Plant Science 179 (2010) 407–422

Contents lists available at ScienceDirect

Plant Science

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

Review

High throughput DNA sequencing: The new sequencing revolution

Michel Delseny[a,*], Bin Han[b], Yue Ie Hsing[c]

[a] Laboratoire Génome et Développement des Plantes, UMR5096 CNRS-IRD-UP, University of Perpignan, 58 av. Paul Alduy, 66860 Perpignan, France
[b] National Centre for Gene Research, Chinese Academy of Sciences, 500 Caobao Road, Shanghai 200233, China
[c] Institute of Plant and Microbial Biology, Academia Sinica, 128 Academia Road, sec 2, Nankang, Taipei 11529, Taiwan

ARTICLE INFO

Article history:
Received 7 February 2010
Received in revised form 23 July 2010
Accepted 26 July 2010
Available online 3 August 2010

Keywords:
DNA sequencing
Plant genome
Genome structure and evolution
Transposable elements
Epigenome sequencing
Transcriptome sequencing

ABSTRACT

Improvements in technology have rapidly changed the field of DNA sequencing. These improvements are boosted by bio-medical research. Plant science has benefited from this breakthrough, and a number of plant genomes are now available, new biological questions can be approached and new breeding strategies can be designed. The first part of this review aims to briefly describe the principles of the new sequencing methods, many of which are already used in plant laboratories. The second part summarizes the state of plant genome sequencing and illustrates the achievements in the last few years. Although already impressive, these results represent only the beginning of a new genomic era in plant science. Finally we describe some of the exciting discoveries in the structure and evolution of plant genomes made possible by genome sequencing in terms of biodiversity, genome expression and epigenetic regulations. All of these findings have already influenced plant breeding and biodiversity protection. Finally we discuss current trends, challenges and perspectives.

© 2010 Elsevier Ireland Ltd. All rights reserved.

Contents

1. Introduction 408
2. The next-generation sequencing technologies: basic principles 408
   2.1. Sequencing by synthesis after amplification 408
      2.1.1. Pyrosequencing and 454 technology 408
      2.1.2. Illumina (Solexa) technology 410
      2.1.3. SOLiD (Applied Biosystems) and Polonator technologies 410
   2.2. Sequencing from single DNA molecules 410
      2.2.1. Helicos technology 410
      2.2.2. Real-time sequencing with Pacific Biosciences technology 410
      2.2.3. Other emerging technologies 411
   2.3. Advantages and limitations of the new-generation sequencing technologies 411
      2.3.1. Advantages 411
      2.3.2. Limitations and challenges 411
      2.3.3. Improving efficiency and throughput 412
3. An overview of current plant genome projects 413
   3.1. ESTs and BESs 413
   3.2. Plant genome sequence projects 413
      3.2.1. Complete or almost complete genomes 413
      3.2.2. On-going projects 414
      3.2.3. Annotation of the sequences 414

* Corresponding author. Tel.: +33 468 668 848; fax: +33 468 668 499.
E-mail addresses: delseny@univ-perp.fr (M. Delseny), bhan@ncgr.ac.cn (B. Han), bohsing@gate.sinica.edu.tw (Y.I. Hsing).

0168-9452/$ – see front matter © 2010 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.plantsci.2010.07.019

4. Applications of classical sequencing and NGSTs to solve biological problems 415
4.1. Structure and evolution of plant genomes 415
4.2. Transposon inventory and dynamics 416
4.3. Population structure, genetic diversity and genotyping 416
4.4. Targeted sequencing 417
4.5. Sequencing the epigenome 418
4.6. Sequencing the transcriptome 418
5. Conclusion and perspectives 419
Acknowledgements 419
Appendix A. Supplementary data 419
References 419

1. Introduction

During the last 25 years DNA sequencing has completely changed our vision of biology and particularly plant biology. It has been possible to characterize a large number of genes by their nucleotide sequences, thus providing a shortcut to the corresponding protein sequences and their functions. Information on gene polymorphisms has facilitated genetic mapping, gene cloning and the understanding of evolutionary relationships and has allowed for the initiation of biodiversity studies.

The most popular sequencing method has been the Sanger method [1], described in all textbooks. Since its conception, the method has been continually improved. When combined with the use of robots and with concomitant progress in cloning strategies and physical mapping, the method has allowed for sequencing larger DNA fragments and, finally, complete genomes. As a result, a series of landmark genomes were obtained: *Haemophilus influenzae*, *Saccharomyces cerevisiae*, *Escherichia coli*, *Caenorabditis elegans*, *Drosophila melanogaster*, *Arabidopsis thaliana* and, finally, *Homo sapiens* and rice [2–4]. The deciphering of these genomes led to the era of functional genomics and completely modified biological investigation. It also demonstrated that with enough money and human resources, the genomes of other species could be obtained. However, this technology remained tedious and expensive. Sequencing of each of the above eukaryotic genomes cost several million US dollars and mobilized hundreds of scientists all over the world.

These limiting factors have prompted the development of new technologies that allow for many more samples to be analysed at the same time, without prior cloning or mapping work, and at much lower cost. This development has opened new avenues in biology to solve important questions that could not be answered with classical sequencing. The driving force to improve sequencing technology has been human biology and personalized medicine, with the objective being the sequencing of an individual's human genome costing a mere $1000 or less. This goal has not yet been achieved but seems close.

In this paper we review some of the next-generation sequencing technologies (NGST), which are already commercialized, and recent trends. We describe the current situation of genome sequencing in higher plants, to illustrate how the sequencing of plant genomes is modifying our vision of plant evolution, and how NGSTs can be applied to various domains of plant biology, and we identify some of the challenges in developing these technologies.

