3

Sequencing Technologies and Their Use in Plant Biotechnology and Breeding

Victor Llaca
Dupont Agricultural Biotechnology, Pioneer Hi-Bred International,
Wilmington, Delaware,
USA

1. Introduction

The development of DNA sequencing strategies has been a high priority in genetics research since the discovery of the structure of DNA and the basic molecular mechanisms of heredity. However, it was not until the works by Maxam and Gilbert (1977), and Sanger (Sanger *et al*, 1977), that the first practical sequencing methods were developed and implemented on a large scale. The first isolation and sequencing of a plant cDNA by Bedbrook and colleagues a few years later initiated the field of Plant Molecular Genetics (Bedbrook *et al,* 1980). Plant biotechnology started shortly thereafter with the successful integration of recombinant DNA and sequencing techniques to generate the first transgenic plants using *Agrobacterium* (Fraley *et al,* 1983; Herrera-Estrella *et al,* 1983). The first genetic map in plants based on restriction fragment length polymorphisms (RFLPs; Bernatzky & Tanksley, 1986) enabled the capture of genetic variation and started the era of molecular marker-assisted plant breeding. Since then, sequencing methodologies have been essential tools in plant research. They have allowed the characterization and modification of genes and metabolic pathways, as well as the use of genetic variation for studies in species diversity, marker-assisted selection (MAS), germplasm characterization and seed purity. The determination of the reference genomes in *Arabidopsis thaliana*, rice and maize using Sanger sequencing strategies constituted major milestones that enabled the analysis of genome architecture and gene characterization in plants (The Arabidopsis Genome Initiative, 2001; International Rice Genome Project, 2005; Schnable *et al,* 2009). More recently, the development and increasing availability of multiple Next-Generation sequencing (NGS) technologies minimized research limitations and bottlenecks based on sequence information (Metzker, 2010; Glenn, 2011). It is difficult to overstate the influence that these massively parallel systems have had in our understanding of plant genomes and in the expansion, acceleration and diversification of breeding and biotechnology projects. At the same time, this influence tends to understate the importance that capillary Sanger sequencing still has in day-by-day research and development work. This review provides a description of major sequencing technologies that are available today, their use as well as future prospects in basic plant genetics research, biotechnology and breeding in crop plants.

2. Current sequencing technologies

The development of recent sequencing technologies has generated a remarkable increase, by orders of magnitude, in sequencing throughput with a corresponding drop in cost per base. A simple exercise to comprehend the scale of acceleration in sequencing is to look back at the state of the art of sequencing in 1980. At that time, earlier improvements in Sanger and Maxam-Gilbert methodologies had initiated the wide use of sequencing in research laboratories around the world. Then, typical sequencing throughputs per slab gel run were under 10,000 bp. During the period from 1980 to 2005 sequencing platforms based on Sanger chemistry had a 500 to 1,000-fold increase, to more than 5 Mbp per run. The number of reads that could be processed, quality, read length and analysis all improved and were optimized, propelled by the development of the human genome project (Barnhart, 1989). While these technological advances were certainly impressive, they dwarf when compared to the acceleration in sequencing capacity after 2005. At that time, novel ultra-high throughput technologies started to become commercially available. From 2005 through the second half of 2011, the throughput per run had increased an additional 100,000-fold, or 5 orders of magnitude. This acceleration has been unprecedented in science and technology. It has outpaced Moore's law that famously predicted that the number of transistors in a computer processor would double every two years (Moore, 1965, Figure 1). This fast increase in sequencing capacity has had important consequences in analysis and logistics, and has changed expectations in all aspects of plant genetics, breeding and biotechnology.

The new chemistries and platforms, broadly described as Next-Generation sequencing (NGS) technologies, take advantage of diverse chemistries and detection approaches. While some of these technologies appear to have little in common with each other, they share key characteristics. All NGS technologies are massively parallel systems relying on the immobilization of millions of, up to billions of DNA templates in a solid surface. They do not use electrophoresis, relying instead on *in situ* base detection and extension. With the exception of one system, developed by Helicos, NGS platforms need to amplify the templates and use one of several PCR-based approaches. One additional characteristic of NGS systems that took more than one early-adopting institution unprepared, is the increased need for computer power and storage necessary to process and retain the massive data produced. Currently there are 5 companies commercializing one or more NGS platforms. However, there are only three NGS technologies, Roche 454, Illumina and ABI SOLiD, that account for the vast majority of usage in plant research and are widely available in academic institutions, private research centers and service-providing companies. As it will be emphasized in the next sections, these platforms have different input and output characteristics that make them more or less advantageous to specific applications. Finally, one 'Third Generation' sequencing platform has recently become commercially available from Pacific Biosciences. Third generation technologies are also massively parallel systems although they use single-molecule DNA templates, real-time detection and are able to generate longer reads faster. The expectation is that third generation machines will eventually produce large numbers of high-quality reads with an average of several kilobase-pairs from a single molecule.

This section provides a brief description and comparison of the most relevant sequencing technologies available in plant biotechnology and breeding. For additional, in-depth technical information, the reader is advised to refer to also other reviews available, including those from Glenn (2011), and Metzker (2010).

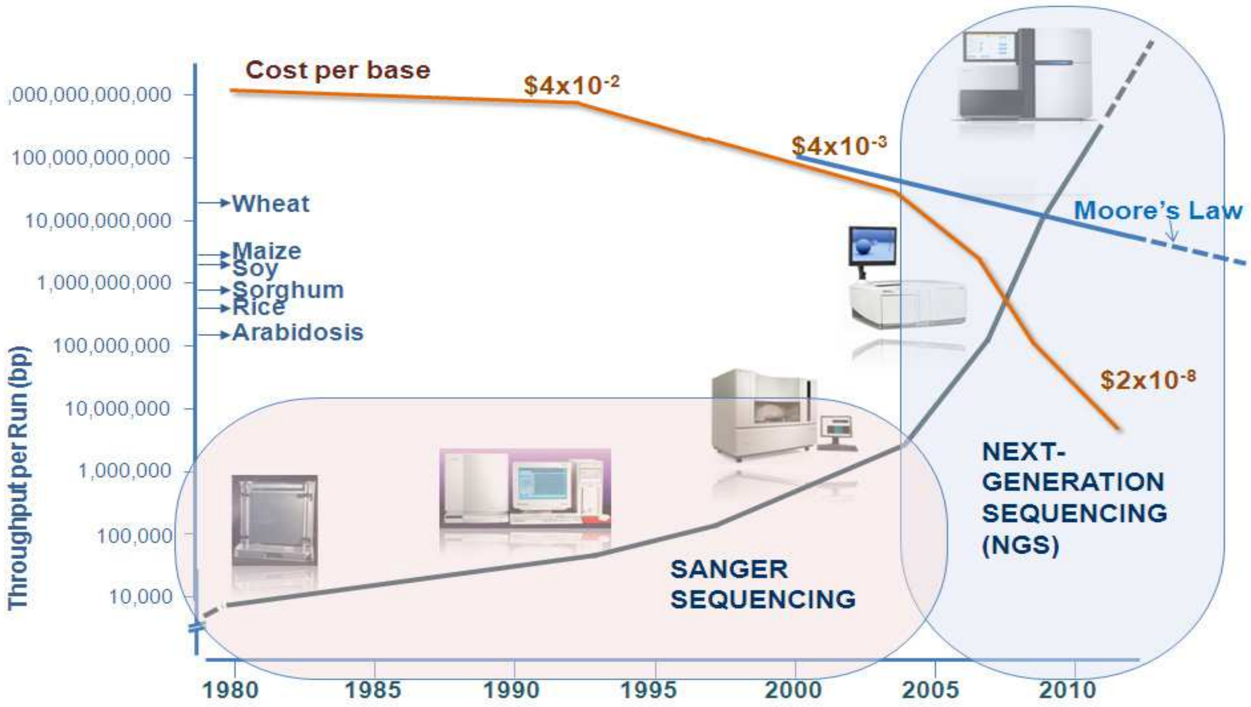


Fig. 1. Increase in maximum throughput per run in sequencing platforms from 1980 to 2011. Throughput per run based on Stratton *et al* (2009) and Glenn (2011).

2.1 Sanger sequence analyzers

For more than 30 years and until recently, sequencing based on the Sanger and Maxam-Gilbert chemistries were the only practical methods to routinely determine DNA sequences in plants and other biological systems. During the 80's and 90's, Sanger-based platforms increased throughput by orders of magnitude, and became the method of choice, while the Maxam-Gilbert method remained a low-throughput process. The development of automated Sanger systems was greatly facilitated by technical innovations such as thermal cycle-sequencing and single-tube reactions in combination with fluorescence-tagged terminator chemistry (Trainor, 1990). Additional improvements in parallelization, quality, read length, and cost-effectiveness were achieved by the development of automatic base-calling and capillary electrophoresis. In the current version of Sanger sequencing a mixture of primer, DNA polymerase, deoxinucleotides (dNTPs) and a proportion of dideoxynucleotide terminators (ddNTP), each labeled with a different fluorescent dye, are combined with the DNA template. During the thermal cycling reaction, DNA molecules are extended from templates and randomly terminated by the occasional incorporation of a labeled ddNTP. DNA is then cleaned up and denatured. Detection is achieved by laser excitation of the fluorescent labels after capillary-based electrophoresis separation of the extension products. The differences in dye excitation generate a "four color" system that is easily translated by a computer to generate the sequence. Modern Sanger sequencers like the Applied Biosystems ABI3730 have reached a high level of sophistication and can achieve routine read-lengths close to 900 bp and per-base 'raw' accuracies of 99.99% or higher (Shendure & Ji, 2008). The ABI 3730xl analyzer can run 96 or 384 samples every 2-10 hours, generating approximately 100,000 bases of raw sequence at a cost of a few hundred dollars.

2.2 Roche 454

The 454 platform (now owned by Roche) was the first NGS platform available as a standalone system. DNA templates need to be prepared by emulsion PCR and bound to beads, with 1–2 million beads deposited into wells in a titanium-covered plate. The Roche 454 technology is based on Pyrosequencing and additional beads that have sulphurylase and luciferase attached to them are also loaded into the same wells to generate the light production reaction. DNA polymerase reactions are performed in cycles but, unlike Sanger, there are no terminators. Instead, one single dNTP is alternated in every cycle in limiting amounts. Fluorescence after the reaction indicates the incorporation of the specific dNTP used in the cycle (Metzker, 2010). Because the intensity of the light peaks is proportional to the number of bases of the same type together in the template, the fluorescence can be used to determine the length of homopolymers, although accuracy decreases considerably with homopolymer length. The current 454 chemistry is able to produce the longest reads of any NGS system, about 700 bp, approaching those generated by Sanger reads. However, 454 systems can sequence several megabases for less than 100 dollars.

