolymer tract increases. As a result, insertion-deletion in homopolymer tracts represents the dominant error type for the 454 system. Longer homopolymers lead to higher error rates, as the number of insertion-deletion errors significantly increases after the 7th base in A/T homopolymer stretches is sequenced (Wicker et al. 2006).

Illumina GAIIx

The latest version of the Illumina sequencing system (referred to as the "Genome Analyzer IIx" or "GAIIx") from Illumina (San Diego, CA) yields more than 160 million sequences whose size typically varies between 36 and 76 bps. Clustered amplicons are generated by immobilizing adapter-flanked DNA fragments on a flow cell surface, coated by a dense lawn of covalently attached primers, whose sequences are complementary to the sequences of the adapters flanking the DNA fragments. A solid phase amplification protocol, also called "bridge amplification" (Fedurco et al. 2006), generates up to 1,000 copies of each DNA fragment, grouped together into "clusters", with densities of up to 10 million clusters per square centimeter. After cluster formation, sequencing is performed using a sequencing-by-synthesis approach with reversible fluorescent terminator deoxyribonucleotides, in which each cycle consists of single-base extension followed by image acquisition after the release of a fluorescent signal whose nature is used to determine the identity of the incorporated base. These nucleotides also contain cleavable "blocking agents" at the 3′ hydroxyl position: chemical moieties allowing the incorporation of one base only during each cycle. Fluorescent labels and blocking agents are cleaved after single-base extension and image acquisition, allowing for the next cycle to occur (Turcatti et al. 2008).

Read lengths can be limited by several noise factors, including phasing, in which imperfections in

single-base extension and cleavage chemistries accumulate at each cycle, leading to: (1) strands of various lengths within a DNA cluster, thus attenuating the signal delivered, (2) signal decay, as fluorescent signal intensities decrease as a function of cycle number and (3) fluorophore cross-talk, which becomes more preponderant in later cycles and can introduce substantial bias towards specific bases (Erlich et al. 2008). All three of these factors increase the probability of base calling errors in long reads. In 2008, Illumina introduced new reagents that improve cleavage chemistry, thereby allowing longer read sequencing (up to 76 bps) and reducing the risk of errors. Currently, average error rates on the Illumina GAIIx platform are on the order of 0.5% (Bentley et al. 2008).

Applied Biosystems SOLiD

This platform developed by Applied Biosystems (Foster City, CA) differs from the two previous platforms in that it uses a sequencing-by-ligation approach (Shendure et al. 2005) to sequence tens to hundreds of millions of adapter-flanked DNA fragments which are immobilized on the surface of 1 μM paramagnetic beads and amplified via emulsion PCR (Dressman et al. 2003). After breaking the emulsion, the clonally amplified beads are selectively recovered and deposited onto a glass slide array for sequencing. A universal primer whose last base is complementary to the *n*th position of the adapter sequence then is hybridized to the array. At each cycle, a set of degenerate fluorescently labeled octamers, whose identity is correlated with the identity of the first 2 bases within the octamer, compete for ligation to the DNA fragment. After ligation and image acquisition, the octamer is chemically cleaved between position 5 and 6 to remove the fluorescent label, and subsequent cycles of octamer ligation, image acquisition and cleavage are performed, enabling sequencing of every 5th base, with the number of cycles determining the overall read length. A second round, in which the elongated primer is removed and a new universal primer whose last base is complementary to the $(n-1)$th position of the adapter sequence is added, is followed by 3 more rounds of primer reset to complete the sequence of the DNA fragment. The SOLiD system features a two-base encoding mechanism, in which two adjacent bases, rather than a single base, are correlated with a fluorescent signal. Because each signal is used to label four possible dinucleotides, the identity of a base is determined by analyzing the 16 dinucleotide combinations of two signals ("color spaces") which result from two successive ligations. As a consequence, each base is interrogated twice, once as the first base and once as the second base, in two independent rounds of primer hybridization. This two-base encoding system allows sequencing errors to be more promptly identified, since a sequencing error would be detected in one ligation reaction only, rather than two successive ones (McKernan et al. 2009).

Dover system polonator

Dover System's Polonator is a second-generation sequencing platform now available to early-access users that was developed from collaboration between George Church's laboratory and the Danaher Corporation (Washington, DC). The most interesting aspect of that platform is its "open source" development model, where all aspects of the system are freely available and modifiable by its users to improve and extend the instrument, potentially enabling faster innovations and lower development costs than other platforms. The Polonator in its present version generates paired-end reads whose cumulative length is 28 bps using bead-based emulsion PCR and sequencing by ligation approach.

Third-generation sequencing platforms

The field of next-generation sequencing is evolving at a breathtaking pace, with current systems rapidly improving to longer reads and higher throughputs. New systems, already tagged with the "third-generation sequencing" moniker, are being developed, promising even longer reads and higher throughput per sample when compared to existing second-generation platforms (Rusk 2009). A major feature of third-generation sequencing systems is the use of single DNA molecules, rather than clustered amplicons, as templates for sequencing, thereby eliminating the risk of phasing errors encountered by second-generation sequencing systems.

One such third-generation system, the HeliScope, commercialized by Helicos BioSciences (Cambridge, MA) relies on an array-based sequencing-by-

synthesis approach (Harris et al. 2008), in which adapter-flanked DNA fragments prepared by poly(A) tailing (not PCR amplification, as required by all previous systems) are hybridized to poly(T) oligomers immobilized on the surface of a flow cell, and sequenced by releasing in a sequential order single types of fluorescently labeled nucleotides. Multiple cycles of single-base extension generate reads whose average length is 25 bps or greater (Harris et al. 2008). By using single DNA molecules as the template for the sequencing reaction, rather than clusters of molecules, a total higher density of sequenced DNA strands on the flow cell surface translates into a higher sequencing output per run when compared to current second-generation sequencing systems. Helicos BioSciences also recently reported the development of a new direct single molecule RNA sequencing technique without prior conversion of messenger RNA to cDNA (Ozsolak et al. 2009). The average read length was around 20 nucleotides while the error rate was approximately 4%. Even though the prototype flow cell used in the experiment generated thousands of reads rather than the 600–800 millions reads generated on the HeliScope, the technique remains highly scalable partly because of a similar average number of reads per unit of surface area on the flow cell.

In addition to single-molecule sequencing, the other major feature of third-generation sequencing systems is the possibility to sequence longer reads than current second-generation sequencing platforms. While relatively short DNA sequences remain a staple of the HeliScope platform, two-third-generation single-molecule sequencing systems, developed by Pacific Biosciences (Menlo Park, CA) and Oxford Nanopore Technologies (Oxford, UK), promise to generate sequences thousands of nucleotides in length. The single-molecule real time (SMRT^{TM}) technology developed by Pacific Biosciences (Eid et al. 2009) uses single DNA polymerase molecules attached to the bottom surface of nanometer-scale holes to incorporate, in a real-time fashion, fluorescently labeled nucleotides to the elongated strand of DNA. The very small diameter of the holes, favoring the natural diffusion of nucleotides and of fluorescent dyes following their cleavage, allows DNA polymerases to incorporate bases in a real-time fashion at a speed of tens per second, and the continuous excitation and detection of fluorescent signal following each incorporation event. In addition, the fluorescent dye is attached to the phosphate backbone rather than the base, allowing its natural real-time release following the incorporation of the nucleotide to the elongating strand. The company's proposed commercial release date for their sequencing instrument is the second half of 2010. The system developed by Oxford Nanopore Technologies relies on protein nanopores (Maglia et al. 2008) for the real-time detection of DNA bases (Clarke et al. 2009). An exonuclease attached to the surface of the nanopore is responsible for capturing DNA strands and cleaving individual bases from the end of DNA strands. As single bases fall into the pore, an engineered protein sensor covalently attached to the inner surface of the nanopore transiently binds the bases as they pass through, blocking an electrical current that runs through the pore and generating a change in conductivity characteristic to each type of bases (A, C, G, T and methylcytosine). While the company has not yet disclosed when the technology will become commercially available, the fact that this system does not require any fluorescent labeling of the DNA or expensive detection system makes it likely that it will be cheaper than other third-generation sequencing systems, and its ability to read methylcytosine offers great promises for epigenomic studies that otherwise could require chemical modification of DNA prior to sequencing.