2. The next-generation sequencing technologies: basic principles

NGSTs are evolving rapidly and have been described in several recent reviews [5–9], including in the plant domain [6,7]. Therefore this section presents only the basic principles. More details are found in these reviews and at the websites of the companies marketing the technologies and instruments. All NGSTs have benefited, to various extent, from new developments in imaging, automation, microprocessing and nanotechnologies, domains that have developed independently of biology. Various ways of reducing costs include avoiding cloning, miniaturizing reactions, use of new chemical procedures, and use of massively parallel sequencing. New sequencing technologies can be grouped into several classes: sequencing by hybridization, sequencing by synthesis from amplified molecules distributed in microarrays and sequencing single molecules. Fig. 1 presents a flow chart of the different methods. However, sequencing by hybridization is a re-sequencing technology that now seems of limited value as compared with the other technologies and will not be described here.

2.1. Sequencing by synthesis after amplification

In this approach, DNA fragments are amplified in clusters, denatured and distributed on microarrays or in microtiter plates that are introduced into a flow cell where the sequencing reactions take place. A primer is extended cyclically by one or a few nucleotides at a time, and the sequence is read at each step of the DNA synthesis. This strategy differs from the Sanger method whereby a whole range of partial copies of the DNA molecules are first synthesized and then analysed. The various methods differ in the strategy used to amplify the sequences, the chemistry used, and the length of the reads. The methods have in common the possibility of sequencing up to several million DNA fragments in parallel.

2.1.1. Pyrosequencing and 454 technology

Roche's 454 technology ([5,8,10] and http://www.454.com/) was the first to be marketed in 2004. It introduced several important innovations. Small DNA fragments (300–800 bp) are ligated to adapters, separated into single strands and bound to small DNA capture beads under conditions favouring a single fragment per bead. The fragments are then amplified by a technique dubbed "PCR-emulsion", whereby each bead is isolated within a droplet of a PCR reaction mixture – in oil emulsion. At the end of amplification, each bead carries several million copies of a unique DNA fragment. This step avoids the cloning of each DNA fragment and the tedious work of colony picking and preparing DNA templates. Then, the emulsion is broken, the DNA is denatured and the beads are deposited in the wells of a PicoTiterPlate. The plate contains millions of wells that are individual reactors for the sequencing reactions that are catalysed by the *Bacillus stearothermophilus* (*Bst*) DNA-polymerase. This device is placed in a flow cell into which reagents are injected. The diameter of the wells is manufactured so that only one bead can be accepted in each well. When a nucleotide is added to the primer by the DNA-polymerase, a pyrophosphate molecule is released. This pyrophosphate is converted into ATP by a sulfurylase and the ATP is used to produce a chemiluminescent signal by the luciferase reaction. This method was therefore

Illumina Exhibit 1058, Page 2 of 16
Illumina v. Molecular Loop Biosciences - IPR2024-00964

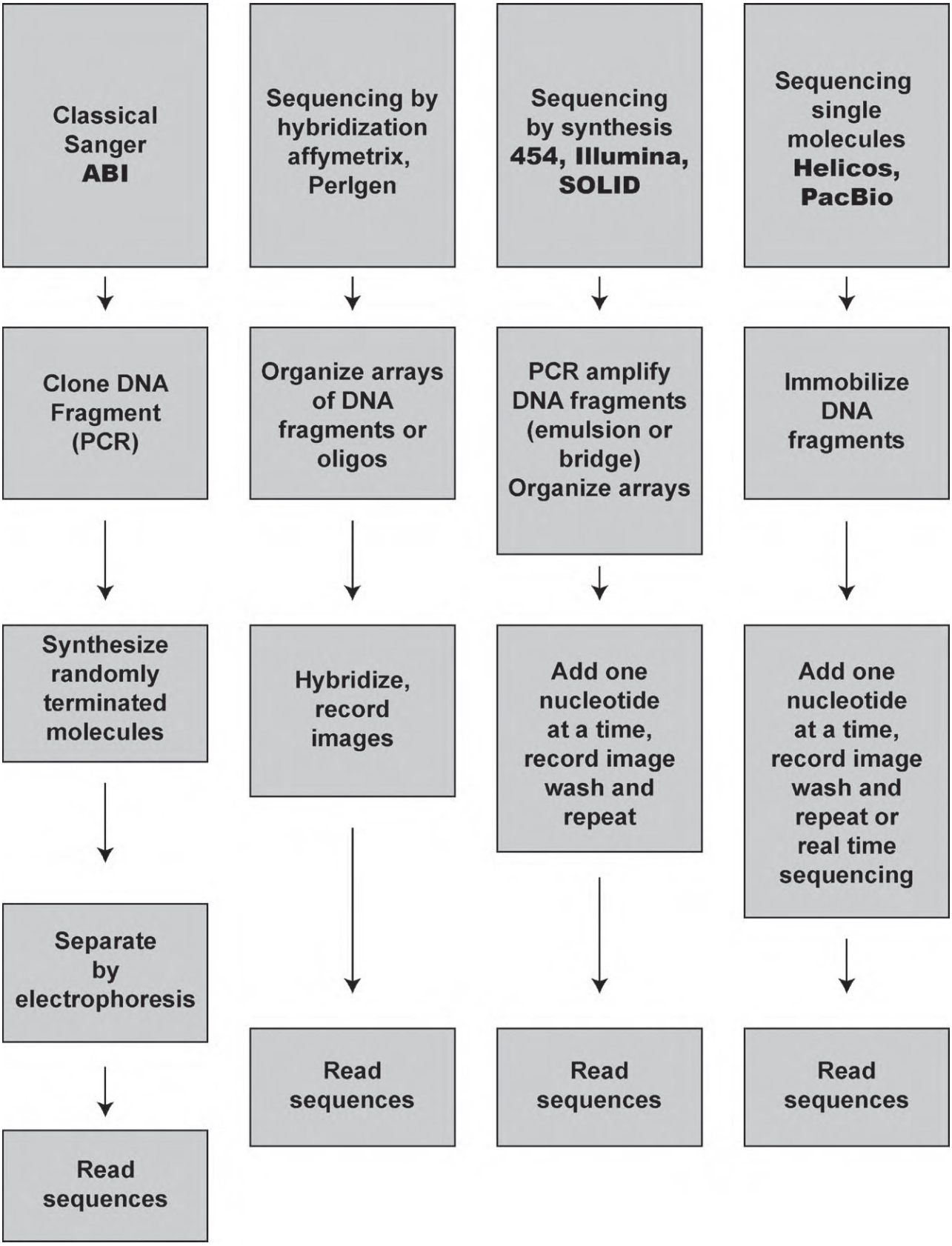

Fig. 1. A flow chart of the different types of sequencing methods.