2.3 Illumina

The Solexa platform (now owned by Illumina) has become the most widely used NGS system in Plant biotechnology and breeding. Illumina captures template DNA that has been ligated to specific adapters in a flow cell, a glass enclosure similar in size to a microscope slide, with a dense lawn of primers. The template is then amplified into clusters of identical molecules, or polonies, and sequenced in cycles using DNA polymerase. Terminator dNTPs in the reaction are labeled with different fluorescent labels and detection is by optical fluorescence. As only terminators are used, only one base can be incorporated in one cluster in every cycle. After the reaction is imaged in four different fluorescence levels, the dye and terminator group is cleaved off and another round of dye-labeled terminators is added. The total number of cycles determine the length of the read and is currently up to 101 or 151, for a total of 101 or 151 bases incorporated, respectively. At the time of writing this review, this technology was able to yield the highest throughput of any system, with one of the highest raw accuracies. One major disadvantage is the short read it produces. However, paired-end protocols virtually double the read per template and facilitate some applications that were originally out of the reach of the technology. The Illumina HiSeq 2000 sequencer is currently able to sequence up to 540-600 Gbp in a single 2-flow cell, 8.5-day run at a cost of about 2 cents per Mbp (http://www.illumina.com/systems/hiseq_2000.ilmn).

2.4 Life Technologies SOLiD

ABI (now part of Life Technologies) has commercialized the SOLiD (Support Oligonucleotide Ligation Detection) platform. This platform is based on Sequencing by Ligation (SbL) chemistry. SbL is a cyclic method but differs fundamentally from other cyclic NGS chemistries in its use of DNA ligase instead of polymerase, and two-base-encoded probes instead of individual bases as units. In SbL, a fluorescently labeled 2-base probe hybridizes to its complementary sequence adjacent to the primed template and ligated. Non-ligated probes are then washed away, followed by fluorescent detection. In SOLiD, every cycle (probe hybridization, ligation, detection, and probe cleavage) is repeated ten times to yield ten color calls spaced in five-base intervals. The extension product is removed and additional ligation rounds are performed with an n–1 primer, which moves the calls by one position. Color calls from the ligation rounds are then ordered into a linear sequence to decode the DNA sequence (Metzker, 2010). SOLiD has similar throughput and cost per base to Illumina. It also has the best raw accuracy among commercial NGS systems.

2.5 Life Technologies Ion Torrent

Ion torrent is the commercial name for a new NGS platform now owned by Life Technologies (Rothberg *et al,* 2011; http://www.iontorrent.com). At the time of writing this chapter, the system was not widely used in plant research and its use elsewhere had been described in a limited number of published research papers (e.g. Miller *et al,* 2011). However, with recent upgrades, fast turnaround times and affordability, the system is finding its way into research laboratories. Currently, its usefulness in being evaluated for a number of applications in plant biotechnology and breeding. Ion Torrent differs from other NGS in that its chemistry does not require fluorescence or chemiluminescence, and for that matter optics (e.g. a CCD camera) to work. Beads, each carrying PCR clones from a single original fragment, are subjected to polymerase synthesis using standard dNTPs on an ion chip. The ion chip is a massively parallel semiconductor-sensing device that contains ion-sensitive, field-effect transistor-based sensors (tiny pH meters, essentially), coupled to more than one million wells where the polymerization reaction occurs. Cycles of reactions including one single nucleotide are produced, in a way that is analogous to the Roche 454 system. In each cycle, the electronic detection of changes in pH due to the release of a proton during base incorporation indicates that a base has been incorporated. The IonTorrent has the lowest throughput but also the fastest turnaround times of all commercially available NGS systems. The current Ion Torrent chip can yield several hundred thousand reads with an average length of about 100 bp in less than 2 hours.

Platform	**3730xl**	**5500xl SOLiD**	**454 FLX Titanium**	**HiSeq 2000**	**GAIIx**	**MiSeq**	**Ion Torrent**
Company	**ABI**	**ABI**	**Roche**	**Illumina**	**Illumina**	**Illumina**	**Life Tech.**
Chemistry	Dideoxy	SbL	PS	SbS	SbS	SbS	pH
Amplification	Biol/PCR	EmPCR	EmPCR	BrPCR	BrPCR	BrPCR	EmPCR
Detection	Fluor.	Fluor.	Fluor.	Fluor.	Fluor.	Fluor.	pH
Run Time (days)	0.08	8	0.5	8	14	1.1	0.08
Max. Aver. Length (bp)	900	60x2	700	101x2	151x2	151x2	100
Max. TP/run (Gbp)	0.00008	310	0.8	600	100	1	0.1
Max.Reads/Run(Million)	0.000096	5,167	1	3,000	320	3	1
TP per 24hr (Gbp)	0.00064	45	1	75	7	1	2.4
Raw Error range (%)	0.01	0.01	1-3	0.1	0.1	0.1	(1)*
Equip.Cost (xUS$1,000)	**150**	**600**	**300**	**690**	**350**	**125**	**60**
Cost per Mbp (US$)	**4,000**	**0.05**	**8**	**0.02**	**0.1**	**0.7**	**10**

SBS: Sequencing by synthesis, SbL: Sequencing by ligation, PS: Pyrosequencing, EmPCR: Emulsion PCR, Biol: Biological cloning, Fluoresc.: Fluorescence, BrPCR: Bridge PCR, TP: Throughput.

Table 1. Comparison of current sequence technologies.

2.6 Other NGS platforms, Helicos Heliscope, Polonator

There are other NGS systems that have been marketed in the last few years, however, they have had limited use in plant sciences. Helicos developed the first commercial single-molecule sequencer, called HeliScope. However, very few units were sold due to the cost of the machine, on-site requirements and other considerations. Currently, Helicos provides sequencing as a service. One additional company, Azco-Biotech is marketing the Max-Seq Genome sequencer (http://www.azcobiotech.com/instruments/maxseq.php). This commercial version of the academic, open-source Polonator can run either sequencing by synthesis or sequencing by ligation protocols, similar to Illumina and SOLiD, respectively, although it generates shorter reads, 35- or 55-bp-long.

2.7 Pacific Biosciences and the 3rd generation

Pacific Biosciences has launched the PacBio RS platform, considered the first commercially available 3rd-Generation system. The first early-access instruments were deployed in late 2010 and the first commercial batch became available by mid-2011. The PacBio system is based on SMRT, a single-molecule sequencing chemistry with real time detection. The sequencing cell has DNA polymerases attached to nanowells and exposed to single molecule templates and labeled NTPs. No terminators are used, although conditions are set to slow polymerization to a level that can be detected by a CCD camera. Each dNTP has a unique fluorescent label that is detected and then cleaved off during synthesis. Polymerization is detected as it happens, several bases per second. Because of this real-time detection and the enzyme processivity, this method has the potential to generate reads in excess of 10 kilobases in a few minutes. The potential of a technology that is able to sequence single molecules and produce long reads is immense. However, the PacBio technology may need to overcome a number of technical challenges before it reaches a widespread use in plant sciences. Average read length in current outputs exceeds 1 Kbp although single-pass error rate has been reported to be 15%, considerably higher than other sequencing platforms (Glenn, 2011). One major source of errors consists of deletions produced during detection. As will be discussed later, improvements in raw quality and further gains in read length will broaden the range of optimal applications for PacBio.

3. Applications in plant research

Sequencing platforms have different combinations of throughput, cost, read length, number of reads and raw accuracy. Their effective use in plant research and development programs depends on matching the best Sanger, NGS or Third Generation platform to specific applications (Morozova & Marra, 2008; Schuster, 2008; Varshney *et al,* 2009). One common misconception about Sanger-based systems is that they have, or will soon become obsolete. On the contrary, Sanger capillary systems are still the most widely used sequencers in routine molecular biology applications and are not likely to disappear in the near future. While their number of optimal applications has decreased, Sanger sequencers remain essential in many. The characteristics of capillary Sanger systems make them better suited for confirmatory sequencing in recombinant DNA technology, when the need to determine specific targets at low throughput makes them cost-effective. They are also best in low- to medium-throughput low-complexity shotgun and targeted sequencing experiments, where the use of highly-parallel random sequencing is impractical. Currently, no other chemistry or technology can match Sanger's combination of length and quality that remain the gold standard of sequencing.

Most sequencing applications can be divided into 2 categories: 1)*de novo* sequencing, and 2) resequencing. In the case of *de novo* sequencing, reads are obtained from an unknown sequence and either assembled to reconstruct this sequence or compared directly to reads from other unknown sequences. In the case of resequencing, reads are mapped or aligned to a known reference sequence. *De novo* applications are usually slower and more computer-intensive than resequencing, but are needed to reconstruct genomes and transcriptomes in species with unknown genomes. Major resequencing applications include polymorphism discovery and transcription profiling. This section emphasizes the use of new massive sequencing technologies and how they have recently been deployed in *de novo* and resequencing applications in plant research.

3.1 Physical maps and reference genomes

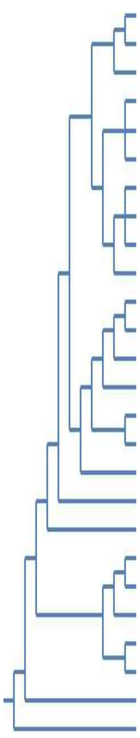
It is not surprising that considerable effort has been given during the last 15 years to the sequencing of plant genomes. The determination of nuclear and organellar genomes enables the identification of genes, regulatory elements, and the analysis of genome structure. This information improves our understanding of the role of genes in development and evolution, and facilitates the discovery of related genes and functions across species (Messing & Llaca, 1998; Feuillet *et al,* 2011). Reference genomes are also important tools in the identification, analysis and exploitation of genetic diversity of an organism in plant population genetics and breeding (Varshney *et al,* 2009; Edwards & Batley, 2010; Jackson *et al,* 2011). The sequencing of the human genome and other vertebrates in the 90's provided the technological pathway for the initial sequencing of genomes in plants (International Human Genome Sequencing Consortium, 2001, Venter *et al,* 2001). However, the structure of plant genomes poses additional challenges. Plant genomes are characterized by higher proportions of highly repetitive DNA and by the presence of segmental duplications or full genome duplications due to polyploidization events. The 1C genome content in Maize, for example, is smaller than in humans but consists of higher proportions and larger tracks of high-copy elements such as retrotransposable elements. Only a small fraction of the genome corresponds to exons and regulatory regions, usually in low-copy DNA islands that harbor single genes or small groups of genes (Schnable *et al,* 2009; Llaca *et al,* 2011). The average genome size in plants is larger than humans, approximately 5.8 Gbp, and they have a wider size distribution than mammals (Bennett & Leitch, http://data.kew.org/cvalues.). Some important crops like hexaploid wheat can have genomes that are more than 4 times the size of the human genome.