In 2010, Complete Genomics (Mountain View, CA) plans to deliver genome-wide sequencing services solely for the human genome. This company claims that their technology now enables the sequencing of complete human genomes for $4,400 in consumable costs (Drmanac et al. 2009). Their technology is based on the use of its combinatorial sequencing-by-ligation chemistry on individual DNA "nanoballs". Each nanoball contains two 13 bps and two 26 bps genomic DNA inserts separated by short known sequences used to anneal specific sequencing primers. Their sequencing-by-ligation strategy generates short 10 bp insert sequences in both directions from the known sequences, therefore creating non-contiguous 66 bps sequences (accounting for overlap between inserts) per nanoball in ultra-high density DNA nanoarrays, totaling hundreds of Gbps per run. The single molecule technology developed by VisiGen Biotechnologies (Houston, TX) uses fluorescence resonance energy transfer (FRET) to measure in real-time fashion interactions between a labeled

polymerase and gamma-phosphate labeled nucleotides during DNA strand synthesis. The company had planned to launch a service based around its technology by the end of 2009, and to start selling reagents and instruments in 2011.

Other companies, such as Intelligent Bio-Systems, ZS Genetics, Reveo, LightSpeed Genomics and NABsys, are developing technologies that will become available in the more distant future.

Current applications of next-generation sequencing for single nucleotide polymorphism discovery in plants

Reducing the complexity of a plant genome

The overall size and structure of plant genomes constitute a major obstacle to conventional sequencing methods. The size of plant genomes varies widely relative to mammalian genomes (Arumuganathan and Earle 1991), with estimated total genome size ranging from 98 Mbps for *Fragaria viridis* to ~125 Gbps for *Fritillaria asyriaca*, which covers approximately three orders of magnitude (Grover et al. 2008; http://data.kew.org/cvalues/introduction.html). While plant genes are significantly smaller on average relative to mammalian genes (Sakharkar et al. 2004; Ren et al. 2006), the wide variation of genome size among plant species is due to the occurrence of both polyploidy and the elevated content of highly repetitive DNA sequences primarily due to the expansion of transposable elements (SanMiguel et al. 1996; Rostoks et al. 2002; Bennetzen et al. 2005). The abundance of repetitive elements in plants and their high homology, based upon this rapid transposable element expansion, both represent a significant challenge for whole genome *de novo* assembly of sequencing data and for the alignment of whole genome sequencing data to a reference DNA sequence (Chaisson et al. 2004; Whiteford et al. 2005; Pop and Salzberg 2007; Sundquist et al. 2007).

This problem is further compounded by the typical short lengths of the sequences generated with second-generation sequencing. As a result, researchers have come up with various strategies to reduce the complexity of large plant genomes prior to second-generation sequencing. cDNA libraries (Barbazuk et al. 2007; Trick et al. 2009) are an effective way to target exonic regions of a genome and avoid sequencing repetitive regions (transcribed sequences derived from transposable elements are infrequently observed in cDNA libraries). However, the overall gene representation in cDNA libraries is dependent upon the temporal and spatial patterns of expression that occur during development and in response to environmental stimuli, genes having very low expression require extremely deep rates of sequencing (Moskal et al. 2007). Thus, normalization strategies have been employed for cDNA libraries to reduce the occurrence of highly transcribed sequences and, in the process, enrich the library with sequences having lower transcription rates (Shcheglov et al. 2007). Other gene-enrichment techniques are based on the distinct methylation pattern of plant genomes (Rabinowicz et al. 2003; Rabinowicz et al. 2005). Methylation is a common DNA modification that appears to be ubiquitous in plants and has been shown to be critically important in silencing transposable elements and the regulation of gene development (Martienssen 1998). In plant genomes, a methyl group is covalently attached to the ring structure of a cytosine residue (5-methylcytosine (^{m}C)) (Raleigh and Wilson 1986; Dila et al. 1990), and this modification is observed primarily within ^{m}CpG dinucleotides and ^{m}CpNpG trinucleotides. Inactive transposons display dense methylation while genic sequences generally have comparatively low rates of DNA methylation (Rabinowicz et al. 2005). This preferential methylation of repeats has been used to generate gene-enriched DNA libraries where repetitive elements were largely absent. This enrichment of hypomethylated genic DNA focuses the alignment of short reads to low-copy or genic regions of a given reference genome and minimizes the content of hypermethylated repetitive sequence. In one methyl-filtration technique, genomic shotgun libraries are generated using *E. coli mcrBC*$^{+}$ host strains (Rabinowicz et al. 1999; Palmer et al. 2003). McrBC restricts DNA at (G/A)^{m}C, thus methylated ("repetitive") DNA can be largely excluded from the libraries. A simpler methyl-filtration technique uses the in vitro digestion of genomic DNA with a 5-methylcytosine-sensitive restriction endonuclease followed by the fractionation of the repetitive DNA from the genic sequences by gel electrophoresis (Fellers 2008; Gore et al. 2009). Large undigested fragments will contain large blocks of hypermethylated DNA (primarily transposable element-related

sequences), while smaller fragments will correspond to gene-enriched low copy DNA fractions that are hypomethylated. Another available gene-enrichment approach that has been used primarily with human and other mammalian DNA samples focuses on the use of microarray-based selection of specific loci prior to sequencing with second-generation sequencing platforms (Albert et al. 2007; Hodges et al. 2007; Okou et al. 2007). Reduction in complexity is achieved by direct hybridization capture of segments of the genome onto arrays followed by elution and sequencing of the captured fraction. While this approach can reduce the time and resources required for polymorphism detection (Ng et al. 2009), it relies exclusively on using existing sequencing information for designing custom array oligonucleotides and sequences that are duplicated in a genome cannot be individually targeted. Its effectiveness in using hybridization to capture short DNA fragments from repeat-rich plant genomic DNA also has not yet been established. Finally, other approaches, based on the multiplex amplification of thousands of target sequences in a single tube, have been described in human (Li et al. 2009; Tewhey et al. 2009) but may represent a challenging alternative in plants due to the presence of highly homologous sequences in the genome.

Bioinformatic analysis for variant discovery

Variant discovery with second-generation sequencing technologies requires either the alignment of short reads to a reference sequence (Barbazuk et al. 2007; Hillier et al. 2008; Ossowski et al. 2008; Van Tassel et al. 2008; Wang et al. 2008; Ahn et al. 2009; Amaral et al. 2009; Eck et al. 2009; Gore et al. 2009; Huang et al. 2009; Ramos et al. 2009), or the *de novo* assembly of short reads and comparison of contigs (Novaes et al. 2008; Bundock et al. 2009; Maughan et al. 2009). The scale of the data generated and the short size of the reads often mean that traditional tools used for Sanger sequencing data are not well suited for second-generation sequencing data, and, as a result, a variety of computational tools have been developed for their analysis.