named "pyrosequencing". First a nucleotide is injected into the flow cell. A chemiluminescent signal is detected in the microwells where the nucleotide has been incorporated and a camera records the signal. Unincorporated nucleotides are then washed away and are replaced by a second nucleotide and the ordered incorporation/recording/washing cycle is repeated with the four nucleotides until the primer has been sufficiently extended. The intensity of the signal is proportional to the number of nucleotides that have been incorporated by the DNA-polymerase. Although the initial machines could read only ~110 bp, current machines with improved design and chemistry can routinely generate 400–500 bp reads and up to 600 Mbp of raw sequence per 10-h run. The latest improvements should allow for ~1000 bp reads.

2.1.2. Illumina (Solexa) technology

The Illumina Genome Analyser (http://www.illumina.com) uses a technology initially developed by the Solexa company. Small DNA fragments are amplified on a solid substrate by "bridge PCR". The DNA fragments are ligated with 5′ and 3′ adaptors and denatured, thus resulting in an *in vitro* library of DNA fragments. The fragments are hybridized to a series of forward and reverse primers immobilized on the substrate that correspond to the adaptors used to prepare the library. During PCR amplification with the *Bst* DNA-polymerase, the amplicons resulting from a given fragment remain attached to the substrate, thus forming a cluster of ~1000 copies of a single fragment. Several million clusters can be accumulated within each of the independent arrays (channels) present in the flow cell, where the sequencing reactions take place. The amplicons are made single-stranded, and a universal sequencing primer is hybridized and extended with modified nucleotides [5,8]. Each modified nucleotide acts as a terminator: it is labelled with a distinct fluorochrome that is chemically cleavable, so that DNA synthesis is reversibly stopped at each position. An image is recorded at each cycle in four channels so that the nucleotide incorporated in each independent cluster can be determined. Then the cleavable moieties of the incorporated nucleotides are removed, and the cycle is repeated, thus allowing the incorporation of the next nucleotide. Read lengths were initially limited to 32–40 bp, but most of the recent machines can read up to 75–150 bp and allow for >100 Gbp of raw sequence to be produced per 4–9 day run.

2.1.3. SOLiD (Applied Biosystems) and Polonator technologies

The initial technology was described in 2005 [5,8,11] and the first machines were released by Applied Biosystems (http://appliedbiosystems.com) in 2007. Adaptor-flanked DNA fragments are denatured and fixed on magnetic beads, one per bead. The library is amplified by emulsion PCR, and each bead carrying a cluster of amplified template is covalently attached to the surface of a glass slide, which is inserted into a flow cell. Each spot on the array is dubbed a "polony" by analogy with the bacterial colonies on a plate. The sequencing strategy relies on ligation of oligonucleotides thus the name *S*equencing by *O*ligo *Li*gation and *D*etection (SOLiD). Sequencing is initiated with a first universal primer (n) complementary to the adaptors used for preparing the DNA fragment library. Then the template is hybridized with a set of 5′-fluorescent semi-degenerate 8-mer oligonucleotides in which the two 3′-terminal bases (positions 1 and 2, di-base) are fixed and all the other nucleotides are random. Each fluorochrome defines a set of 4 of the possible 16 di-bases so that, with this coding system, each base is defined twice. When an oligonucleotide anneals to the template immediately adjacent to the primer, it is ligated with T4-DNA-ligase. The fluorescent signal is recorded in the four channels and identifies the first di-base in positions 1 and 2 in the target sequence. The ligated oligonucleotide is cleaved after base 5, thus releasing the fluorochrome, and a new ligation cycle is carried out to identify positions 6 and 7 of the target sequence. The cycle is repeated several times to allow for stepwise extension of the sequence. To determine the other positions in the sequence, the extended DNA strand is melted and washed away and a new round of annealing/ligation is carried out with a new primer ($n-1$). This new round determines positions 0 and 1, then 5 and 6, 10 and 11 and so on. Additional rounds with primers $n-2$, $n-3$ and $n-4$ allow for determining all the positions in the target sequence. In this system each base is read twice, in independent primer rounds with a different di-base colour code. The machine initially produced ~3000 Mbp of raw sequences in the form of short 35 bp sequences, but can now read 50 bp and produce >50 Gbp per 7–14 day run.

In February 2008 another sequencing machine, the Polonator G.007, involving a similar strategy (http://www.polonator.org), was released at a significantly lower price than any other instrument. However, shorter reads (26 bp) and less raw sequence (12 Gbp) are produced per run. A series of improvements in the ligation strategy have recently been reported by Complete Genomics Inc. (http://www.completegenomics.com/) and have been validated by sequencing of the genomes of three humans [12].

2.2. Sequencing from single DNA molecules

In contrast to the previous technologies, this group of techniques does not require amplification of DNA fragment libraries and can be used to sequence any DNA molecule directly. They are sometimes named third-generation sequencing methods.

2.2.1. Helicos technology

The Helicos platform (http://helicosbio.com) was the first to propose direct sequencing of a single DNA molecule. This breakthrough is essentially due to improved chemistry with highly fluorescent modified nucleotides, which act as reversible terminators. Tails of poly(A) are added to the DNA fragments, which are captured on the surface of an array coated with oligo(dT). Billions of fragments are thus randomly bound to a surface divided into 50 channels and can be directly sequenced by synthesis with a DNA-polymerase when the array is introduced into a flow cell. At each cycle, a given fluorescent nucleotide is injected into the flow cell to allow for template-dependent single-nucleotide extensions of the primers on a fraction of the fragments. The image is recorded, excess unincorporated nucleotides are washed away and the fluorescent moiety of the incorporated nucleotide is removed. This cycle is repeated with each of the four nucleotides until the sequence read is about 30 nt. A machine can produce 30–40 Gbp per 8 day run. The first genomes to be sequenced by this method were M13 bacteriophage [13] and *E. coli*.