The first completed reference plant genomes, Arabidopsis and Rice, were from model plant species with small genomes, approximately 4% and 12% the size of the human genome. The genomes were produced by Sanger-based shotgun sequencing of overlapping bacterial artificial chromosomes (BACs) (The Arabidopsis Genome Initiative, 2001; International Rice Genome Project, 2005). This BAC-by-BAC approach requires the initial construction, fingerprinting and physical mapping of large numbers of random BACs (Soderlund *et al,* 1997; Ding *et al,* 2001). A subset of BACs is selected based on a minimum tiling path and shotgun libraries are individually constructed from each BAC and completed by subclone end-sequencing and assembly. Finally, BAC sequences are completed using a targeted approach aimed at closing sequencing gaps and finishing low-quality regions. This process, albeit slow and time consuming produced the only two references considered finished to date. These projects were performed by large collaborative consortia and took several years of fingerprinting and sequencing work. The cost of the Arabidopsis genome project has been estimated at US$70 million (Feuillet *et al,* 2011). In maize, a draft reference genome was completed from the inbred line B73 using a similar approach, although no gap closure or low quality finishing steps were completed. The maize draft genome, a highly valuable genetic resource available to the plant research community, was accomplished by multiple laboratories at an estimated cost of tens of millions in a joint NSF/DOE/USDA program. The three BAC-by-BAC sequencing projects mentioned above benefited from working in small units (BACs), which minimized problems associated by misassembly of highly repetitive DNA. One important consideration about BAC-by-BAC genomes is that they are not really complete. They have representation gaps in regions that are "unclonable" under the conditions used to prepare the BAC libraries. Many of these unclonable regions correspond to tandem repeats such as telomeric sequences and other repetitive regions, although it may also include gene space (Schnable *et al, 2009*). Furthermore, even in BAC-by-BAC approaches, the complexity of many plant genomes of moderate size such as maize prevent the creation of a complete physical assembly and there are some regions that may still lay in unassigned regions.

The high cost, long time, and logistics of BAC-by-BAC projects led many groups to adopt an alternative strategy also previously implemented in humans and other vertebrates: Whole-Genome Sequencing (WGS; Venter *et al,* 2001). In WGS, whole genomic DNA is randomly sheared and the fragments are end-sequenced and assembled. This strategy has improved with the use of multiple genomic libraries with different insert sizes and improved assembly software, which can identify such constraints in clone size. Not surprisingly, the first WGS, Sanger-based draft genomes were obtained from small genomes with relatively small amounts of repetitive DNA, including *Populus* (Tuskan *et al,* 2006), Grape (Jailon *et al, 2007*), and Papaya (Ming *et al, 2008*). More recent refinements enabled the sequencing of larger genomes such as *Sorghum bicolor* (~730 Mbp; Paterson *et al, 2009*) and soybean, an ancestral tetraploid (1.1 Mbp; Schmutz *et al, 2010*). The cost and time to accomplish these projects is reduced in comparison to BAC-by-BAC projects, although still considerable. In the case of soybean, the largest plant genome completed by Sanger WGS, sequencing was done by a team of 18 institutions and a total of more than 15 million Sanger reads were produced and assembled from multiple libraries with average sizes ranging from 3.3 Kb to 135 kb (Schmutz *et al, 2010*). In general, WGS approaches are effective in the determination of gene space in small and medium size plant genomes. However, reduction in time and cost is achieved at the expense of assembly fidelity in repetitive regions and expanded need for computational resources. WGS-based approaches increase potential assembly artifacts due to haplotype and homeolog collapse in regions with high identity. This may lead to large numbers of scaffolds to be mapped.

The use of NGS platforms in WGS projects has improved the ability to rapidly determine reference genomes at the expense of overall assembly quality, especially in high copy and duplicated regions. The potato reference genome (The Potato Genome Sequencing Consortium, 2011) was successfully constructed using a combination of Illumina, 454 and

Sanger reads. The implementation of hybrid methods using Roche 454 sequencing in combination with Sanger sequences has been effective in reducing overall cost and time to generate high-quality sequences in gene space regions (Rounsley, 2009). Examples of hybrid references are cucumber (Huang *et al,* 2009) and apple (Velasco *et al, 2*010). The use of NGS-only WGS assemblies, especially based on Illumina or Solid reads, can reduce cost and time by orders of magnitude in relation to Sanger or Hybrid strategies. Medium-size genomes such as maize, can be covered up to 200-fold in a single 9-day run in an Illumina HiSeq2000 system for under $30,000, for example. However, correct mapping and *de novo* assembly of these shotgun short reads has been problematic. Short reads have raised concern about their ability to accurately assemble genomes with high abundance of near identical repetitive sequences and gene duplication. The difficulty of using shotgun short read data for *de novo* assembly has also been a challenge in humans and other animals, but it is exacerbated in plants due to the higher proportion of highly repetitive DNA, segmental duplications and polyploidization. However, improvements have been made recently by using strategies that rely on paired-end reads and mate pairs, the use of multiple libraries with different insert sizes and the development of software with algorithms use end-sequence distance information from these libraries. Using these strategies, contig size, particularly in gene-rich regions, has increased considerably (Li *et al*, 2010). As read length in NGS continues to expand (e.g. Illumina platforms can perform 150-bp paired ends, and Roche 454 has released a long read chemistry), assembly will be improved.



Plant/species	SeqTotal/Est size (Mb)	Chrom.	Technology	Strategy	Reference
Cassava *(Manihot esculenta)*	533 / 760	8	454 + Sanger	WGS	www.phytozome.net
Castor Bean *(Ricinus communis)*	350 / 400	10	Sanger	WGS	Chan et al, 2010
Poplar *(Populus Trichocarpa)*	410 / 485	19	Sanger	WGS*	Tuskan et al, 2007
Medicago truncatula	214 / 307	8	Sanger	BbB	www.phytozome.net
Lotus japonica	315 / 472	6	Sanger	WGS, BbB	Sato et al, 2008
Soy *(Glycine max)*	950 / 1,100	20	Sanger	WGS	Schmutz et al, 2010
Apple *(Malus x domestica)*	603 / 742	8	454 + Sanger	WGS	Velasco et al, 2010
Woodland Strawberry *(F. vesca)*	209 / 240	7	454+Illumina+SOLiD	WGS	Shulaev et al, 2010
Peach tree *(Prunus persica)*	227 / 269	8	Sanger	WGS	www.rosaceae.org
Cucumber *(Cucumis sativus)*	203 / 880	14	Illumina + Sanger	WGS	Huang et al, 2009
Arabidopsis thaliana	115 / 125	5	Sanger	BbB	The Arab. Genome. Init., 2001
Arabidopsis Lyrata	207 / 207	8	Sanger	WGS	Hu et al, 2011
Papaya *(Carica papaya)*	135 / 367	9	Sanger	WGS	Ming et al, 2008
Chocolate *(Theobroma cacao)*	326 / 430	10	454 + Sanger	WGS	Argout et al, 2011
Sweet Orange *(Citrus sinensis)*	319 / 380	9	454 + Sanger	WGS	www.citrusgenomedb.org
Mandarin *(C. clementina)*	296 / 370	9	Sanger	WGS	www.citrusgenomedb.org
Eucalypt *(Eucalyptus grandis)*	641 / 650	22	Sanger	WGS	www.phytozome.net
Grape *(Vitis vinifera)*	715 / 416	19	Sanger	WGS	Jaillon et al, 2007
Potato *(Solanum tuberosum)*	727 / 844	12	454+Illumina+Sanger	WGS	Potato Gen. Seq. Cons., 2011
Sorghum *(Sorghum bicolor)*	730 / 734	10	Sanger	WGS	Paterson et al, 2009
Corn *(Zea mays)*	2,300 / 2,650	10	Sanger	BbB	Schnable et al, 2009
Foxtail Millet *(Setaria Italica)*	405 / 515		Sanger	WGS	www.phytozome.net
Rice *(Oryza sativa)*	389 / 400	12	Sanger	BbB	Intl Rice Gen. Seq Proj., 2005
Brachypodium distachyon	272 / 355	5	Sanger	WGS	Intl. Brachypodium Init., 2009
Selaginella moellendorffii	215 / 86	27	Sanger	WGS	www.phytozome.net
Physcomitrella patens	480 / 518	27	Sanger	WGS	Rensing et al, 2008

WGS – Whole genome sequencing, BbB – BAC by BAC.

Fig. 2. Reference Plant Genomes. The list includes genomes publicly available.

The use of NGS to sequence genomes in a BAC-by-BAC, or a pooled BAC approach can be facilitated by the use of new physical mapping technologies such as whole genome profiling (WGP). This process allows the physical mapping of BACs using a restriction-based fingerprinting approach analogous to high information content gel electrophoresis. In this system BAC clones are pooled, then DNA from the pools are prepared and digested with a restriction endonuclease. Tags are then added to the ends of the fragments and the labeled fragments are end-sequenced using Illumina chemistry. The sequence data are processed and analyzed by an optimized FPC software program to build BAC contigs across the genome. The sequence data obtained during WGP can be combined with BAC, BAC pool, and or Whole Genome Sequencing data. (Steuernagel 2009; van Oeveren *et al, 20*11).

Regardless the assembly strategy or sequence technology used, the completion of reference genomes for most plants remains a big challenge. All publicly available completed references indicated in Figure 2 correspond to plant species with below-average genome sizes. The challenge of sequencing full genomes with vast amounts of duplications and continuous high copy transposable elements remains inaccessible with the current technology. The major technological breakthrough required here is the improvement of 3rd Generation technologies able to produce long reads. Such long reads can then be used to improve contig length in combination with other technologies, or by themselves.

3.2 Development of pan-genomes

The significant sequence diversity and the high structural polymorphism observed in important plant models such as maize, highlights a serious limitation in the use of a single reference genome as a sufficient representative of a species. There is accumulating evidence that large differences in 1C DNA content observed between closely-related species, or between subspecies, landraces and lines in the same species correspond not only to differences in repetitive, non-coding DNA but also to gene content (Morgante *et al,* 2007; Llaca *et al,* 2011). Deep resequencing and the addition of *de novo* assembly of non-reference genomes is necessary to capture gene space included in larger structural variations (i.e. CNVs, PAVs, and large indels). With improvements in long-read sequencing technologies and assembly software and strategies, the creation of reference "pan-genomes" for certain species will be an important resource in plant genetics research

3.3 Genome surveys and partial genome assembly

The use of genome survey sequence (GSS) and partial targeted *de novo* assembly strategies can be useful in gene discovery research projects involving non-model plants species, or species with significant sequence diversity and high structural polymorphism. Maize and other crops exhibit pan-genomes that can be considerably larger than the standard available reference genome, with presence-absence variation (PAV) polymorphisms including both non-genic regions and gene space. Partial *de novo* assembly is also useful for gene discovery in large genomes or any non-model species where a region of interest can be genetically mapped and a partial physical map can be derived from the genetic map. Target regions may be included in one single BAC clone or in a series of overlapping BAC clones, determined by fingerprinting or by known probes that are used to develop assemblies. Rounsley *et al*. (2009) sequenced and assembled a 19Mbp region of chromosome 3 in rice, using the 454 reads and was able to generate scaffolds with size ranging from 243 Kbp to 518 Kbp. Similar approaches have been used in cacao (Feltus *et al,* 2011) and Barley (Steuernagel *et al,* 2009).