The alignment of short reads to a reference sequence allows the discovery of different types of sequence variations, including single nucleotide polymorphisms (SNPs), short insertion/deletions (indels) and copy number variants (CNVs). The accuracy of read alignment (and the calling of a variant) can vary significantly with the efficiency of basecalling and the presence of indels or erroneous base calls generated during sequencing. Also, the unique architecture of plant genomes often means that a large proportion of short reads will align to several possible genomic locations. As a result, analyses can be greatly facilitated when a reduced (i.e. gene-enriched) version of the genome is sequenced and when read pairs are utilized, as pairing information and the distance between pair mates usually provide additional layers of information to increase the confidence of the alignment. Alignments also can be greatly enhanced when using a combination of factors that include base quality scores and the number of short reads covering a region of interest.

Traditional tools for *de novo* assembly of Sanger sequencing data, such as Phrap (http://www.phrap.org), typically work by capturing all possible overlaps between reads and sorting reads based on different metrics that include length of the overlap and presence of high-quality base discrepancies. Such approaches rapidly become computationally intensive with large next-generation sequencing datasets. Short read lengths may lead to small numbers of short overlaps, making it difficult to separate clusters of repetitive sequences from true overlaps. As a result, most *de novo* assemblers for short reads utilize a different algorithm, based on the de Bruijn graph approach (Pevzner et al. 2001), in which reads are sorted into words of *k*-nucleotides, or *k*-mers, and mapped through a graph, going from one word to the next in a determined order. The use of *k*-mers, and not reads, as the fundamental data structure allows a better handling of naturally redundant next-generation sequencing datasets and a better identification of repetitive sequences.

A variety of tools that are beyond the scope of this paper have been developed for alignment and *de novo* assembly of short reads (for examples, see Warren et al. 2007; Dohm et al. 2007; Jeck et al. 2007; De Bona et al. 2008; Chaisson and Pevzner 2008; Li et al. 2008a; Smith et al. 2008; Zerbino and Birney 2008; Rumble et al. 2009). Also, it must be noted that new software tools are released at a rapid pace and the tools described in this chapter at best represent a snapshot of what is available at this particular

moment. Some tools whose output is suitable for variant discovery applications are described below.

Second-generation sequencing technology providers have developed computational tools adapted to their own sequencing platform. The GS De Novo Assembler, from Roche, is able to align and assemble *de novo* reads generated on the 454 FLX. Roche also developed others tools which can be used for the discovery of variants. The GS Reference Mapper maps reads to a reference sequence and generate a consensus sequence that can be used to view SNPs and short indels (up to 50 bps) when compared to the reference sequence. The GS Amplicon Variant Analyzer aligns reads from amplicons to a reference sequence and identify variants and their respective frequencies in large pools of samples, allowing for the detection of low-frequency (<1%) variants. ELAND, developed by Illumina for its next-generation sequencing platform, aligns short reads to reference sequences provided by the end user. Its output can be linked to the CASAVA software package, a post-processing analysis pipeline developed by Illumina to analyze data from reads aligned to the reference sequence and to identify homozygous or heterozygous SNPs from the alignment.

MAQ (Li et al. 2008b) is a public alignment tool that works with either the Illumina or SOLiD platforms. It performs ungapped alignments to align single short reads and then produces a consensus sequence (referred to as a mapped assembly) by calling each base mapped to the reference sequence. MAQ takes into account the quality scores associated with short reads when producing a consensus sequence and chooses the place in the reference sequence with the minimum sum of the scores of the mismatched bases to align reads and call variants. MAQ also assigns Phred-like scores at each position along the consensus, therefore minimizing the chance of false positives when calling variants, and has the ability to call heterozygotes. In addition to aligning single short reads, MAQ also has the ability to detect short indels with "paired-end" reads by carrying out local Smith-Waterman alignments on unmapped reads when the mate is already mapped.

Bowtie (Langmead et al. 2009) is the first of a new generation of ultra-fast alignment tools that use a unique indexing strategy with a very low memory footprint, known as the Burrows-Wheeler transform (Li and Durbin 2009), to align short reads to large genomes. Bowtie becomes less effective when the best match to the reference genome is an inexact one, and therefore may be mostly geared towards mammalian genome re-sequencing projects. Nevertheless, Bowtie's output format can be converted to the format used by MAQ, allowing faster alignments that can be used for variant detection with MAQ.

PyroBayes (Quinlan et al. 2008) is a modified version of the PolyBayes software and is specifically aimed at basecalling pyrosequencing data generated with the 454 technology. PyroBayes is based on the principle that the native 454 base caller assigns quality scores that are not high enough for SNP calling in low-coverage situations. PyroBayes produces higher (more accurate) quality scores, making more high-quality base calls and therefore allowing more accurate SNP detection in re-sequencing applications.

The ssahaSNP (http://www.sanger.ac.uk/Software/analysis/ssahaSNP) alignment tool detects homozygous SNPS and indels after aligning short reads to a reference genome. Alignments are performed first by aligning matching *k*-mer words to the reference genome, locating the position of the reads in the genome and filtering in the process repetitive elements by ignoring words with high occurrence. Full sequence alignments then are performed and putative SNPs are detected, taking into account quality scores (allocated during sequence alignment) at the SNP position and at the neighboring base positions.

Variant discovery in plants with second-generation sequencing technologies

As noted previously, plant genomic architecture tends to be complicated by the presence of highly repetitive transposable elements and the presence of paralogous sequences, particularly in polyploid species. For those more complex genomes, detection of genetic variants via second-generation sequencing platforms has been focused on either sampling transcript libraries or the hypomethylated fraction of the genome.

By sequencing cDNA libraries generated from a range of individuals and comparing overlapping reads in a pair-wise fashion, SNPs present in exonic sequences can be easily identified. The rate of SNP discovery from cDNA libraries is a product of (1) sequencing depth, (2) frequency of variation within

exonic regions, and (3) temporal and spatial variation of transcription between individuals. Sequencing depth is a primary concern for non-normalized cDNA libraries where the rate of uncovering lower expressed novel transcripts relates to resequencing more highly expressed transcripts. The normalization of cDNA libraries which is intended to remove the wide variations in expression offers a much better rate of novel SNP discovery per kilobase sequenced.

Likewise, enriching for hypomethylated DNA using methylation sensitivity and size selection will reduce the total content of repetitive sequences and focus the sequencing effort on the gene-rich content. The sampling of the hypomethylated sequences will generally comprise a greater total amount of DNA relative to transcript-based sequencing approaches as the intronic and flanking promoter and terminator sequences will be included. These non-coding regions that are associated with genes can provide additional novel variants when compared with exonic, and particularly the coding regions contained within, that can be under negative selection for mutation. Additionally, unlike non-normalized transcript-based sequencing strategies, the hypomethylated fraction of DNA (minus any methylation differences among tissues sampled or across individuals) will be represented in roughly equal proportions in the sequencing library. Sequencing for variant discovery in hypomethylated libraries requires additional sequencing capacity to provide suitable coverage for variant discovery relative to transcript-based approaches given the larger sampling of the genome; however by sequencing the regions flanking exons, the total number of variants discovered and their distribution are expected to be proportionally higher and result in greater genome coverage of identified variations.

The predominant class of genetic variation in plants is the single nucleotide polymorphism (SNP). SNPs offer the highest resolution to create haplotypes and study the association of heritable traits with underlying genetic variation and are also stable over generations making them ideal markers for whole genome analysis or complex trait mapping. Additionally, provided a suitable rate of polymorphism exists within a species, ultra-high density genome wide SNP maps can be created and utilized for whole genome association studies. The recent combination of reduced representation strategies with next generation sequencing platforms has been successful in the rapid and robust identification of SNPs and small indels in a range of plant species with complex genomes.