2.2.2. Real-time sequencing with Pacific Biosciences technology

Pacific Biosciences (http://pacificbiosciences.com) developed the single molecule real-time (SMRT^{TM}) sequencing technology, involving a chip with several thousand nanoscale wells whereby a single ϕ29-phage DNA-polymerase is immobilized and bound to a single primed DNA template. Hexaphosphate nucleotides labelled on the phosphate moiety with four distinct fluorochromes are added. The nucleotide residence time on the polymerase active site is of the order of a millisecond and allows for a fluorescent pulse to be generated and recorded in real time. The geometry of the nanowell determines a limited detection zone with a volume of 10^{-21} L, conferring both high sensitivity and low-back ground. The formation of a phosphodiester bond by the DNA-polymerase releases the polyphosphate fluorochrome from the detection zone, and a dark interphase is observed until the next nucleotide is positioned on the active site of the polymerase. Three to five nucleotides are incorporated per second [14]. Although up to 10 kb can be read in a single run, routine reads are ~1 kb for 15-min runs. Such a

Illumina Exhibit 1058, Page 4 of 16
Illumina v. Molecular Loop Biosciences - IPR2024-00964

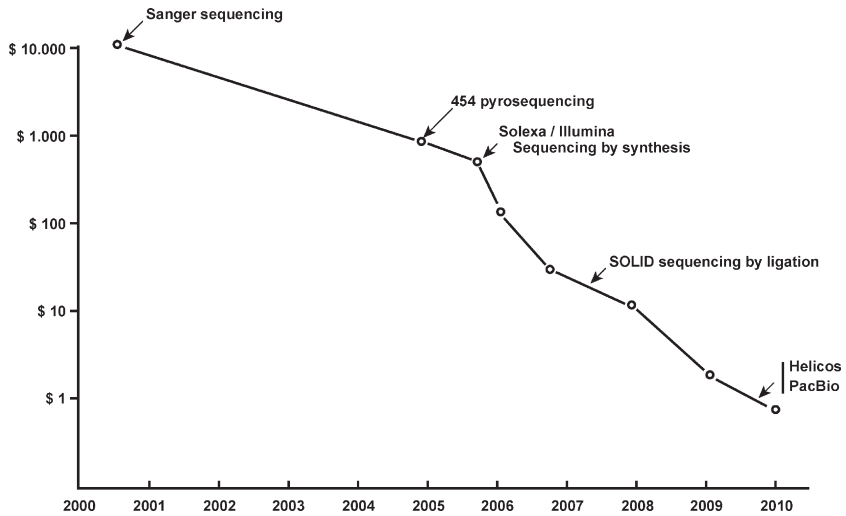

Fig. 2. An estimate of the evolution of sequencing costs over the last 10 years. Costs are given for sequencing a megabase using a logarithmic scale. This curve is adapted from [15]. Time of introduction of new technologies is indicated.

capacity should allow this machine to read several-fold coverage of a human genome in one run.

2.2.3. Other emerging technologies

Several other companies are developing new strategies, based on nano technologies [9], but none has yet reached the commercialization stage.

Life Technologies (http://lifetechnologies.com) is developing an instrument based on real time sequencing whereby nucleotide incorporation is detected by monitoring the interaction between a fluorophore-bearing polymerase and γ-phosphate fluorescent nucleotides by fluorescent resonance energy transfer (FRET). Performance (1.5 kbp reads and 20-min runs) is similar to that of the PacBio machine.

Oxford Nanopore Technologies (http://www.nanoporetech.com) has developed a completely different strategy whereby the DNA molecule to be sequenced is "de-constructed" by an exonuclease in a nanopore. Sequential release of each base induces a disturbance in an electric current through the nanopore that is characteristic of each base type. This electric detection system requires no labeling of the DNA or sophisticated optical device, which is an important breakthrough.

The Ion Torrent systems (http://www.iontorrent.com) uses a semiconductor-based high-density array of microwells whereby each nucleotide incorporation by a DNA-polymerase is specifically recorded by measuring hydrogen ions released as a by-product of DNA synthesis. Currently, the Ion Torrent systems claims 100–200 bp reads for each 1 to 2-h run.

2.3. Advantages and limitations of the new-generation sequencing technologies

2.3.1. Advantages

The major advantage of all the NGSTs is that their throughput is much higher than that of classical sequencing. The most recent machines have a run throughput of >100 Gbp as compared with the 70–100 kbp capacity of the first automated sequencing machines [8]. This capacity was achieved by massive parallel sequencing of hundreds of thousands or millions of templates. A second critical advantage is that the preliminary and tedious cloning work is eliminated and substituted by PCR amplification of DNA fragments. In the most recent technologies, even this step is eliminated, because single DNA molecules are directly sequenced, thus further reducing representation bias in template libraries.

The reduction of reaction volumes and massive parallelism have also reduced the volumes of reagents needed and overall costs. Because of continuous improvement of the technology within the last 10 years, the sequencing costs have decreased from ~$10,000/Mbp with the Sanger method to ~$60 with 454 and ~$1–$2 with the last generation of instruments. The evolution of this cost is illustrated in Fig. 2 [15] and the cost should continue to decrease in the coming months and years. Nevertheless, the instruments remain expensive with costs ranging from $200,000 to $1,000,000. However, new machines without sophisticated optical detection devices should be much cheaper.

Altogether, NGSTs have simplified the sequencing strategies, reduced artefacts, tremendously reduced the speed at which a genome can be sequenced and reduced the cost of sequencing by several orders of magnitude, thus allowing many more samples to be analysed.

2.3.2. Limitations and challenges

Although advantageous, all these technologies have a number of limitations that, at least initially, restricted their use to some applications.