3.4 Plant-associate genomics

Genome sequence information from plant pathogen, comensal and mutualistic species is an important resource for plant improvement. Knowledge on gene content, expression and diversity of plant-associated organisms helps our understanding of the basis of their interactions with plants and in developing strategies to modify such interactions. Sequencing the genomes of mutualistic endophytes can help in the modification of nitrogen fixation and other processes, which are essential for developing a more sustainable agriculture. Genomic resources created from pathogens and their non-pathogenic relatives provide not only targets to develop increased disease resistance and pest control, but improved mechanisms for gene transfer in plant biotechnology as well (Wood *et al,* 2001).

Due to their simplicity, DNA and RNA viruses were among the first pathogens to be sequenced. Strains for more than 700 plant viruses, ranging from 1.2 to 30 Kb, have been completely sequenced to date (http://www.ncbi.nlm.nih.gov/genomes/). However, most recent advances in plant-associated genomes have been related to bacterial pathogens. They represent an important data mining resource for plant pathologists regarding some of the most devastating agricultural diseases (Stavrinides, 2009). From a technical point of view, they are also simpler to sequence than plant genomes and better suited for current technologies. Most bacterial genomes are approximately 5 Mbp, approximately 1,000-fold smaller than an average plant genome, with a relatively simple structure, often consisting of a single amplicon, although other amplicons such as megaplasmids and plasmids are frequently present. Bacterial genomes have little or no highly-repetitive elements. Due to their importance and relative technical simplicity, a considerable number of bacterial pathogens had already been sequenced by Sanger methodologies before the advent of NGS. Among the first published genomes are the citrus chlorosis agent *Xylella fastidiosa* (Simpson *et al,* 2000), *Pseudomonas syringae* pv tomato DC3000 (Buell *et al, 20*03), and *Ralstonia solanacearum* (Salanoubat *et al,* 2002). Currently, there are approximately 50 public finished and draft genomes from pathogenic bacteria, plus an even larger number currently in progress or unpublished. The smallest genome is that of *Phytoplasma asteris*, an obligated intracellular pathogen (0.9 Mbp; Oshima *et al,* 2004) and the largest (10.4 Mbp) is from *Streptomyces scabies* 87-22, the causal agent of Potato scab (http://www.sanger.ac.uk/). Complete genomes of other, non-pathogenic, mutualistic bacteria have also been sequenced. Krause *et al* (2006) reported the complete 4.3 Mbp genome of *Azoarcus* sp. strain BH72. Azoarcus is a mutualistic endophyte in cereal species, supplying biologically fixed nitrogen to its host while colonizing plants in high numbers without eliciting disease. Unlike related species that are pathogens, Azoarcus shows a lack of genes that pathogenic bacteria possess that degrade plant cell walls.

Fungi and stramenopiles are eukaryotic groups that include important plant pathogens and mutualists and have been the focus of genome sequencing projects. Stramanemopiles such as *Phytophtora spp*. are fungus-like eukaryotes although they are more closely related to diatoms. Genomes from at least 15 fungal and stramenopile pathogens are publicly available (see the Phytopathogen Genomics Resource, http://cpgr.plantbiology.msu.edu/ for a comprehensive list). Among them is *Phytophtora infestan*s, the stramanemopile causal agent of the Irish Potato Famine in the 19th century. The 240 Mbp sequence was assembled using Sanger sequencing (Haas *et al*, 2009). Martin *et al* (2008) sequenced the 65-Mbp genome of the fungus *Laccaria bicolor* that is part of a mycorrhizal symbiosis. The adoption of NGS for *de novo* sequencing of prokaryotic and fungal plant pathogens has been effective, especially when using a combination of 454 and Illumina reads (Reindhart *et al, 2009;* DiGuistini *et al, 2009*).

The biological relevance and economic importance of certain nematodes and insects has made them desirable targets for genome sequencing. Currently, publicly available genomes have been produced by Sanger-based WGS, including drafts for the Northern root-knot nematode *Meloidogyne hapla*, (Opperman *et al*, 2008). The genomes of the Aphid *Acyrthosiphon pisum* and the beetle *Tribolium castaneum*, which produce damage in crops and stored grains, have been completed (International Aphid Genomics Consortium, 2010; Tribolium Genome Sequencing Consortium, 2008).

Within two degrees of separation, genome sequencing of bacterial and fungal species that are pathogens of plant-infesting insects and nematodes are important resources in developing effective and safer strategies for pest resistance. One of such resources is the complete sequencing of replicons comprising the genome of *Bacillus thuringiensis*, which expresses insecticidal crystal proteins that have been used to engineer insect-resistant crops (Roh *et al*, 2007; Challacombe *et al,* 2007).

3.5 Metagenomics

The previous section underscores the importance of understanding the genomics of plant-associated microbiota. However, there are multiple interactions between plants and non-characterized microorganisms, some of them still to be discovered, that cannot be grown in cultures in the laboratory. (Riesenfeld *et al,* 2004; Allen *et al,* 2009). There is increasing evidence of the effect these organisms have in traits such as disease resistance and nitrogen utilization (Handelsman, 2004). Metagenomic studies of these microbial communities can exploit the availability of DNA amplification techniques and highly-parallel, clone-free NGS sequencers to sequence part of their genomes (Chen & Pachter 2005; Leveau, 2007). There are two major roles that high-throughput sequencing technologies can play in metagenomes applied to agriculture. The most common role is the mass sequencing of environmental (e.g. soil, water) samples to provide a systems-biology view of the microbiota under study. This type of study focuses on the genetic diversity and interactions between large numbers of plant associates and plants (Krober *et al,* 2009). Roche 454 Pyrosequencing of small subunit (16S) ribosomal RNA amplicons (pyrotags) is a method for profiling microbial communities that provides deep coverage with low cost, although it is complicated by several artifacts, including chimeric sequences caused by PCR amplication and sequencing errors. Illumina protocols have also been developed for the sequencing of “itags” derived from 16S hypervariable regions, for deep metagenomics analysis (Degnan & Ochman, 2011).

A second trend in modern metagenomics involves its exploitation for the discovery of biomolecules with novel properties. Current discovery strategies involve the screening of metagenomic libraries. Jin *et al* (2007) identified a novel EPSP gene with high resistance to glyphosate and potential use in plant biotechnology by screening a metagenomic library derived from a glyphosate-polluted area. However, the use of ultra-high throughput sequencing that could lead to simpler strategies is currently limited by the length of reads in NGS systems. Long reads are needed to generate full-sequence information within a single read. The use of sequencing approaches in biomolecule discovery will be feasible with the improvement of longer-read, 3rd-Generation technologies such as PacBio.

3.6 Genomic variant discovery for marker development

Linkage mapping, diversity and evolutionary studies in plants rely on the ability to identify and analyze single nucleotide and insertion-deletion polymorphisms (SNPs and Indels), which can be directly related to differences in a phenotype of interest, be genetically linked to its causative factor, or indicate relationships between individuals in populations (Rafalski, 2002). The implementation of high-throughput PCR-based marker technologies (e.g. Taqman) and improvements in Sanger sequencing throughput increased the limits for both the number of markers as well as samples in marker-related studies. These changes have enabled new applications in linkage and association mapping analysis, marker assisted selection (MAS) and characterization of germplasm. They have also facilitated fingerprinting and determination of seed purity. More recently, the emergence of NGS has enabled genome-wide discovery of polymorphisms on a massive scale. The Roche 454 system has been used effectively for polymorphism discovery (Gore *et al*, 2009a), although the higher throughput and lower cost of Illumina and SOLiD technologies make them particularly well suited for major programs when a reference genome is available (Deschamps & Campbell, 2010).

In species such as Arabidopsis and Rice, which have a small genome and an available reference genome, NGS-based genome-wide variant discovery can be simply accomplished by WGS (Ossowski *et al,* 2008; Huang *et al,* 2010). In medium- to large-sized genomes, where the proportion of gene space is reduced and much of the sequence is repetitive, the use of reduced-representation strategies can improve cost effectiveness. Reduced representation strategies involve the selection of specific regions of the genome to reduce complexity and increase coverage for the selected regions. Several enrichment strategies can be used to reduce genome representation. These approaches can utilize previous knowledge about the genome or region of interest. Examples of knowledge-driven enrichment include multiplex long-range PCR, molecular inversion probes (MIP), and sequence capture (Mamanova *et al,* 2010). These methods are usually preferred when a specific region or gene family is targeted. However, random approaches based on restriction digestion and transcriptome sequencing are more adequate in most genome-wide projects (Deschamps & Campbell, 2010). The use of methylation-sensitive enzymes or endonucleases that preferentially cut in low copy DNA have been particularly successful when used in strategies to identify large sets of SNPs in maize and soybean varieties (Gore *et al, 2009*b; Deschamps *et al,* 2009; Hyten *et al*, 2010). Illumina-based SNP discovery strategies using reduced representation libraries (RRLs) have been described by Deschamps and colleagues in soybean. By combining a 6-bp methylation-sensitive and one 4bp-restriction endonuclease they demonstrated enrichment for gene space, and considerable reduction of repetitive DNA reads (Deschamps *et al,* 2009; Fellers, 2008). Other methyl-filtration methods for reduced representation consist of digesting DNA with the endonuclease mcrBC (Gore *et al, 2009*; Palmer *et al,* 2003). Sequences derived from cDNA have been used in both Sanger and NGS sequences to reduce representation for polymorphism detection (Barbazuk *et al, 2007*; Trick *et al, 2009*). One major advantage to this method is the direct targeting of exonic DNA, which increases the chance of detecting functional SNPs, especially when used in conjunction with cDNA normalization methods. However, it can also constrain SNPs within a relatively small number of genes expressed in the tissue and stage used. Both standard whole-genome and RRL approaches usually yield a massive amount of polymorphisms, at a scale that is beyond prior Sanger-based projects. For example, Lam *et al* (2010) reported the genome resequencing in 31 wild and cultivated soy varieties, which led to the identification of more than 10 million SNPs in total, where more than 1 million of them were in genic regions. Nelson *et al* (2011) resequenced 8 sorghum (*Sorghum bicolor*) accessions using a reduced-representation approach in an Illumina system and identified 283,000 SNPs. With seemingly unlimited numbers of SNPs, current bottlenecks have been shifted from the discovery phase to marker assay development and validation.