SNP discovery using 454 sequencing

Zea mays (maize) possesses a rather complex genome due to its recent reduction from an autotetraploid to a functional diploid which, in turn, has produced a large number of highly similar paralogous sequences (Swigonova et al. 2004; Emrich et al. 2007a). Additionally, the majority of the genome is comprised of repetitive sequences, primarily due to the rapid explosion of retrotransposons which are present as extended nested insertions between genic regions (Meyers et al. 2001; SanMiguel and Vitte 2008). Two separate 454 sequencing efforts that reduced the complexity of the maize genome prior to sequencing have been recently described for the purpose of SNP discovery.

Barbazuk et al. (2007) developed a protocol where SNP discovery was performed by 454 sequencing cDNA libraries derived from the laser capture microdissection (LCM) of maize shoot apical meristems (SAMs). Two maize genotypes, B73 and Mo17, were chosen because these genotypes are parents for the publicly available recombinant inbred line (RIL) genetic map and a comprehensive set of gene-enriched genomic assemblies from B73 were available. LCM-derived cDNA libraries from maize were previously shown to have a good diversity of transcripts which would provide a basis for a genome-wide distribution of SNPs (Emrich et al. 2007b). A single 454 GS20 run was performed to sequence the LCM SAM libraries for each genotype. These two separate runs generated a total of 260,887 high quality reads for B73 and 287,917 for Mo17. The strategy for SNP detection was to align the GS20 reads against existing maize B73 genome assemblies to generate multiple sequence alignments (MSAs). Iterative pairwise comparisons among the aligned reads would reveal SNPs in the transcribed sequences.

Coverage is a factor in robust SNP discovery from transcripts using the 454 GS20 platform because the error rate was estimated at the time to be $\sim$1.5% (Emrich et al. 2007b). Thus true SNP discovery is augmented by imposing minimum copy number requirements for GS20 reads ($\geq$2 reads for B73 and

≥3 reads for Mo17) when aligned to the B73 reference (essentially the third B73 read). If SNP discovery were employed using lower coverage, sequencing errors present in single copy reads would be identified as SNPs and thereby reduce the rate of true SNP discovery. A second requirement imposed was that all reads in the MSA were identical within the genotype (monoallelism) but polymorphism was observed between the genotypes. This filtering step removes variation observed within a genotype; these variants are presumably due to the presence of highly similar paralogous sequences within the genome (Emrich et al. 2007a). Using a coverage parameter of three increases the chances of observing the paralog-based variation relative to a coverage requirement of two.

Using these filtering parameters, a total of 2,017 SNPs were found from pairwise comparisons between B73 and Mo17 EST sequences aligned to the B73 reference genome assembly. Sanger-based validation was performed by resequencing 96 SNP-containing locus in B73 and Mo17 and screening for the presence of the variation; a total of 85 SNPs were confirmed (a rate of 88.5%) indicating that the strategy for identifying SNPs from a transcript library is robust.

A second protocol developed by Gore et al. (2009) described the enrichment of the gene space of maize using B73 and Mo17. This strategy relied upon isolating the hypomethylated fraction of the genome by (1) digesting the genome to completion with a restriction enzyme sensitive to the methylation status of the cytosine residue in a CG dinucleotide (*HpaII*—5′-CCGG-3′) and (2) recovery of size selected fragments (between 100 and 600 bp) from an agarose gel. The hypomethylated DNA obtained by size selection was concatenated for subsequent amplification, size selected, and nebulized to generate small insert libraries for sequencing on the 454 GS FLX platform. Several libraries were generated from different tissues and compared against a single unfiltered whole genome shotgun control library to assess the effectiveness of the reduction protocol. The unfiltered library contained 63.1% repetitive content and the three filtered libraries had 3.9, 4.5, and 31.4% repetitive content demonstrating that methylation filtration reduces the repetitive content significantly. This result mirrors previously reported protocols developed for the Sanger biochemistry (Rabinowicz et al. 2003). Using a computational approach to filter out variants derived from highly similar paralogous sequences from these methylation filtered libraries (and hence not true allelic SNPs), a total of 126,683 SNPs were identified between the B73 and Mo17 genotypes. The rate of false SNP discovery rate was found to be ~15% which is consistent with the previously transcript-based discovery effort (Barbazuk et al. 2007; Gore et al. 2009).

The two SNP discovery projects reviewed above were done in the relatively well-characterized species of maize that has extensive genomic resources. However, the vast majority of plant species lack any significant genomic resources. As an example, *Eucalyptus grandis* has practically no genomic resources in either transcripts or genomic sequence. However, this species is agronomically important for wood, pulp and biofuels (FAO 2000). For this SNP discovery effort, equivalent amounts of tissue from 21 different *Eucalyptus* genotypes were pooled together prior to mRNA extraction. The pooled mRNA samples were then used to generate normalized cDNA libraries for sequencing. Unlike maize, where existing reference transcript assemblies or genomic assemblies were used to generate the multiple sequence alignments, for *Eucalyptus*, the individual sequences from the normalized libraries were assembled directly into transcript assemblies and SNP detection done from pairwise comparisons between the assembled reads *in silico*. As with the maize analysis, a minimum coverage requirement of at least two reads aligning to the consensus sequence assembly must contain the varying allele (i.e. the SNP) and at least two reads must contain the consensus allele. Overall, two 454 G20 runs and a single 454 FLX run were performed to generate a total of 148.4 Mbps of EST sequence which were the basis for transcript assemblies. When analyzing the assemblies, 28,652 SNPs (note: no indels were included in this total) were detected and, with an added threshold of the rare allele required to be present >10%, the number of true SNPs was reduced to 23,742. Sanger confirmation of 337 SNP-containing loci found that 279 (82.8%) were validated.

Another SNP discovery project in species lacking robust genomic resources was performed in the highly heterozygous and polyploid genome of sugarcane (Bundock et al. 2009). Sugarcane is a hybrid of *Saccharum officinarum* and *Saccharum spontaneum*,

followed by backcrossing of this hybrid to *S. officinarum* (Bundock et al. 2009). According to the authors, the use of public EST sequences for SNP discovery had been mostly inefficient largely because of the difficulty in generating large numbers of DNA templates with the necessary level of purity for a traditional Sanger resequencing approach. For this particular study, 307 PCR amplifications were performed on the two parents of a sugarcane mapping population (IJ76-514 × Q165), using tentative consensus sequences from sugarcane candidate genes as templates for primer design. Equimolar amounts of all PCR products were pooled together and sequenced at both ends with a 454 FLX sequencing system. On average, more than 90,000 454 reads were obtained for both parents and clustered with the CAP3 program (Huang and Madan 1999). The resulting CAP3 contigs were analyzed for candidate SNPs with the software package PolyBayes (Marth et al. 1999). Under a minimum coverage requirement of 25 reads at a frequency of 4% or more, or 21 reads at a frequency of 5% or more, 1,632 and 1,013 candidates SNPs were discovered in the Q165 and IJ76-514 parents, respectively. 209 candidate SNP sites tested out of 225 (93%) were validated as polymorphic by Sequenom assays.