Because most NGSTs produce short reads, they have been initially restricted to re-sequencing new strains or new varieties of a previously sequenced genome to detect point mutations. Repeat regions which are abundant in plant genomes represent another problem because they are difficult to assemble unambiguously when their size exceeds the average read length.

Although the cloning steps have been eliminated, constructions of fragment libraries remain tricky and involve several steps of fragmentation, adaptor ligation and PCR amplification. Each step can introduce representation bias and artefacts, which must be evaluated carefully. Particularly, the size of the fragments needs to be controlled and optimized for efficient sequencing depending on the technology used.

Other limitations are sequencing errors and artefacts, each method having its own weakness. So far, the Sanger method has the highest accuracy because od the high fidelity of *E. coli* DNA-polymerase and remains the "gold standard". Mistakes occur with short homopolymers with the 454 technology because multiple incorporation of the same nucleotide instead of single incorporation can occur at each round. Modified nucleotides can also cause mis-incorporation or block further incorporation if the florescent moiety cannot be completely removed. Errors occur with the SOLiD method even though each base is double-checked. Other difficul-

Illumina Exhibit 1058, Page 5 of 16
Illumina v. Molecular Loop Biosciences - IPR2024-00964

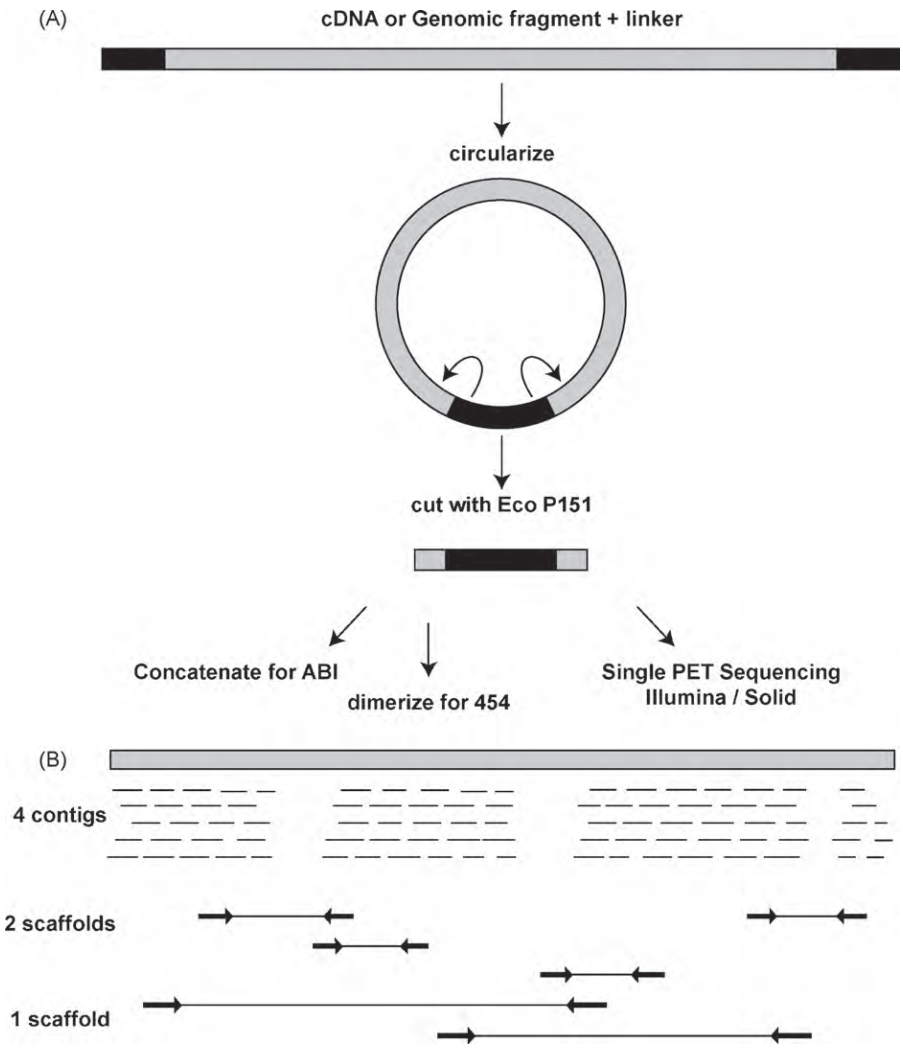

Fig. 3. A) Schematic drawing of the paired-end technology. Adaptors and genome fragments are represented respectively by the black and grey lines. B) Strategy for sequencing large DNA fragments: short reads are assembled into contigs. A high coverage is required. In the next steps, paired-ends derived from larger fragments are used to assemble contigs into scaffolds.

ties are decays in the fluorescent signal and dephasing between the various fragments sequenced in parallel. As a result, the error rate is usually higher at the end of the sequence. All companies have developed quality controls and continually strive to improve protocols and software. Errors are best identified by rare variants in redundant sequencing. The error rate becomes an important issue when the goal of sequencing is precisely to detect rare variants within a gene or a genome as was illustrated in several articles [16–19]. Increasing coverage is one way to identify sequencing errors versus true variants, but more sophisticated statistics, incorporating as much information as possible (e.g., Bayesian techniques that incorporate prior probabilities of observing substitution patterns), will become increasingly important. When a reference genome is available, the short reads are first aligned with it, and all those that do not satisfactorily match are simply discarded from further analysis. An alternative is to combine different methodologies for matching: for example errors in homopolymers with the 454 methods can be corrected by low coverage with Illumina or SOLiD sequencing. The false discovery rate can be estimated by comparing a "gold standard" limited sequence (usually a few Mbp) obtained by Sanger sequencing and the same sequence obtained by an NGST.

The next major challenge is the assembly of short reads into longer sequences. Initially the short reads were aligned to the reference genome, and only unambiguous assignments were considered. This situation eliminated most of the repeated sequences and limited the technology to re-sequencing but is a major drawback for sequencing new, large and complex genomes such as those of higher plants. Although companies provided the software to align short reads and reference sequences, new algorithms were also developed to improve assembly, the key issue in *de novo* sequencing [5,7]. These programs are adapted to each situation (e.g. Newbler is better for 454 data assemblies, whereas Velvet is better for Illumina data). Several new programs overcome some of the difficulties of *de novo* assembly [20–23]. Assembly can also be improved by using reduced representation libraries [22].