In plant species where high-quality reference genomes are not available, variant discovery using an NGS resequencing approach can still be accomplished by using alternative references, such as high-quality transcriptome assemblies (see section 3.9) or *de novo* partial assemblies of individual BACs or BAC contigs (see section 3.5). However, both strategies carry some additional risk and validation must include potential for detection of repetitive sequences of paralogous genes. An alternative is an annotation-based strategy, as described in the wheat relative *Aegilops tauschii* by You *et al.* (2011). The genome size is more than 4Gbp and has a large proportion (more than 90%) of repetitive DNA. They produced Roche 454 shotgun reads at low genome coverage from one genotype and identified single-copy sequences and repeat junctions from repetitive sequences as well as sequences shared by paralogous genes. SOLiD and Solexa reads were then generated from another genotype and were mapped to the annotated Roche 454 reads. In this case, 454 reads provide a DNA "context" surrounding the putative SNPs, which can be used to generate genome-wide markers. They were able to identify nearly 500,000 SNPs with a validation rate higher than 81%.

3.7 QTL and eQTL mapping, hapmaps and WGAS

In plants, most agronomically important traits are quantitative. Plant yield, flowering time, sugar content, disease resistance and fruit weight are examples of quantitative traits, which result from the segregation of many genes and are influenced by environmental interactions (Paran & Zamir, 2003). While quantitative traits have been studied for more than 100 years, the mapping of the underlying quantitative trait loci (QTL) could only be accomplished after the development of sequencing methodologies, molecular markers and improved statistical methods. Furthermore, until 2005, only a small fraction of mapped plant QTL had been cloned (Salvi and Tuberosa, 2005; Frary, 2000). One major difficulty was the low resolution of available mapping strategies. Before the advent of NGS platforms, most QTL identification and cloning was based on linkage mapping strategies. In linkage mapping, polymorphisms are identified between two parents and then followed in a large segregating population. The linkage of different regions of the genome to the individual phenotypes can be then inferred statistically by identifying recombinants that show phenotypic differences in the trait of interest. One drawback to linkage mapping is the low resolution that results from relatively few recombinants generated from two original parents in a limited number of generations. Even in cases of QTL with large effects in the total genetic variance, intervals can encompass a large genetic and physical distance and require walking through several megabase-pairs of sequence, with a large number of potential candidates (Yano *et al*. 1997; El-Assal *et al*. 2001; Liu *et al*. 2002). In maize, recent linkage mapping studies have identified QTL with relatively large effects in oil content (Zheng *et al,* 2008) and root architecture (Ruta *et al,* 2010).

The development of high-throughput genotyping technologies and later the emergence of NGS platforms has enabled the use of genome-wide association studies (GWAS) and bulked segregant analysis to map plant QTL (Rafalski, 2010; Schneeberger & Weigel, 2011). Unlike linkage mapping, GWAS exploit the natural diversity generated by multi-generational recombination events in a population or panel (Risch & Merikangas, 1996; Yu & Buckler, 2006; Belo *et al,* 2008; Nordborg & Weigel, 2008). These results in increased resolution compared to linkage mapping populations, as long as enough markers are provided: GWAS may require hundreds of thousands or even millions of genetic markers to achieve sufficient coverage. Before NGS, such marker density was unfeasible and linkage disequilibrium, or association, mapping studies needed to focus on polymorphisms in candidate genes that were suspected to have roles in controlling phenotypic variation for one specific trait of interest (Thornsberry *et al, 20*01). In plants, availability of NGS and the ability to create lines of individuals with identical or near identical background offer the potential to create public GWAS resources that can be accessed by multiple groups and rapidly resolve complex traits. Plant GWAS can be performed in large numbers of samples in replicated trials using inbreds and recombinant inbred lines (RILs) (Zhu *et al,* 2008). One or more research groups can then analyze one or many traits in multiple environments. The most important GWAS resource in maize is a collection of recombinant inbred lines derived from a nested association mapping (NAM) population (Gore *et al,* 2009). The maize NAM population is a collection of 5,000 RILs in sets of 200, derived from one of 25 populations. (Each of the 200 RILs is derived from one F2 plant from a cross between one of 25 inbred lines to B73.) The original inbred lines that were used as founders of the NAM have been resequenced using a NGS reduced-representation approach. Such resequencing surveys (HapMaps) include a high quality data set consisting of 1.4 million SNPs and 200,000 indels spanning the 5,000 inbred lines. Seeds from the RILs can be used to grow and phenotype plants for any trait of interest (McMullen *et al,* 2009). Recent studies, all derived from the same NAM resource, demonstrate the effectiveness of this approach to identify and characterize QTL. Buckler *et al* (2009) identified 50 loci that contribute to variation in the genetic architecture of flowering time, with many loci showing small additive effects. Tian *et al* (2011) also identified large numbers of QTL with small effects determining leaf architecture. Poland *et al* (2011) identified candidate genes for resistance to northern leaf blight in 29 loci, which included QTL with small additive effects. Kump *et al* (2011), identified QTL for Southern Leaf blight. HapMaps in Rice have been reported by Huang *et al* (2010) by resequencing at low coverage a total of 517 landraces that yielded a total of 3.6 million SNPs. The study identified QTL with minor and major contributions to phenotypic variance for drought tolerance, spikelet number and 12 additional agronomic traits. In *Medicago truncatula*, Branca *et al* (2011) detected more than 3 million SNPs in 26 inbred lines to study the genetics of traits related to symbiosis and nodulation. In *Arabidopsis thaliana*, the 1,001 Genomes Project, started in 2008 aims at discovering polymorphisms in that number of wild accessions (Weigel and Mott, 2009; http://1001genomes.org/). The complete genome sequences of over 80 accessions have already been released and inbred lines have been generated from each accession.

Finally, the determination of genome variants and transcription profiling by NGS approaches can be used effectively in the determination of expression quantitative loci (eQTLs; Damerval *et al,* 1994). Variation in the expression of transcripts, when measured across a segregating population, can be used to map regions with *cis*- and *trans*- effects (Holloway & Li, 2010). The development of massively parallel sequencing technologies has replaced microarrays as the method of choice for eQTL analsysis (Holloway *et al,* 2011; West *et al*, 2007). Using NGS, Swanson-Wagner *et al* (2009) identified ~4,000 eQTLs in reciprocal crosses between the maize inbred lines B73 and Mo17, most of them acting in *trans* and regulated exclusively by the paternally transmitted allele.

3.8 Genotyping by sequencing

The value of NGS-driven massive polymorphisms discovery can be seriously restricted by cost and time limitations in the design, validation and deployment of molecular markers. With the falling cost of NGS there is an increased interest in genotyping-by-sequencing (GbS), where the obtained sequence differences are used directly as markers for analysis. As described in section 3.7, Maize NAM and other GWAS comunity-based resources already make use of GbS. However, such panels have limited utility beyond their populations or panels. A number of reduced-representation GbS protocols have been reported that can be applied to other population or panels for linkage, association, bulked segregant analysis, fingerprinting, diversity and other studies. Depending on the details of the project and the available resources, sequences can be mapped to a reference. However, in large genomes or other genomes with no reference available, the consensus of reads flanking the polymorphism can be used as a partial reference or polymorphic reads can simply be treated as dominant markers (Elshire *et al,* 2011). Construction of a low-density GbS linkage map using Restriction-Site-Associated DNA (RAD) has been reported in barley (Chutimanitsakun *et al,* 2010). The use of simpler and highly multiplexed protocols, however, is required in most cases to make GbS cost and time-effective. The bottom line is that an all-inclusive cost per sample is lower than those provided by other available genotyping platforms. Cost estimates need to include the considerable computational and bioinformatic resources needed for GbS data analysis. Using a simple reduced-representation procedure based on *Ape*KI restriction digestion, Elshire *et al* (2011) identified and mapped approximately 200,000 polymorphisms in the 2 parents and the 276 RILs from the maize IBM (B73 x Mo17) mapping population at an estimated cost of $29.00 per sample. With the same protocols and using a single Illumina run, they can process up to 672 samples, taking the actual data collection cost to well under US$20.00 per sample. The low cost per base and the high numbers of reads produced per run make the Illumina and SOLiD systems more suitable for GbS.

3.9 Transcriptome assembly and profiling

The sequencing of DNA products synthesized from total and mRNA isolates (cDNA) has been crucial in gene expression analysis, discovery and determination of alternative splicing forms of genes (isoforms). In the case of organisms with a genome sequence available, cDNA sequencing has facilitated the annotation of splicing sites and untranslated regions (UTRs), as well as improved gene prediction algorithms (Brautigam & Gowik, 2010). As indicated before, transcriptome sequencing can also be deployed as a reduced-representation strategy to identify polymorphisms for marker development and genotyping. Before the advent of NGS, multiple Sanger sequencing strategies were developed for the quantitative and qualitative analysis of mRNA expression. The need for direct quantitative analysis on gene expression led to the development of profiling strategies such as Serial-Analysis-of-Gene-Expression (SAGE; Velculescu *et al,* 1995). On the other hand, the creation of large consortia dedicated to providing end-sequence for individual clones from cDNA libraries enabled gene discovery, annotation and expression on a large scale (Rafalski *et al,* 1998). These efforts have yielded more than 22 million Sanger-based expressed sequenced tags (ESTs) from more than 40 plant species. The largest datasets correspond to arabidopsis, soybean, rice, maize and wheat, each of which contains more than 1 million entries (http://www.ncbi.nlm.nih.gov/dbEST/). There are also EST databases available for at least 42 fungal, stramenopile and nematode phytopathogen species (http://cpgr.plantbiology.msu.edu). Sequence information derived from these databases has been utilized to develop expression microarrays, which are aimed at establishing relative abundance of known genes in large numbers of samples and tissues within the same species or among related species (Rensink and Buell, 2005). While such arrays have been effective in providing gene expression data, they are inheritably biased in their design and have limitations in resolution and in their ability to differentiate between individual genes within families. The highly parallel, short-read NGS technologies such as Illumina and SOLiD have allowed the development of transcription profiling strategies that are more sensitive and accurate than SAGE or microarrays. Initial NGS strategies for transcription profiling had their roots in the innovative, now obsolete massively parallel signature sequencing (MPSS) technology. This technology, owned and provided as a service by Lynx Therapeutics, consisted in the generation and sequencing of short 17-bp unique tags, or signatures, from 3'-UTRs of transcripts at high coverage (Simon *et al,* 2009). It provided unparalleled resolution generating over a million signature sequences per experiment, although the cost of every experiment was considerable (Reinartz *et al,* 2002). With Illumina, tag-based “digital” expression profiling protocols became relatively simple and achieved higher resolution than MPSS at a fraction of the cost (Wang *et al*, 2010) .