A recent study demonstrated the use of 454 pyrosequencing in amaranth (*Amaranthus caudatus*), another species lacking significant genomic or transcriptomic resources, to sequence indexed digested DNA fragments from four pooled mapping parents, followed by the *de novo* assembly of each indexed sequencing data sets and the detection of SNPs by way of comparing the resulting contig sequences (Maughan et al. 2009). The complexity of the genome first was reduced by double-digesting genomic DNA with the *EcoRI* and *BfaI* restriction endonucleases, ligating restriction site-specific adaptors to the DNA fragments then removing via biotin-streptavidin paramagnetic bead separation DNA fragments containing the *BfaI* restriction sites at both ends. Indexed barcodes then were added to the remaining fragments via PCR and the resulting amplicons were size-selected via gel electrophoresis and sequenced on the 454 FLX platform. For each mapping parent, the resulting sequences were assembled with the GS De Novo Assembler. SNPs were detected in large (>200 bps) contigs if (1) the coverage depth at the SNP position was 10× or higher; (2) the minor allele represented at least 30% of all alleles observed; and (3) 90% of the alleles were associated to a unique barcode sequence. Under those conditions, the number of SNPs detected by pairwise comparison of each mapping parents varied from 140 to 11,047. Sanger confirmation of 35 of those SNPs indicated that 34 (97%) were validated.

These results demonstrate that robust and rapid SNP discovery strategies can be employed in a species without any existing genomics resources and the longer reads of the 454 FLX will augment the ability to directly assemble reads into genomic or transcript assemblies for intra-species or inter-species sequence comparison.

SNP discovery using Illumina sequencing

The read lengths produced by the Illumina platform are significantly shorter than those of the 454 GS20 or FLX system. However, this platform generates a far greater total of sequence per run. The discovery of new SNPs can be facilitated when extensive genomic or EST sequence data are available. Short next-generation sequencing reads can be aligned to a reference sequence and variants are detected by comparing the reads to the reference genotype. The Illumina sequencing system is particularly well-suited for such large-scale variant discovery projects mostly because of its high sequencing throughput that potentially can result in multiple overlapping reads and a confident assignment for a given variant. As a result, the Illumina platform has been used for a variety of genome-wide variant discovery projects (Hillier et al. 2008; Ossowski et al. 2008; Van Tassel et al. 2008; Wang et al. 2008; Ahn et al. 2009; Amaral et al. 2009; Eck et al. 2009; Huang et al. 2009; Ramos et al. 2009), where a high quality reference genome sequence was available.

In the absence of extensive genomic or EST resources, genome-wide SNP detection can be envisioned, in which a set of reference sequences is created by assembling *de novo* Illumina sequencing reads and aligning individual Illumina reads to the resulting contigs (Kerstens et al. 2009). *De novo* assembly of genomes of a smaller total size (e.g. some bacterial species) or large insert clones (e.g. bacterial artificial chromosomes) is possible with sufficient coverage (Pop and Salzberg 2007). Additionally, if a species has a compact genome with

relatively low rates of repetitive content, then the whole genome can be shotgun sequenced with acceptable rates of sequencing loss to tags that map to multiple locations. However, for larger genomes that have a higher content of repetitive sequences and the presence of paralogous sequences, whole genome assembly with short reads is not tractable, and alternative resequencing approaches for uncovering variation must be explored.

The *Arabidopsis thaliana* (Arabidopsis) genome is quite compact among plant species with an estimated size of 125 Mb and the total repetitive content of the genome is <10% (The Arabidopsis Genome Initiative 2000). Three genotypes (i.e. Col-0, Bur-0, Tsu-1) of Arabidopsis were deeply resequenced using a whole genome shotgun strategy to ascertain relative rates of variation to the reference whole genome assembly for Arabidopsis (constructed from the Col-0 genotype). The estimated coverage after filtering out low quality reads and contamination by organellar DNA ranged from 15-fold to 25-fold. Similar to an initial shotgun strategy for *Caernohabditis elegans*, alignment of short reads from each of the genotypes to the reference assembly could be performed to easily identify genetic variation (Hillier et al. 2008). Interestingly, the short reads generated from the Col-0 accession did find a limited amount of variation (1,172 SNPs and 1,287 indels) relative to the reference Col-0 assembly derived from a BAC-by-BAC resequencing strategy. By varying filtering strategies, these differences were attributed to either subtle variation between the individuals within this accession or actual mistakes in the whole genome assembly. This ability to use high rates of coverage to uncover very low rates of polymorphism was also reported by the authors to uncover causal SNPs in EMS-mutagenized lines of *Arabidopsis* as well as opportunities to resolve areas of localized polymorphism relative to a reference assembly. The ability to reveal very low rates of variation potentially has interesting implications for uncovering induced or existing variation in TILLING projects. For the other two *Arabidopsis* accessions when compared with the Col-0 accession, the deep coverage of reads allowed for robust SNP discovery with 823,325 non-redundant SNPs and 79,961 non-redundant 1–3 bp indels. These variants can be combined with studies on the phenotypic variation in these accessions for functional gene discovery.

Rice, like *Arabidopsis*, has a reference genome assembly created by a BAC-by-BAC approach for the cultivar *Nipponbare*. Additionally, a whole genome shotgun assembly exists for the 93-11 Chinese super-hybrid cultivar. A set of 150 F_{11} Single Seed Descent Recombinant Inbred Lines (SSDRILs) derived from a cross between *Nipponbare* and 93-11 was whole genome shotgun sequenced on an Illumina Genome Analyzer using indexed libraries (Huang et al. 2009). A 5′ 3 bp index code was used to assign any given read to its cognate RIL and the next 33 bp of the short reads were then mapped to both of the parental genome sequences for SNP detection. Given that the SSD RILs are expected to be well over 99% homozygous for any given locus, the sequence in tag can be assigned unambiguously to either parent (or both if the tag is monomorphic between the genotypes). Using this strategy of mapping SNPs from the short reads onto the physical assemblies, the recombination break-points in the RILs can be observed in much greater detail when compared with using genotypic markers (i.e. SNP markers, SSRs, RFLPs, etc.). As expected with SSD RILs, for any given chromosomal region, SNPs were predominantly from one of the parents while a few SNPs scattered within from the other parent. These scattered SNPs were hypothesized to be primarily the products of sequencing errors – which is consistent with the elevated rates of sequencing errors in next-generation sequencing systems relative to the conventional Sanger platform. This direct genotyping method using short reads and SNP assignment allowed the authors to characterize the recombination breakpoints to within 40 kb (on average) and the authors reported that the actual sequence run time for this analysis took approximately 2 weeks. When compared with a previous analysis of the same RIL population, these data provided a 35-fold increase in resolution of the recombination breakpoints.

In contrast to these two whole genome approaches which make use of a rather complete reference assembly, the Illumina sequencing platform has also been successfully used to identify SNPs in *Brassica napus* (oilseed rape), which is an allotetraploid species within the *Brassicaceae* family and has limited genomic resources (Trick et al. 2009a). Oilseed rape is the product of a spontaneous hybridization of the "A" genome from *B. rapa* and the "C" genome from *B. oleracea* (Palmer et al. 1983; Parkin

et al. 1995). Thus, within *B. napa*, polymorphism can exist between cultivars on homologs (inter-cultivar polymorphism) or variation can be observed within the individual on the homeologs (inter-homeolog polymorphism, or "hemi-SNP"). Short read ESTs from two *B.* napus cultivars were generated from non-normalized leaf cDNA libraries and were aligned via MAQ to a set of ~94,000 transcript assemblies derived from a set of ~810,000 ESTs from a range of diverse *Brassica* species (Trick et al. 2009b). These two *B. napus* cultivars chosen for sequencing and SNP identification were known from RFLP analysis to be relatively divergent and were both available as doubled haploid lines, i.e. they are completely homozygous. Under the most stringent filtering parameter, the aligned short reads were then required to have a minimum of 8-fold coverage for each cultivar at any given base prior to intra- and inter-cultivar SNP discovery. If a polymorphism is observed when comparing the reads from within a cultivar, this was indicative of the SNP being observed between two homeologous sequences, one on the "A" genome and the other on the "C" genome, and termed a hemi-SNP. Conversely, a simple SNP would have all reads monomorphic for one cultivar and all reads from the other cultivar as monomorphic but those reads would be polymorphic with respect to each other. For this analysis, a total of 23,330 putative SNPs were associated with 9,265 unigenes with the vast majority, 21,259 (91.2%) being hemi-SNPs and 2,071 inter-cultivar SNPs.