Finally a new computational challenge is emerging: the rate of producing raw sequence is now faster than that of the rate of increasing storage capacity of computers. As a consequence, new ways of storing and mining data must be developed; an example is "cloud computing"[24].

2.3.3. *Improving efficiency and throughput*

All companies and sequencing centres regularly update instruments, chemistry and protocols [25].

The major difficulty in sequencing genomes is the assembly of reads into longer contigs. A popular solution with classical sequencing was to sequence the two ends of the same fragment [2,26]. This strategy is now used to construct libraries for NGSTs whereby

Illumina Exhibit 1058, Page 6 of 16
Illumina v. Molecular Loop Biosciences - IPR2024-00964

sequencing paired-ends of fragments of variable size is possible [27–28]. As illustrated in Fig. 3A, paired-end libraries can be generated by circularizing fragments with a linker containing two recognition sites for a type III restriction enzyme such as Eco P151 which cuts 25–27 bp downstream from its recognition site. Paired-end fragments can be sequenced with various instruments, but longer runs are required. Paired-ends allow for assembly of small contigs into larger scaffolds. Scaffolds are assemblies of sequenced contigs that can be defined as neighbours by the presence of "face-to-face" paired-end sequences in two contigs (Fig. 3B).

The high throughput of NGST machines means that their capacity is often too high for sequencing only one library, and techniques have been developed to allow sequencing of several libraries simultaneously. The problem is then to recognize the particular library that a given sequence came from. To solve this difficulty, each library can be bar-coded by a distinct, short oligonucleotide [29]. Most companies now include bar-coding in their protocols, which usually allows sequencing up to 96 samples. This multiplexing process can still be improved for analysis of many more samples by using the position information of different libraries with the same bar-code [30] or by making overlapping pools, whereby each individual library is re-sequenced in different pools [31].

3. An overview of current plant genome projects

This section summarizes the acquisition of sequence information in higher plants, with Sanger sequencing and NGSTs. Several types of sequences information have been accumulated over the last 20 years: short sequence tags represented by expressed sequence tags (ESTs) and BAC-end sequences (BESs) and long assembled genomic sequences. This information is archived in several major databases (see Supplementary Table 1).

3.1. ESTs and BESs

ESTs are single-pass sequences of cDNAs chosen at random in libraries prepared from various tissues, organs or in different physiological conditions. They target only the expressed part of the genome. Most of the sequences in the dbEST database are >600 bp and result from Sanger sequencing. They are generally highly redundant and, unless normalisation has been carried out, their relative numbers reflect the relative abundance of the corresponding mRNAs. Redundancy is useful to assemble overlapping ESTs into contigs, or UniGenes, and eventually into full-length cDNA sequences [32]. These sequences have allowed for production of a cost-effective catalogue of the most abundantly expressed genes before any genome was sequenced. More than eight million plant ESTs are now available in dbEST (Table 1). Currently (July 2010), five plants have >1 million ESTs each, 17 between 700,000 and 200,000, 9 between 200,000 and 100,000 and another 27 more than 50,000. UniGene sets have been organized for most of them, for a preliminary hint of the gene number in these species. NGSTs now allow for producing ESTs at a much faster rate [7]. As an example, 20 million 35-bp Solexa reads were obtained from two cultivars each of *Brassica napus* [33].

ESTs turned out to be essential for annotation of the *Arabidopsis* and rice genomes and to prepare DNA chips before the genomic sequence became available. They will be equally necessary when genomic sequences from more plants are determined. In parallel with EST sequencing, an effort has been made to produce full-length cDNA (FL-cDNA) collections. These are essential not only for annotation, but also for further biological studies and gene manipulations. So far, collections of FL-cDNA produced by Sanger sequencing have been reported only for *Arabidopsis*, rice, soybean and maize [34–35].

Table 1
Top plant species for EST entries in dbEST on 09.07.2010 (http://www.ncbi.nlm.nih.gov/dbEST/).

	EST (09/07/2010 release)	UniGene sets
Total entries in dbEST	66274817	
Homo sapiens	8301249	123200
Zea mays	2019105	97486
Arabidopsis thaliana	1527299	30579
Glycine max	1459639	33001
Oryza sativa	1249124	40978
Triticum aestivum	1071199	40870
Brassica napus	643884	27139
Hordeum vulgare	501616	23595
Panicum virgatum	442269	20973
Phaseolus coccineus	391150	
Vitis vinifera	362193	22083
Pinus taeda	328628	18079
Malus x domestica	324429	23731
Nicotiana tabacum	317769	24069
Picea glauca	313110	22472
Solanum lycopersicum	296955	18346
Medicago truncatula	269238	18785
Gossypium hirsutum	268786	20671
Lotus japonicus	242432	17185
Mimulus guttatus	231095	
Sorghum bicolor	209828	14057
Citrus sinensis	208909	15818
9 species	>100000	
27 species	>50000	

The nine species with >100,000 entries are *Saccharum officinalis*, *Vigna unguiculata* (Cowpea), *Picea sitchensis* (Sitka spruce), *Theobroma cacao*, *Brassica rapa* (Chinese cabbage), *Helianthus annuus*, *Brachipodium distachyon*, *Citrus clementina*, *Capsicum annuum*. The 27 species with >50,000 EST are *Solanum melongena*, *Populus trichocarpa*, *Artemisia annua*, *Arachis hypogea*, *Aquilegia formosa x Aquilegia pubescens*, *Phaseolus vulgaris*, *Raphanus sativus*, *Raphanus raphanistrum*, *Lactuca sativa*, *Manihot esculentua* (Cassava), *Prunus persica*, *Carica papaya*, *Populus tremula x Populus tremuloides*, *Festuca pratensis*, *Gossypium raimondii*, *Ricinus communis*, *Poncirus trifoliata*, *Ipomea nil*, *Cryptomeria japonica*, *Nicotiana benthamiana*, *Citrus reticulata*, *Coffea canephora*, *Lactuca serriola*, *Avena barbata*, *Cichorium intybus*, *Populus trichocarpa x Populus deltoides*, and *Populus nigra*.