The use of shotgun sequencing of cDNA using the Roche 454 analyzer has provided relatively long reads and high coverage for gene discovery, annotation and polymorphism discovery in both model and non-model plant species (Barbazuck *et al,* 2007; Emrich *et al,* 2007). More recently, the increasing gains in throughput, as well as improvement in shotgun RNA sequencing (RNA-seq) strategies and analysis software have expanded the potential of Illumina and SOLiD platforms for full transcriptome analysis, and replaced the use of the tag-based expression profiling approach. In RNA-seq, total or messenger RNA is fragmented and converted into cDNA. Alternatively, it is first converted into cDNA and then fragmented. Adaptors are attached to one or both ends, and sequenced as single- or paired-ends (Wang *et al,* 2009; Margerat & Bahler, 2010). Depending on the genomic resources available for the organism of interest, the resulting sequences can be aligned to either a reference genome or reference transcripts. Alternatively, genes can be assembled *de novo*. In either case, cDNA sequencing provides considerably more information on the transcriptome, including gene structure, expression levels, presence of multiple isoforms and sequence polymorphism. Unlike microarray-based hybridization, it does not depend upon previous knowledge of potential genes. A considerable number of RNA-Seq projects have been made in major crop species. Severin *et al* (2010), Zenoni *et al* (2010) and Zhang *et* *al* (2010), all applied Illumina-based RNA-Seq on multiple tissues and stages in soy, grape and rice, respectively, and aligned transcript reads to their respective reference genome sequences. Li *et al* (2011) used Illumina in multiple stages along a leaf developmental gradient and in mature bundle sheath and mesophyll cells.

3.10 Small RNA characterization

Small RNAs (sRNA) are non-protein-coding small RNA molecules ranging from 20 to 30 nt that have a role in development, genome maintenance and plant responses to environmental stresses (Simon *et al, 2*009). Most sRNAs belong to two major groups: 1) microRNAs (miRNA) are about 21 nt and usually have a post-transcriptional regulatory role by directing cleavage of a specific transcript, 2) short interfering RNAs (siRNA) are usually 24 nt-long and influence *de novo* methylation or other modifications to silence genes (Vaucheret 2006). The finding of their prevalence in low-molecular-weight fractions of total RNA in animals and plants predated the development of NGS. However, the use of MPSS greatly expanded resolution and later became clear that short-read NGS technologies such as Illumina or SOLiD had optimal characteristics in sRNA analysis (Zhang *et al*, 2009). Roche 454 sequencing has also been used in sRNA analysis (Gonzales-Ibeas *et al,* 2011).

3.11 Epigenomics

In plants and other multicellular organisms, cell differentiation is driven by variation of epigenetic marks encoded on the DNA or chromatin. Such variations can be stable, or heritable, but do not change the underlying DNA sequence of the genome (Zhang & Jeltsch, 2010). Furthermore, there is accumulating evidence that transgenerationally inherited epigenetic variants (epialleles) have a significant effect in differential gene expression in plant and animal populations (Reinders *et al,* 2009; Johannes *et al,* 2009). Biochemical alterations such as cytosine methylation polymorphisms and differential histone deacetylation are epigenetic marks that can play a critical role in development (Schöb & Grossniklaus, 2006; Chen & Tian, 2007). In plants, 5-methylcytosine can be present at symmetric CG sites but also can be located at CHG sites as well as in asymmetric CHH locations (where H can be A, C or T). Methylation at CG sites usually occurs symmetrically on both strands and is heritable, maintained by specific types of methyltransferases that recognize hemimethylated sites created during replication. Methylation of CHG and CHH is established and maintained by additional methyltransferases (Schob & Grossniklaus, 2006). One important consideration is that both the epigenome and methylome will be considerably larger than the genome of an organism. As a major part of the epigenome, the methylome consists of the sum of genome and methylation states at every cytosine location. Multiple states coexist in the same individual, depending on cell types, tissues, developmental stages or environments. Adding additional complexity is the fact that methylation at one position may be partial within the same cell type. Similar to the transcriptome, methylome analysis has, therefore, a quantitative component in addition to a qualitative one.

Different Sanger and NGS strategies have been developed over the years that can directly or indirectly identify epigenetic marks and patterns. Before NGS, epigenetic studies were mostly limited to individual genes or sets of candidate genes or regions. One exception is the extensive work done in arabidopsis by Zhang *et al* (2007), which provided the first genome-wide study in plants and considerable information on methylation distribution and effect in gene expression. The use of NGS technologies coupled with bisulfite conversion, restriction digestion, or immunoprecipitation strategies are facilitating genome-wide methylome analysis in plants. Of these, approaches based on sodium bisulfite conversion (BC) provide the highest resolution. In BC, denatured DNA is treated with sodium bisulfite, which induces the hydrolytic deamination of cytosine. Subsequent treatment with a desulfonation agent transforms the uracyl-sulfonate intermediate into uracyl. Replication and further amplification of the converted strands will incorporate a thymidine in originally non-methylated C sites. However, 5-methylcytosine residues remain unreactive during conversion and further amplification will retain the original CG pairing at that position (Liu *et al,* 2004; Frommer *et al* 1992). As a consequence, unmethylated and methylated cytosines can be mapped in the two original strands and resolution can reach single base level, as long as the original sequence is known (Zhang *et al*, 2007; Lister *et al*, 2008). The combination of methylome BC-NGS and high-definition transcriptome analysis will have an important role in further characterization of epi-regulation in plants.

4. Future outlook

For more than 30 years, the use of sequencing methods and technologies, in combination with strategies in breeding and molecular genetic modification, has contributed both to our knowledge of plant genetics and to remarkable increases in agricultural productivity. In recent years, agricultural sciences have been in the middle of a second technological revolution in DNA sequencing, driven in large part by the post-human genome goal of affordable genome sequencing for disease research and personalized medicine. The resulting NGS systems have become a “disruptive technology”, radically reducing limitations in sequence information and consequently altering the types of questions and problems that can be addressed (Mardis, 2010). As we have explored in this chapter, these massively parallel sequencing systems have had a dramatic effect in variant and gene discovery, genotyping, and in the characterization of transcriptomes, genomes, epigenomes and metagenomes. The increasing ability to sequence complete genomes from multiple individuals within the same species is providing a more comprehensive view of crop diversity and transgenesis, as well as a better understanding of the effect of mutational and epimutational processes in plant breeding. With all their high throughput and low cost, different NGS platforms have specific flaws and researchers have been playing on their weaknesses and strengths. Short reads, even at very high throughputs, may not be the best output for *de novo* sequencing. (After all, longer Sanger reads are not ideal either to determine the most complex eukaryotic genomes.) Alternatively, the use of systems that are able to produce longer reads but have relatively lower capacity will probably not be effective in some of the most important future applications in molecular breeding such as genotyping-by-sequencing.

The boundaries of sequencing technologies continue to expand, and the quest for more universal sequencers continues. The expected improvements in quality and read length in real-time 3rd-Generation systems have the potential to greatly benefit plant *de novo* sequencing and metagenomics applications. The development of cheaper, portable, easier-to-use machines has the potential to create a decentralization of sequencing and to create entirely new field applications. With shorter technological cycles in sequencing and computing, it is difficult to anticipate the next disruptive technology. While it is likely that 3rd Generation systems will soon become widespread in plant research, there has also been progress toward nanopore-based technologies. Nanopore systems are based on electronic detection of DNA sequence and have the potential of low sample preparation work, high speed, and low cost (Branton *et al*, 2008). Future improvements in sequencing technologies will enable applications that can support discovery and innovation needed to respond to growing population pressure, energy crises, decreasing fresh water availability and climate change (Gepts & Hancock, 2006; Moose & Mumm, 2008). Ultimately, DNA sequencing is one of a number of tools in plant breeding and biotechnology, albeit an essential one. The knowledge of genomic organization, diversity and function of genes in crops needs to be associated to an understanding of plant biology. Effective plant breeding programs need solid statistical strategies and the ability to create, manage and integrate large, heterogeneous sets of data based on phenotype and sequence-derived information.

5. Acknowledgments

The author thanks Antoni Rafalski, Stephane Deschamps, Delia Ines Lugo and Karlene Butler for their critical reading of this manuscript and for many interesting discussions.

6. References

Allen, C., Bent, A. & Charkowski, A. (2009). Underexplored niches in research on plant pathogenic bacteria. *Plant Physiology* 150(4): 1631–1637.

Arabidopsis Genome Initiative (2001). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408(6814): 796-815.

Argout, X. *et al* (2011). The genome of *Theobroma cacao*. *Nature Genetics* 43(2): 101-108.

Barbazuk, W.B. *et al* (2007). SNP discovery via 454 transcriptome sequencing. *Plant Journal,* Vol.51(5): 910–918.

Barnhart, B.J. (1989). DOE Human Genome Program. *Human Genome Quarterly* 1(1): 1.

Bedbrook, J.R., Smith, S.M. & Ellis, R.J. (1980). Molecular cloning and sequencing of cDNA encoding the precursor to the small subunit of chloroplast ribulose 1,5-bisphosphate carboxylase. *Nature* 287(5784): 692–697.

Belo, A. *et al* (2008). Whole genome scan detects an allelic variant of fad2 associated with increased oleic acid levels in maize. *Molecular Genetics & Genomics* 279(1): 1-10.

Bernatsky, R. & Tanksley, S. (1986). Toward a saturated linkage map in tomato based on isozymes and random cDNA sequences. *Genetics* 112(4): 887-898.

Branca, A. *et al* (2011). Whole-genome nucleotide diversity, recombination, and linkage disequilibrium in the model legume *Medicago truncatula*. *Proc. Natl. Acad. Sci. USA* 108(42): E864–E870.

Branton, D. *et al* (2008). The potential and challenges of nanopore sequencing. *Nature Biotechnology* 26(10): 1146-1153.

Brautigam, A. & Gowik, U. (2010). Next generation sequencing as a valuable tool in plant research. *Plant Biology* 12(6): 831-841.

Buckler, E.S. *et al* (2009). The architecture of maize flowering time. *Science* 325(5941): 714-718.

Buell, C.R. *et al* (2003). The complete genome sequence of the Arabidopsis and tomato pathogen *Pseudomonas syringae* pv. tomato DC3000. *Proc. Natl. Acad. Sci. USA* 100(18): 10181–10186.

Challacombe, J.F. *et al* (2007). The complete genome sequence of *Bacillus thuringiensis* Al Hakam. *Journal of Bacteriology* 189(9): 3680-3861.

Chan, A.P. *et al* (2010). Draft genome sequence of the oilseed species *Ricinus communis*. *Nature Biotechnology* 28(9): 951-956.

Chen, K. & Pachter, L. (2005). Bioinformatics for whole-genome shotgun sequencing of microbial communities. *PLoS Computational Biology* 1(2): 106-112.

Chen, Z.J., Tian, L. (2007). Roles of dynamic and reversible histone acetylation in plant development and polyploidy. *Biochimica et Biophysica Acta* 1769(5-6): 295-307.

Chutimanitsakun, Y. *et al* (2011). Construction and application for QTL analysis of a Restriction Site Associated DNA (RAD) linkage map in barley. *BMC Genomics* 12(4): 1-13.