Conclusions

The advent of the second-generation sequencing platforms has now allowed for the ability to read billions of base pairs of sequence within a single or limited number of runs. These sequences are generated in a massively parallel fashion from tens or hundreds of millions of short reads (25–500 bp). For genomes where there are rather good genomic resources including a reference assembly or tens of thousands of transcript assemblies derived from very deep EST sequencing, tag alignment to these references is a straight-forward process and pairwise comparisons among tags can allow for easy variant discovery. For those species lacking significant genomic resources or are more complex (i.e. polyploidy), second-generation sequencing can be used to generate EST reads and subsequent transcript assemblies and identify genetic variations allowing for the rapid development of genomic resources from genic sequences. While this massive increase in the capture of data has come with a drastic reduction in the costs of acquiring the sequence, the quality of the sequence in any given read is more variable than the "gold standard" of Sanger based sequence. This variability in quality is primarily addressed by ensuring deep coverage – in each of the applications for SNP discovery, a threshold for minimum coverage was imposed to minimize the discovery of "false SNPs", or error generated during the creation of the sequence. These coverage requirements directly relate to the component of the genome that is being sequenced. For simpler genomes in the Plant Kingdom, shotgun sequencing is entirely appropriate due to the relatively low rate of repetitive content. Conversely, for genomes with a high repetitive content, in order to generate a sufficient coverage for SNP discovery, reduction strategies such as normalized cDNAs or screening for hypomethylation can focus sequencing resources on gene-rich regions. The possibility of developing ultra-dense sets of SNPs for whole genome association studies or rapid marker development for complex QTL mapping now seems feasible with the advent of the next-generation sequencing platforms and associated bioinformatics tools.

References

Ahn SM, Kim TH, Lee S et al (2009) The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. Genome Res 19:1622–1629

Albert TJ, Molla MN, Muzny DM et al (2007) Direct selection of human genomic loci by microarray hybridization. Nat Methods 4:903–905

Altshuler D, Pollara VJ, Cowles CR et al (2000) An SNP map of the human genome generated by reduced representation shotgun sequencing. Nature 407:513–516

Amaral AJ, Megens HJ, Kerstens HHD et al (2009) Application of massive parallel sequencing to whole genome SNP discovery in the porcine genome. BMC Genomics 10:374

Arumuganathan K, Earle ED (1991) Nuclear DNA content of some important plant species. Plant Mol Biol Rep 9:208–218

Barbazuk WB, Emrich SJ, Chen HD, Li L, Schnable PS (2007) SNP discovery via 454 transcriptome sequencing. Plant J 51:910–918

Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K (2007) High-resolution profiling of histone methylations in the human genome. Cell 129:823–837

Bennetzen JL, Ma J, Devos KM (2005) Mechanisms of recent genome size variation in flowering plants. Ann Bot 95:127–135

Bentley DR, Balasubramanian S, Swerdlow HP et al (2008) Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456:53–59

Buetow KH, Edmonson MN, Cassidy AB (1999) Reliable identification of large numbers of candidate SNPs from public EST data. Nat Genet 21:323–325

Bundock PC, Eliott FG, Ablett G, Benson AD, Casu RE, Aitken KS, Henry RJ (2009) Targeted single nucleotide polymorphism (SNP) discovery in a highly polyploidy plant species using 454 sequencing. Plant Biotechnol J 7:347–354

Chaisson MJ, Pevzner PA (2008) Short read fragment assembly of bacterial genomes. Genome Res 18:324–330

Chaisson M, Pevzner P, Tang H (2004) Fragment assembly with short reads. Bioinformatics 20:2067–2074

Cheung F, Win J, Lang JM, Hamilton J, Vuong H, Leach JE, Kamoun S, Levesque AC, Tisserat N, Buell CR (2008) Analysis of the Pythium ultimum transcriptome using Sanger and pyrosequencing approaches. BMC Genomics 9:542

Choi IY, Hyten DL, Matukimalli LK et al (2007) A soybean transcript map: gene distribution, haplotype and single-nucleotide polymorphism analysis. Genetics 176:685–696

Clarke J, Wu HC, Jayasinghe L, Patel A, Reid S, Bayley H (2009) Continuous base identification for single molecule nanopore DNA sequencing. Nat Nanotechnol 4:265–270

Clifton SW, Mitreva M (2009) Strategies for undertaking expressed sequence tag (EST) projects. Methods Mol Biol 533:13–32

Cokus SJ, Feng S, Zhang X, Chen Z, Merriman B, Haudenschild CD, Pradhan S, Nelson SF, Pellegrini M, Jacobsen SE (2008) Shotgun bisulphate sequencing of the *Arabidopsis* genome reveals DNA methylation patterning. Nature 452:215–219

De Bona F, Ossowski S, Schneeberger K, Rätsch G (2008) Optimal spliced alignments of short sequence reads. Bioinformatics 24:i174

Diguistini S, Liao NY, Platt D et al (2009) De novo genome sequence assembly of a filamentous fungus using Sanger, 454 and Illumina sequence data. Genome Biol 10:R94

Dila D, Sutherland E, Moran L, Slatko B, Raleigh EA (1990) Genetic and sequence organization of the mcrBC locus of *Escherichia coli* K-12. J Bacteriol 172:4888–4900

Dohm JC, Lottaz C, Borodina T, Himmelbauer H (2007) SHARCGS, a fast and highly accurate short-read assembly algorithm for de novo genomic sequencing. Genome Res 17:1697–1706

Dressman D, Yan H, Traverso G, Kinzler KW, Vogelstein B (2003) Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. Proc Natl Acad Sci 100:8817–8822

Drmanac R, Sparks AB, Callow MJ et al. (2009) Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. Science doi:10.1126/Science.1181498

Duran C, Appleby N, Clark T et al (2009) AutoSNPdb: an annotated single nucleotide polymorphism database for crop plants. Nucleic Acids Res 37:D951–D953

Eck SH, Benet-Pages A, Flisikowski K, Meitinger T, Fries R, Strom TM (2009) Whole genome sequencing of a single *Bos taurus* animal for single nucleotide polymorphism discovery. Genome Biol 10:R82

Eid J, Fehr A, Gray J et al (2009) Real-time DNA sequencing from single polymerase molecules. Science 323:133–138

Emrich SJ, Li L, Wen T-J et al (2007a) Nearly identical paralogs: Implications for maize (*Zea mays* L.) genome evolution. Genetics 175:429–439

Emrich SJ, Barbazuk WB, Li L, Schanble PS (2007b) Gene discovery and annotation using LCM-454 transcriptome sequencing. Genome Res 17:69–73

Erlich Y, Mitra PP, delaBastide M, McCombie WR, Hannon GJ (2008) Alta-cyclic: a self-optimizing base caller for next-generation sequencing. Nat Methods 5:679–682

Ewing B, Green P (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res 8:186–194

Fahlgren N, Howell MD, Kasschau KD et al (2007) High-throughput sequencing of *Arabidopsis* microRNAs: evidence for frequenct birth and death of MIRNA genes. PLoS ONE 2:e219