BESs are determined at each end of the insert in a BAC clone. BAC libraries have been prepared for many plant species and BES were determined to help physical mapping and genome sequencing. With large BAC libraries, BESs account for a significant portion of the genome sequence, which is sometimes enough to identify new transposable elements, recognize microsatellites or align related genomes [36–37]. BESs are also useful in assembling genomes because they represent long-range links through paired-ends fragments.

3.2. Plant genome sequence projects

3.2.1. Complete or almost complete genomes

Arabidopsis and rice were the first plant genomes to be sequenced. The strategy involved construction of several BAC libraries, physical mapping, and establishment of a minimum tiling path representing the minimum number of partially overlapping BACs required to completely cover the genome. Individual BAC clones were subcloned into shotgun libraries, sequenced and assembled one after the other [2,4], thus providing a highly accurate sequence. Drafts of the rice genome were established independently by a shotgun strategy [38–39]: random DNA fragments of selected size sequenced from both ends and overlapping sequences then assembled into contigs. When end sequences of longer fragments are available (e.g., BESs), the contigs can be organized into scaffolds. So far, drafts of the genomic sequence of eight additional species have been published: poplar tree, grapevine, papaya, sorghum, cucumber, maize, soybean and *Brachypodium* [40–48]. With the exception of cucumber, for which Illumina machines were extensively used, all these genomes were established using

Illumina Exhibit 1058, Page 7 of 16
Illumina v. Molecular Loop Biosciences - IPR2024-00964

the Sanger method, whole-genome shotgun sequencing, or a combination of whole-genome shotgun and map-based sequencing.

The major advantage of the ordered strategy is the delivery of a high-quality sequence, with a minimal error rate (estimated to be 1/40,000 in *Arabidopsis* and rice) and a minimal number of gaps which are essentially located in centromeric, nucleolar organizer and heterochromatic regions. In contrast, many gaps remain in draft sequences, so they are more difficult to use as a reference because their quality is lower. Quality of an assembly is defined by the contig and scaffold N50 values which refer to the size above which 50% of the total sequence can be found. For example, the estimated error rate in the sorghum genome is 1/10,000 and the genome is assembled into 3304 scaffolds with an N50 of 35 (35 scaffolds with a minimum size of 7 Mbp account for 50% of the assembled sequence) and 12,873 contigs with an N50 of 958 (958 contigs longer than 195 kbp account for half of the genome) [44]. After initial publication of the genome, the sequence and assembly quality should improve and be updated. As a result, all reported genomes have not achieved the same quality and a number of standards have been defined [49]. Coverage by contigs and scaffolds is at least 90% in "high-quality drafts", but with little or no manual review. Further steps to improve quality require extensive manual work to resolve gaps, verify the assemblies, correct sequencing errors and annotate. Because these finishing steps cannot be automated, they are expensive. The "gold standard" is the completely finished sequence, with no gaps and less than one error in 100,000 bp. Only the *A. thaliana* Col-0 and *Oryza sativa* Nippon-bare sequences are close to this point.

3.2.2. On-going projects

The development of NGSTs has encouraged a number of scientists to embark on *de novo* sequencing of larger genomes, including those of plants. Although a realistic endeavour, sequencing a new plant genome remains costly and challenging.

The sequencing of many plant genomes is now in progress, and public labs or private companies have announced the complete sequencing of the genomes of *Medicago truncatula*, *Lotus japonicus*, tomato, potato, apple, *Ricinus communis*, oil palm, cassava, strawberry, *Brassica rapa*, *B. oleracea*, *B. napus*, *Eucalyptus grandis* and peach tree. Sequence information and assembly drafts are available in several databases (Supplementary Table 1). For most species, BAC libraries and detailed genetic maps already exist and are helpful in assembling drafts after shotgun sequencing. The availability of protocols to prepare shotgun paired-end libraries for most of the NGSTs facilitates the construction of scaffolds and contigs. Furthermore, if a single technology is insufficient to completely sequence a plant genome *de novo*, a combination of different technologies can be used. Many projects initiated with the 454 technology are being continued with Illumina or SOLiD platforms and will certainly use the latest methods for finishing [23].

The proof of concept of these sequencing and genome assembly strategies is best illustrated by the recent achievement of several individual human and animal sequences by short read technologies [23], as well as by the sequencing of the cucumber genome [45], a barley chromosome [50] and the *Oryza barthii* genome [51]. The general strategy consists in using high coverage with short reads, completed with lower coverage with longer reads or paired-end reads derived from fragments of different sizes. As an example, 42- to 53-bp Illumina reads provided a 68.3× coverage and Sanger sequencing 3.9× coverage in sequencing the cucumber genome [45].

A non-exhaustive list of on-going projects is given in Supplementary Table 2. In addition to the published high-quality draft genomes, at least 18 others have reached the status of standard assembled draft and another 50 are in progress. A few sequenced genomes of interest to plant biologists include algae, fungi, moss, fern, plant bacteria and plant viruses. Therefore, within the next 5 years, each major cultivated plant genome will be available as reasonably well-assembled drafts anchored to the genetic map. Several genomes of wild species may also be sequenced. Genomes of species at the roots of the phylogenetic tree [52] will help clarify evolutionary relationships. Genomes of weeds will be of particular interest because weeds significantly affect agriculture [53]. This deluge of sequencing data opens a completely new perspective both for biology and agriculture that could not be considered just 5 years ago.