Damerval, C. *et al* (1994). Quantitative trait loci underlying gene product variation: A novel perspective for analyzing genome expression. *Genetics* 137(1): 289-301.

Degnan, P.H. & Ochman, H. (2011). Illumina-based analysis of microbial community diversity. *The ISME Journal.* doi:10.1038/ismej.2011.74. [Epub ahead of print].

Deschamps, S. *et al* (2010). Rapid genome-wide single nucleotide polymorphism discovery in soybean and rice via deep resequencing of reduced representation libraries with the Illumina Genome Analyzer. *The Plant Genome* 3(1): 53-68.

Deschamps, S. & Campbell, M. (2010). Utilization of next-generation sequencing platforms in plant genomics and genetic variant discovery. *Molecular Breeding* 25(4): 553–570.

Diguistini, S. *et al* (2009). *De novo* genome sequence assembly of a filamentous fungus using Sanger, 454 and Illumina sequence data. *Genome Biology* 10(R94): doi:10.1186/gb-2009-10-9-r94.

Ding, Y. *et al* (2001). Five-color-based high-information-content fingerprinting of bacterial artificial chromosome clones using type IIS restriction endonucleases. *Genomics* 74(2): 142-154.

Edwards, D. & Batley, J. (2010). Plant genome sequencing: applications for crop improvement. *Plant Biotechnology Journal* 8(1): 2-9.

El-Assal, S.E.D. *et al* (2001). The cloning of a flowering time QTL reveals a novel allele of CRY2. *Nature Genetics* 29(4): 435–440.

Elshire, R.J. *et al* (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One* 6(5): e19379.

Emrich, S.J. *et al* (2007). Gene discovery and annotation using LCM-454 transcriptome sequencing. *Genome Research* 17(1): 69-73.

Fellers, J.P. (2008). Genome filtering using methylation-sensitive restriction enzymes with six-base pair recognition sites. *The Plant Genome* 1(2): 146-152.

Feltus, F.A. *et al* (2011). Sequencing of a QTL-rich region of the *Theobroma cacao* genome using pooled BACs and the identification of trait specific candidate genes. *BMC Genomics* 12(379): 1-16.

Feuillet, C. *et al* (2011). Crop genome sequencing: lessons and rationales. *Trends in Plant Science* 16(2): 77-88.

Fraley, R.T. *et al* (1983). Expression of bacterial genes in plant cells. *Proc. Natl. Acad. Sci. USA* (80)15: 4803–4807.

Frary, A. *et al* (2000). fw2.2: A quantitative trait locus key to the evolution of tomato fruit size. *Science* 289(5476): 85-88.

Frommer, M. *et al* (1992). A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc. Natl. Acad. Sci. USA* 89(5): 1827–1831.

Gepts, P. & Hancock, J. (2006). The future of plant breeding. *Crop Science* 46(4): 1630–1634.

Glenn, T.C. (2011). Field guide to next-generation DNA sequencers. *Molecular Ecology Resources* 11(5): 759-769.

Gonzalez-Ibeas, D. *et al* (2011). Analysis of the melon (*Cucumis melo*) small RNAome by high-throughput pyrosequencing. *Plant Physiology* 150(4): 1631–1637.

Gore, M.A. *et al* (2009a). A First generation haplotype map of maize. *Science* 326(5956): 1115-1117.

Gore, M.A. *et al* (2009b). Large-scale discovery of gene-enriched SNPs. *The Plant Genome* 2(2): 121-133.

Haas, B.J. *et al* (2009). Genome sequence and analysis of the Irish potato famine pathogen *Phytophthora infestans*. *Nature* 461(7262): 393-398.

Handelsman, J. (2004). Metagenomics: Application of genomics to uncultured microorganisms. *Microbiology and Molecular Biology Reviews* 68(4): 669-685.

Herrera-Estrella, L. *et al* (1983). Expression of chimaeric genes transferred in to plant cells using a Ti plasmid-derived vector. *Nature* 303: 209-213.

Holloway, B. & Li, B. (2010). Expression QTLs: applications for crop improvement. *Molecular Breeding* 26(3): 381-391.

Holloway, B. *et al* (2011). Genome-wide expression quantitative trait loci (eQTL) analysis in maize. *BMC Genomics* 12(336): 1-14.

Hu, T.T. *et al* (2011). The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. *Nature Genetics* 43(5): 476-481.

Huang, S. *et al* (2009). The genome of the cucumber, *Cucumis sativus* L. *Nature Genetics* 41(12): 1275-1281.

Huang, X. *et al* (2010). Genome-wide association studies of 14 agronomic traits in rice landraces. *Nature Genetics,* 42(11): 961-967.

Hyten, D.L. *et al* (2010). High-throughput SNP discovery through deep resequencing of a reduced representation library to anchor and orient scaffolds in the soybean whole genome sequence. *BMC Genomics* 11(38): 1-8.

International Aphid Genomics Consortium (2010). Genome sequence of the pea aphid *Acyrthosiphon pisum. PLoS Biology* 8(2): e1000313.

International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature* 409(6822): 860–921.

International Rice Genome Sequencing Project (2005). The map-based sequence of the rice genome. *Nature* 436(7052): 793-800

Jackson, S.A. *et al (2011).* Sequencing crop genomes: approaches and applications. *The New phytologist* 191(4): 915-925.

Jaillon, O. *et al* (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* 449(7161): 463-467.

Jin, D. *et al* (2007). Identification of a new gene encoding EPSPS with high glyphosate resistance from a metagenomic library. *Current Microbiology* 55(4): 350–355.

Johannes, F. *et al* (2009). Assessing the impact of transgenerational epigenetic variation on complex traits. *PLoS Genetics* 5(6): e1000530.

Krause, A. *et al* (2006). Complete genome of the mutualistic, N2-fixing grass endophyte *Azoarcus sp*. strain BH72. *Nature Biotechnology* 24(11): 1384-1390.

Krober, M. *et al* (2009). Phylogenetic characterization of a biogas plant microbial community integrating clone library 16S-rDNA sequences and metagenome sequence data obtained by 454-pyrosequencing. *Journal of Biotechnology* 142(1): 38-49.

Kump, K.L. *et al* (2011). Genome-wide association study of quantitative resistance to southern leaf blight in the maize nested association mapping population. *Nature Genetics* 43(2): 163-168.

Lam, H.M. *et al* (2010). Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. *Nature Genetics* 42(12): 1053-1059.

Leveau, J.H. (2007). The Magic and menace of metagenomics. *European Journal of Plant Pathology* 119(3): 279–300.

Li, P. *et al* (2011). The developmental dynamics of the maize leaf transcriptome. *Nature Genetics* 42(12): 1060-1067.

Li, R. *et al* (2010). *De novo* assembly of human genomes with massively parallel short read sequencing. *Genome Research* 20(2): 265-272.

Lister, R. *et al* (2008). Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell* 133(3): 523-536.

Liu, J. *et al* (2002). A new class of regulatory genes underlying the cause of pear-shaped tomato fruit. *Proc. Natl. Acad. Sci. USA* 99(20): 13302-13306.

Liu, L. *et al* (2004). Profiling DNA methylation by bisulfite genomic sequencing. Problems and solutions. *Methods in Molecular Biology* 287(1): 169-179.

Llaca, V., Campbell, M. & Deschamps, S. (2011). Genome diversity in maize. *Journal of Botany*, 2011(104172):1-10.

Mamanova, L. *et al* (*2010).* Target-enrichment strategies for next-generation sequencing. *Nature Methods* 7(2): 111-118.

Mardis, E.R. (2010). A decade's perspective on DNA sequencing technology. *Nature* 470(7333): 198-203.

Marguerat, S. & Bahler, J. (2010). RNA-seq: from technology to biology. *Cellular & Molecular Life Sciences* 67(4): 569–579.

Martin, F. *et al* (2008). The genome of *Laccaria bicolor* provides insights into mycorrhizal symbiosis. *Nature* 452(7182): 88-92.

Maxam, A.M. & Gilbert, W. (1977). A new method for sequencing DNA. *Proc. Natl. Acad. Sci. USA* 74(2): 560–564.

McMullen, M.D. *et al* (2009). Genetic properties of the maize nested association mapping population. *Science* 325(5941): 737-740.

Messing, J. & Llaca, V. (1998). Importance of anchor genomes for any plant genome project. *Proc. Natl. Acad. Sci. USA* 95(5): 2017–2020.

Metzker, M.L. (2010). Sequencing technologies - the next generation. *Nature Reviews* 11(1): 31-46.

Miller, W. *et al* (2011). Genetic diversity and population structure of the endangered marsupial *Sarcophilus harrisii* (Tasmanian devil). *Proc. Natl. Acad. Sci. USA* 108(3): 12348-12353.

Ming, R. *et al* (2008). The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). *Nature* 452(7190): 991-996.

Moore, G.E. (1965). Cramming more components onto integrated circuits. *Electronics* 38(8): 114-117.

Moose, S.P. & Mumm, R.H. (2008). Molecular plant breeding as the foundation for 21st century crop improvement. *Plant Physiology* 147(3): 969-977.

Morgante, M., De Paoli, E. & Radovic, S. (2007). Transposable elements and the plant pan-genomes. *Current Opinion in Plant Biology* 10(2): 149–155.

Morozova, O. & Marra, M.A. (2008). Applications of next-generation sequencing technologies in functional genomics. *Genomics* 92(5): 255-264.

Nelson, J.C. *et al* (2011). Single-nucleotide polymorphism discovery by high-throughput sequencing in sorghum. *BMC Genomics* 12(135): 1-14.

Nordborg, M. & Weigel, D. (2008). Next-generation genetics in plants. *Nature* 456(7223): 720-723.

Opperman, C.H. *et al* (2008). Sequence and genetic map of *Meloidogyne hapla*: compact nematode genome for plant parasitism. *Proc. Natl. Acad. Sci. USA* 105(39): 14802-14807.

Oshima, K. *et al* (2004). Reductive evolution suggested from the complete genome sequence of a plant-pathogenic phytoplasma. *Nature Genetics* 36(1): 27-29.

Ossowski, S. *et al* (2008). Sequencing of natural strains of *Arabidopsis thaliana* with short reads. *Genome Research* 18(12): 2024-2033.

Palmer, L.E. *et al* (2003). Maize genome sequencing by methylation filtration. *Science* 302 (5653): 2115-2117.

Paran, I. & Zamir, D. (2003). Quantitative traits in plants: beyond QTL. *Trends In Genetics* 19(6): 303-306.

Paterson, A.H. *et al* (2009). The *Sorghum bicolor* genome and the diversification of grasses. *Nature* 457(7229): 551-556.

Poland, J.A. *et al* (2011). Genome-wide nested association mapping of quantitative resistance to northern leaf blight in maize. *Proc. Natl. Acad. Sci. USA* 108(17): 6893-6898.