FAO (2000) Global forest resources assessment 2000—Main report. FAO Forestry Paper 140

Fedurco M, Romieu A, Williams S, Lawrence I, Turcatti G (2006) BTA, a novel reagent for DNA attachement on glass and efficient generation of solid-phase amplified DNA colonies. Nucleic Acids Res 34:e22

Fellers JP (2008) Genome filtering using methylation-sensitive restriction enzymes with six base pair recognition sites. Plant Genome 1:146–152

Goff SA, Ricke D, Lan TH et al (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). Science 296:92–100

Gore MA, Wright MH, Ersoz ES et al (2009) Large-scale discovery of gene-enriched SNPs. Plant Genome 2:121–133

Grover CE, Hawkins JS, Wendel JF (2008) Phylogenetic insights into the pace and pattern of plant genome size evolution. In: Volff J-N (ed) Plant genomes. Karger, Basel, pp 57–68

Harris TD, Buzby PR, Babcock H et al (2008) Single-molecule DNA sequencing of a viral genome. Science 320:106–109

Hillier LW, Marth GT, Quinlan AR et al (2008) Whole-genome sequencing and variant discovery in *C. elegans*. Nat Methods 5:183–188

Hodges E, Xuan Z, Balija V et al (2007) Genome-wide in situ exon capture for selective resequencing. Nat Genet 39:1522–1527

Huang X, Madan A (1999) CAP3: a DNA sequence assembly program. Genome Res 9:868–877

Huang X, Feng Q, Qian Q et al (2009) High-throughput genotyping by whole-genome resequencing. Genome Res 19:1068–1076

Hunkapiller T, Kaiser RJ, Koop BF, Hood L (1991) Large-scale and automated DNA sequence determination. Science 254:59–67

Initiative TheArabidopsisGenome (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408:796–815

IRGSP (2005) The map-based sequence of the rice genome. Nature 436:793–800

Jaillon O, Aury JM, Noel B et al (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature 449:463–467

Jeck WR, Reinhardt JA, Baltrus DA, Hickenbotham MT, Magrini V, Mardis ER, Dangl JL, Jones CD (2007) Extending assembly of short DNA sequences to handle error. Bioinformatics 23:2942–2944

Johnson DS, Mortazavi A, Myers RM, Wold B (2007) Genome-wide mapping of in vitro protein-DNA interactions. Science 316:1497–1502

Kasschau KD, Fahlgren N, Chapman EJ, Sullivan CM, Cumbie JS, Givan SA, Carrington JC (2007) Genome-wide profiling and analysis of *Arabidopsis* siRNAs. PLoS Biol 5:e57

Kerstens HHD, Crooijmans RPMA, Veenendaal A, Dibbits BW, Chin-A-Woeng TFC, den Dunnen JT, Groenen MAM (2009) Large scale single nucleoptide polymorphism discovery in unsequenced genomes using second generation high throughput sequencing technology: applied to turkey. BMC Genomics 10:479

Lander ES, Linton LM, Birren B et al (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921

Langmead B, Trapnell C, Pop M, Saltzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10:R25

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–1760

Li R, Li Y, Kristiansen K, Wang J (2008a) SOAP: short oligonucleotide alignment program. Bioinformatics 24:713–714

Li H, Ruan J, Durbin R (2008b) Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res 18:1851–1858

Li JB, Gao Y, Aach J et al (2009) Multiplex padlock targeted sequencing reveal human hypermutable CpG variations. Genome Res 19:1606–1615

Lijavetzky D, Cabezas JA, Ibanez A, Rodriguez V, Martinez-Zapater JM (2007) High throughput SNP discovery and genotyping in grapevine (*Vitis vinifera* L.) by combining a re-sequencing approach and SNPlex technology. BMC Genomics 8:424

Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, Ecker JR (2008) Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. Cell 133:523–536

Lu C, Jeong DH, Kulkarni K et al (2008) Genome-wide analysis for discovery of rice microRNAs reveals natural antisense microRNAs (nat-miRNAs). Proc Natl Acad Sci 105:4951–4956

Luckey JA, Drossman H, Kostichka AJ, Mead DA, D'Cunha J, Norris TB, Smith LM (1990) High speed DNA sequencing by capillary electrophoresis. Nucleic Acids Res 18:4417–4421

Maglia G, Restrepo MR, Mikhailova E, Bayley H (2008) Enhanced translocation of single DNA molecules through alpha-hemolysin nanopores by manipulation of internal charge. Proc Natl Acad Sci 105:19720–19725

Margulies M, Egholm M, Altman WE et al (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437:376–380

Marth GT, Korf I, Yandell MD et al (1999) A general approach to single-nucleotide polymorphism discovery. Nat Genet 23:452–456

Martienssen RA (1998) Transposons, DNA methylation and gene control. Trends Genet 14:263–264

Maughan PJ, Yourstone SM, Jellen EN, Udall JA (2009) SNP discovery via genomic reduction, barcoding, and 454 pyrosequencing in amaranth. Plant Genome 2:260–270

McKernan KJ, Peckham HE, Costa G et al. (2009) Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two base encoding. Genome Res doi:10.1101/gr.091868.109

Meyers BC, Tingey SV, Morgante M (2001) Abundance, distribution, and transcriptional activity of repetitive elements in the maize genome. Genome Res 11:1660–1676

Mortazavi A, Williams B, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 5:621–628

Moskal WA Jr, Wu HC, Underwood BA, Wang W, Town CD, Xiao Y (2007) Experimental validation of novel genes predicted in the unannotated regions of the Arabidopsis genome. BMC Genomics 8:18

Ng SB, Turner EH, Robertson PD et al (2009) Targeted capture and massively parallel sequencing of 12 human exomes. Nature 461:272–276

Nobuta K, Lu C, Shrivastava R et al (2008) Distinct size distribution of endogenous siRNAs in maize: evidence from deep sequencing in the mop1–1 mutant. Proc Natl Acad Sci 105:14958–14963

Novaes E, Drost DR, Farmerie WG, Pappas GJ Jr, Grattapaglia D, Sederoff RR, Kirst M (2008) High-throughput gene and SNP discovery in *Eucalyptus grandis*, an uncharacterized genome. BMC Genomics 9:312

Okou DT, Steinberg KM, Middle C, Cutler DJ, Albert TJ, Zwick ME (2007) Microarray-based genomic selection for high-throughput resequencing. Nat Methods 4:907–909

Ossowski S, Schneeberger K, Clark RM et al (2008) Sequencing of natural strains of *Arabidopsis thaliana* with short reads. Genome Res 18:2024–2033

Ozsolak F, Platt AR, Jones DR et al (2009) Direct RNA sequencing. Nature 461:814–818

Palmer JD, Shields CR, Cohen DB, Orton TJ (1983) Chloroplast DNA evolution and the origin of amphidiploid *Brassica* species. Theor Appl Genet 65:181–189

Palmer LE, Rabinowicz PD, O'Shaughnessy AL et al (2003) Maize genome sequencing by methylation filtration. Science 302:2115–2117

Parkin IAP, Sharpe PAG, Keith DJ, Lydiate DJ (1995) Identification of the A and C genomes of the amphidiploid *Brassica napus* (oilseed rape). Genome 38:1122–1131

Parkinson J, Blaxter M (2009) Expressed sequence tags: an overview. Methods Mol Biol 533:1–12

Paterson AH, Bowers JE, Bruggmann R et al (2009) The Sorghum bicolor genome and the diversification of grasses. Nature 457:551–556

Pevzner PA, Tang H, Waterman MS (2001) An Eulerian path approach to DNA fragment assembly. Proc Natl Acad Sci 98:9748–9753

Picoult-Newberg L, Ideker TE, Pohl MG, Taylor SL, Donaldson MA, Nickerson DA, Boyce-Jacino M (1999) Mining SNPs from EST databases. Genome Res 9:167–174

Pop M, Salzberg SL (2007) Bioinformatics challenges of new sequencing technology. Trends Genet 24:142–149

Prober JM, Trainor GL, Dam RJ, Hobbs FW, Robertson CW, Zagursky RJ, Cocuzza AJ, Jensen MA, Baumeister K (1987) A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. Science 238:336–341

Pushkarev D, Neff NF, Quake SR (2009) Single-molecule sequencing of an individual human genome. Nat Biotechnol 27:847–852

Quinlan AR, Stewart DA, Stramberg MP, Marth GT (2008) PyroBayes: an improved base caller for SNP discovery in pyrosequences. Nat Methods 5:179–181

Rabinowicz PD, Schutz K, Dedhia N, Yordan C, Parnell LD, Stein L, McCombie WR, Martienssen RA (1999) Differential methylation of genes and retrotransposons facilitates shotgun sequencing of the maize genome. Nat Genet 23:305–308

Rabinowicz P, McCombie WR, Martienssen RA (2003) Gene enrichment in plant genomic shotgun libraries. Curr Opin Plant Biol 6:150–156

Rabinowicz PD, Citek R, Budiman MA et al (2005) Differential methylation of genes and repeats in land plants. Genome Res 15:1431–1440

Raleigh EA, Wilson G (1986) *Escherichia coli* K-12 restricts DNA containing 5-methylcytosine. Proc Natl Acad Sci 83:9070–9074

Ramos AM, Crooijmans RPMA, Affara AJ et al (2009) Design of a high density SNP genotyping assay in the pig using SNPs identified and characterized by next generation sequencing technology. PLoS ONE 4:e6524

Ren X-Y, Vorst O, Fiers MWEJ et al (2006) In plants, highly expressed genes are the least compact. Trends Genet 22:528–532

Roe BA (2004) Shotgun library construction for DNA sequencing. Methods Mol Biol 255:171–187

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

Rostoks N, Park YJ, Ramakrishna W et al (2002) Genomic sequencing reveals gene content, genomic organization, and recombination relationships in barley. Funct Integr Genomics 2:51–59

Rostoks N, Mudie S, Cardle L et al (2005) Genome-wide SNP discovery and linkage analysis in barley based on genes responsive to abiotic stress. Mol Genet Genomics 274:515–527

Rumble SM, Lacroute P, Dalca AV, Fiume M, Sidow A, Brudno M (2009) SHRIMP: accurate mapping of short color-space reads. PLoS Comput Biol 5:e10000386

Rusk N (2009) Cheap third-generation sequencing. Nat Methods 6:244–245

Sachidanandam R, Weissman D, Schmidt SC et al (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 409:928–933

Sakharkar MK, Chow VTK, Kangueane P (2004) Distributions of exons and introns in the human genome. In Silico Biol 4:387–393

Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci 74:5463–5467

SanMiguel P, Vitte C (2008) The LTR-retrotransposons of maize. In: Bennetzen JL, Hake S (eds) Handbook of maize-volume II: domestication, genetics and genomics. Springer, Netherlands, p 307

SanMiguel P, Tikhonov A, Jin YK et al (1996) Nested retrotransposons in the intergenic regions of the maize genome. Science 274:765–768

Shcheglov AS, Zhulidov PA, Bogdanova EA, Shagin DA (2007) Normalization of cDNA libraries. In: Buzdin AA, Lukyanov SA (eds) Nucleic acids hybridizations: modern applications. Springer, Netherlands, pp 97–102

Shendure J, Ji H (2008) Next-generation DNA sequencing. Nat Biotechnol 26:1135–1145

Shendure J, Porreca GJ, Reppas NB et al (2005) Accurate multiplex polony sequencing of an evolved bacterial genome. Science 309:1728–1732

Smailus DE, Marziali A, Dextras P et al (2006) Simple, robust methods for high-throughput nanoliter-scale DNA sequencing. Genome Res 15:1447–1450

Smith LM, Sanders JZ, Kaiser RJ, Hughes P, Dodd C, Connell CR, Heiner C, Kent SB, Hood LE (1986) Fluorescence detection in automated DNA sequence analysis. Nature 321:674–679

Smith AD, Xuan Z, Zhang MQ (2008) Using quality scores and longer reads improves accuracy of Solexa read mapping. BMC Bioinformatics 9:128

Sultan M, Schulz MH, Richard H et al (2008) A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. Science 321:956–960

Sundquist A, Ronaghu M, Tang H et al (2007) Whole-genome sequencing and assembly with high-throughput, short-read technologies. PLoS One 2:484

Sunkar R, Zhou X, Zheng Y, Zhang W, Zhu JK (2008) Identification of novel and candidate miRNAs in rice by high throughput sequencing. BMC Plant Biol 8:25

Swerdlow H, Gesteland R (1990) Cappilary gel electrophoresis for rapid, high resolution DNA sequencing. Nucleic Acids Res 18:1415–1419

Swigonova Z, Lai J, Ma J et al (2004) Close split of sorghum and maize genome progenitors. Genome Res 14:1916–1923

Tewhey R, Warner JB, Nakano M et al. (2009) Microdroplet-based PCR enrichment for large-scale targeted sequencing. Nat Biotechnol doi:10.1038/nbt.1583

Trick M, Long Y, Meng J, Bancroft I (2009a) Single nucleotide polymorphism (SNP) discovery in the polyploidy *Brassica napus* using Solexa transcriptome sequencing. Plant Biotechnol J 7:334–346

Trick M, Cheung F, Drou N, Fraser F, Lobenhofer EK, Hurban P, Magusin A, Town CD, Bancroft I (2009b) A newly-developed community microarray resource for transcriptome profiling in *Brassica* species enables the

confirmation of *Brassica*-specific expressed sequences. BMC Plant Biol 9:50

Turcatti G, Romieu A, Fedurco M, Tairi AP (2008) A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis. Nucleic Acids Res 36:e25

Tuskan GA, DiFazio S, Jansson S et al (2006) The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). Science 313:1596–1604

Useche FJ, Gao G, Hanafey M, Rafalski A (2001) High-throughput identification, database storage and analysis of SNPs in EST sequences. Genome Inform 12:194–203

Van Tassel CP, Smith TPL, Matukumalli LK et al (2008) SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. Nat Methods 5:247–252

Velasco R, Zharkikh A, Troggio M et al (2007) A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. PLoS ONE 2:e1326

Wang J, Wang W, Li R et al (2008) The diploid sequence of an Asian individual. Nature 456:60–65

Warren RL, Sutton GG, Jones SJ, Holt RA (2007) Assembling millions of short DNA sequences using SSAKE. Bioinformatics 23:500–501

Wheeler DA, Srinivasan M, Egholm M et al (2008) The complete genome of an individual by massively parallel DNA sequencing. Nature 452:872–876

Whiteford N, Haslam N, Weber G et al (2005) An analysis of the feasibility of short read sequencing. Nucleic Acids Res 33:e171

Wicker T, Schlagenhauf E, Graner A, Close TJ, Keller B, Stein N (2006) 454 sequencing put to the test using the complex genome of barley. BMC Genomics 7:275

Wilhelm B, Marguerat S, Watt S, Schubert F, Wood V, Goodhead I, Penkett C, Rogers J, Bähler J (2008) Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. Nature 453:1239–1243

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

Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18:821–829