Rafalski, A. *et al* (1998). New experimental and computational approaches to the analysis of gene expression. *Acta Biochimica Polonica* 45(4): 929-934.

Rafalski, A. (2002). Applications of single nucleotide polymorphisms in crop genetics. *Current Opinion in Plant Biology* 5(2): 94-100.

Rafalski, A. (2010). Association genetics in crop improvement. *Current Opinion in Plant Biology* 13(2): 174-180.

Reinartz, J. *et al* (2002). Massively parallel signature sequencing (MPSS) as a tool for in-depth quantitative gene expression profiling in all organisms. *Briefings in Functional Genomics & Proteomics* 1(1): 95-104.

Reinders, J. & Paszkowski, J. (2009). Unlocking the Arabidopsis epigenome. *Epigenetics* 4(8): 557-563.

Reinhardt, J.A. *et al* (2009). *De novo* assembly using low-coverage short read sequence data from the pathogen *Pseudomonas syringae* pv. oryzae. *Genome Research* 19(2): 294-305.

Rensing, S.A. *et al* (2008). The *Physcomitrella* genome reveals evolutionary insights into the conquest of land by plants. *Science* 319(5859): 64-69.

Rensink, W.A. & Buell CR. (2005). Microarray expression profiling resources for plant genomics. *Trends in Plant Science* 10(12): 603-609.

Riesenfeld, C., Schloss PD, & Handelsman J. (2004). Metagenomics: genomic analysis of microbial communities. *Annual Review of Genetics* 38: 525–552.

Risch, N. & Merikangas, K. (1996). The future of genetic studies of complex human diseases. *Science* 273(5281): 1516-1517.

Roh, J.Y. *et al* (2007). *Bacillus thuringiensis* as a specific, safe, and effective tool for insect pest control. *Journal of microbiology and biotechnology* 17(4): 547–559.

Rothberg, J.M. *et al* (2011). An integrated semiconductor device enabling non-optical genome sequencing. *Nature* 475(7356): 348-352.

Rounsley, S. *et al* (2009). *De novo* Next Generation Sequencing of Plant Genomes. *Rice* 2(1): 35-43.

Ruta, N. *et al* (*2010).* QTLs for the elongation of axile and lateral roots of maize in response to low water potential. *Theoretical & Applied Genetics* 120(3): 621-631.

Salanoubat, M. *et al* (2002). Genome sequence of the plant pathogen *Ralstonia solanacearum*. *Nature* 415(6871): 497-502.

Salvi, S. & Tuberosa, R. (2005). To clone or not to clone plant QTLs: present and future challenges. *Trends in Plant Science* 10(6): 297-304.

Sanger, F., Nicklen, S. & Coulson, A.R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. USA* 74(12): 5463–5467.

Sato, S. *et al* (2008). Genome of the legume, Lotus japonicus. *DNA Research* 15(4): 227-239.

Schmutz, J. *et al* (2010). Genome sequence of the palaeop soybean. *Nature* 463(7278): 178-183.

Schnable, P.S. *et al* (2009). The B73 maize genome: complexity, diversity, and dynamics. *Science* 326(5956): 1112-1115.

Schneeberger, K. & Weigel, D. (2011). Fast-forward genetics enabled by new sequencing technologies. *Trends in Plant Science* 16(5): 282-288.

Schöb, H. & Grossniklaus, U. (2006). The first high-resolution DNA "methylome". *Cell* 126(6): 1025-1028.

Schuster, S.C. (2008). Next-generation sequencing transforms today's biology. *Nature Methods* 5(1): 16-18.

Severin, A.J. *et al* (2010). RNA-Seq Atlas of *Glycine max*: a guide to the soybean transcriptome. *BMC Plant Biology* 10(160): 1-16.

Shendure, J. & Ji, H. (2008). Next-generation DNA sequencing. *Nature Biotechnology* 26(10): 1135-1145.

Shulaev, V. *et al (2011).* The genome of woodland strawberry (*Fragaria vesca*). *Nature Genetics* 43(2): 109-116.

Simon, S.A. *et al* (2009). Short-read sequencing technologies for transcriptional analyses. *Annual Review of Plant Biology* 60: 305-333.

Simpson, A.J. *et al* (2000). The genome sequence of the plant pathogen *Xylella fastidiosa*. *Nature* 406(6792): 151-157.

Soderlund, C., Longden, I. & Mott, R. (1997). FPC: a system for building contigs from restriction fingerprinted clones. *Comparative and Applied Biosciences* 13(5): 523-535.

Stavrinides, J. (2009). Origin and evolution of phytopathogenic bacteria. In: *Plant Pathogenic Bacteria: Genomics and Molecular Biology*. R.W. Jackson (Ed). 330. Caister Academic Press. ISBN 9781904455370. UK.

Steuernagel, B. *et al* (2009). *De novo* 454 sequencing of barcoded BAC pools for comprehensive gene survey and genome analysis in the complex genome of barley. *BMC Genomics* 10(547): 1-15.

Stratton, M.R., Campbell, P.J. & Futreal, P.A. (2009). The Cancer Genome. *Nature* 458(7239): 719-724.

Swanson-Wagner, R.A. *et al* (2009). Paternal dominance of trans-eQTL influences gene expression patterns in maize hybrids. *Science* 326(5956): 1118-1120.

The international Brachypodium Initiative. (2010). Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature* 463(7282): 763-768.

The Potato Genome Sequencing Consortium. (2011). Genome sequence and analysis of the tuber crop potato. *Nature* 475(7355): 189-195.

Thornsberry, J.M. *et al* (2001). Dwarf8 polymorphisms associate with variation in flowering time. *Nature Genetics* 28(3): 286–289.

Tian, F. *et al* (2011). Genome-wide association study of leaf architecture in the maize nested association mapping population. *Nature Genetics* 43(2): 159-162.

Trainor, G.L. (1990). DNA sequencing, Automation and the Human Genome. *Analytical Chemistry* 62(5): 418-426.

Tribolium Genome Sequencing Consortium (2008). The genome of the model beetle and pest *Tribolium castaneum*. *Nature* 452(7190): 949-955.

Trick, M. *et al* (2009). Single nucleotide polymorphism (SNP) discovery in the polyploid *Brassica napus* using Solexa transcriptome sequencing. *Plant Biotechnology Journal* 7(4): 334–346.

Tuskan, G.A. *et al* (2006). The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* 313(5793): 1596-1604.

van Oeveren, J. (2011). Sequence-based physical mapping of complex genomes by whole genome profiling. *Genome Research* 21(4): 618-625.

Varshney, R.K. *et al* (2009). Next-generation sequencing technologies and their implications for crop genetics and breeding. *Trends in Biotechnology* 27(9): 522-530.

Vaucheret, H. (2006). Post-transcriptional small RNA pathways in plants: Mechanisms and regulations. *Genes & Development* 20(7): 759-771.

Velasco, R. *et al* (2010). The genome of the domesticated apple (*Malus x domestica* Borkh.). *Nature Genetics* 42(10): 833-839.

Velculescu, V.E. *et al* (1995). Serial analysis of gene expression. *Science* 270(5235): 484–487.

Venter, J.C. *et al* (2001). Sequence of the human genome. *Science* 291(5507): 1304–1351.

Wang, Z., Gerstein, M. & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics* 10(1): 57–63.

Wang, L., Li, P. & Brutnell, T.P. (2010). Exploring plant transcriptomes using ultra high-throughput sequencing. *Briefings in Functional Genomics* 9(2): 118-128.

Weigel, D. & Mott, R. (2009). The 1001 genomes project for *Arabidopsis thaliana*. *Genome Biology* 10(5): 107.

West, M.A. *et al* (2007). Global eQTL mapping reveals the complex genetic architecture of transcript-level variation in Arabidopsis. *Genetics* 175(3): 1441-1450.

Wood, D.W. *et al* (2001). The genome of the natural genetic engineer *Agrobacterium tumefaciens* C58. *Science* 294(5550): 2317-2323.

Yano, M. *et al* (1997). Identification of quantitative trait loci controlling heading date in rice using a high-density linkage map. *Theoretical and Applied Genetics* 95: 1025–1032.

You, F.M. *et al* (2011). Annotation-based genome-wide SNP discovery in the large and complex *Aegilops tauschii* genome using next-generation sequencing without a reference genome sequence. *BMC Genomics* 12(59): 1-19.

Yu, J. & Buckler, E.S. (2006). Genetic association mapping and genome organization of maize. *Current Opinions in Biotechnology* 17(2): 155-160.

Zenoni, S. *et al* (2010). Characterization of transcriptional complexity during berry development in *Vitis vinifera* using RNA-Seq. *Plant Physiology* 152(4): 1787-1795.

Zhang, L. *et al* (2009). A genome-wide characterization of microRNA genes in maize. *PLoS Genetics* 5(11), e1000716.

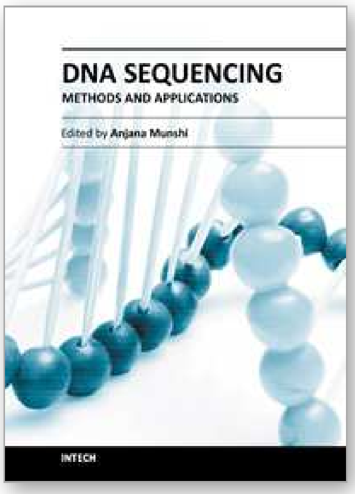
Zhang, G. *et al* (2010). Deep RNA sequencing at single base-pair resolution reveals high complexity of the rice transcriptome. *Genome Research* 20(5): 646–654.

Zhang, X. *et al* (2007). Genome-wide high-resolution mapping and functional analysis of DNA methylation in arabidopsis. *Cell* 126(6): 1189-1201.

Zhang, Y. & Jeltsch, A. (2010). The Application of Next Generation Sequencing in DNA Methylation Analysis. *Genes* 1(1): 85-101.

Zheng, P. *et al* (2008). A phenylalanine in DGAT is a key determinant of oil content and composition in maize. *Nature Genetics* 40(3): 367-372.

Zhu, C. *et al* (2008). Status and prospects of association mapping in plants. *The Plant Genome Journal* 1(1): 5-20.



DNA Sequencing - Methods and Applications
Edited by Dr. Anjana Munshi

ISBN 978-953-51-0564-0
Hard cover, 174 pages
Publisher InTech
Published online 20, April, 2012
Published in print edition April, 2012

This book illustrates methods of DNA sequencing and its application in plant, animal and medical sciences. It has two distinct sections. The one includes 2 chapters devoted to the DNA sequencing methods and the second includes 6 chapters focusing on various applications of this technology. The content of the articles presented in the book is guided by the knowledge and experience of the contributing authors. This book is intended to serve as an important resource and review to the researchers in the field of DNA sequencing.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821