Although NGSTs represent a tremendous advance, a number of challenging problems still have to be solved. Even though the cost of sequencing is very low, considerable effort needs to be put into assembly and analysis of the sequence, and this remains expensive. A first challenge is certainly to increase quality, both of the sequences and their assembly. This challenge requires better and new assembly algorithms but might be solved more rapidly by technologies allowing for much longer sequences to be read. A second challenge is to interpret and annotate these sequences. Finally, this new resource must be used to learn more about plant biology to improve crops and better manage the environment.

3.2.3. Annotation of the sequences

Once a sequence has been determined, it should be annotated. A first layer of annotation (or physical annotation) is to determine the number of genes and their precise limits and structure. Another important issue is the identification of repeated sequences and various types of transposable elements. The quality of annotation depends at least in part on the quality of the sequence itself. The first step in annotation is an automated *ab initio* annotation. The process involves various software programs to predict genes along the sequence, but these programs must be trained to adapt to the specifics of a given genome. This *ab initio* annotation is completed by evidence-based annotation carried out on genome browser platforms that integrate all sources of information such as ESTs, FL-cDNAs, gene ontology, transposons, protein sequences and structural motifs, or homologies with other species to predict gene models [54]. Finally, this first automatic step is completed by a much more time-consuming manual annotation by experts in gene families. The first software programs to be developed resulted in considerable confusion: neighbouring genes were fused or single genes were split, intron/exon predictions were not correct, non-protein coding genes were not predicted, and many transposon-derived sequences were initially annotated as genes. In addition, because of alternative splicing, a single gene can give rise to several gene models. This situations explains why the predicted gene numbers in *Arabidopsis* evolved from ∼26,000 [2] to ∼33,520 [55,56], including 27,380 protein-coding genes, 1310 non-coding RNA genes and 4830 pseudogenes or transposable element genes and why the rice gene number fluctuated between ∼56,450 and ∼33,840 [55]. Now, databases of repeat sequences and transposable elements in plants have been established [57–58] and new criteria and tools are available to annotate small RNA genes [59]. With the progress in sequencing new genomes, comparative approaches can also help with annotation [60–62], which can be completed with transcriptomic data, particularly deep RNA sequencing that reveals new exons and new alternative splicing events [63–64], and proteomic data, that sometimes reveal genes that escaped detection [65–66]. This physical annotation process has required almost 10 years to obtain mostly correct structural annotations for the *Arabidopsis* genome with a very high-quality sequence. Yet this process is constantly updated (e.g. TAIR release 9). The experience gained will

Illumina Exhibit 1058, Page 8 of 16
Illumina v. Molecular Loop Biosciences - IPR2024-00964

undoubtedly facilitate annotation of new genomes, but correctly annotating a plant genome sequence will remain a bottleneck for some time.

A second layer of annotation is functional annotation, in which a function is assigned to each gene. This process can also be automated in part with alignment programs that determine homologies with already known genes in the same or another species. This functional annotation is only predictive and needs to be confirmed by "wet" experimentation, which is too often not performed because it is time-consuming. When the *Arabidopsis* sequence was released, about 55% of predicted genes could be assigned a putative function, but less than 5% were experimentally confirmed. Despite a tremendous effort of the whole community to assign a function to each *Arabidopsis* gene before 2010 [67], we are still far from this goal, and at least 20% of the genes still completely lack a precise biochemical or biological function. A call for a similar effort in rice has been made [68] and the situation is even worse for more recently sequenced genomes. The *Arabidopsis* genome annotation is archived on TAIR [56] and useful sites for annotation are described in The *Arabidopsis* Book [69]. Two websites exist for rice (i.e. MSU and RAP [60,70]) and usually one for each new genome (Supplementary Table 1). Another important challenge is to identify regulatory sequences by classical experimentation or comparative genomics.

Besides these major projects aimed at *de novo* sequencing of new genomes and discovering new genes, additional applications of sequencing address biological questions. These are described in the next sections.

4. Applications of classical sequencing and NGSTs to solve biological problems

The availability of several plant genomes has already allowed for addressing a few important questions: How are the plant genomes organized? How do they evolve? And which genes are specific to each species?

NGSTs provide a number of new opportunities such as sequencing several individual genomes within a species to answer questions about genetic diversity at a previously unimaginable scale. The technologies open the way to using sequencing as a genotyping tool. NGSTs also provide transcriptome data with considerably increased depth. Changes in genome structure and sequence modifications in different cells and tissues during development or in response to environmental cues can be investigated. Some of these questions have been recently addressed [6,7,71,72]. In the next sections we focus on what we have learned from analysis of the few available plant genomes and recent developments induced by the widespread use of NGSTs in plant biology laboratories.

4.1. Structure and evolution of plant genomes

A major finding resulted from analysis of the *Arabidopsis* genome: this genome is indeed a patchwork of duplicated segments, presumably resulting from whole-genome duplication (WGD) [2,73]. The various WGD events can be dated by measuring divergence between paralogous pairs of genes (which result from duplication within the ancestral species). In *Arabidopsis*, at least two WGDs were detected, the most recent one being ~20–30 My [74,75]. It rapidly became obvious that *Arabidopsis* was not an isolated case [76] but the most striking observation came from analysis of the grapevine genome [41,42], which was suggested to be derived from a hexaploid ancestor common to all dicots about 120 My ago but had not undergone any recent WGD. Sequencing of papaya and cucumber genomes revealed an organization similar to that of grape [43,45]. More recent WGD has occurred in other lineages: one in poplar and two in *Arabidopsis* and soybean [40,47]. Thus, for an ancestral locus, up to 3 paralogous genes can be observed in grape, papaya and cucumber; 6 in poplar; and 12 in *Arabidopsis.* However, in most cases, several of the copies have been lost during evolution [73,77–79]. Common origin of these genomes is attested by extensive gene co-linearity. Monocot genomes also show evidence of a common WGD preceding the radiation of grasses about 70 My ago, as well as more recent events of polyploidization in maize and wheat [80–82]. These analyses and those of sorghum, maize and *Brachypodium* genomes [44,46,48] have led to the proposal of refined scenarios for the evolution of grass genomes, which suggests an ancestral genome with five chromosomes. In grass, a WGD produced 10 chromosome pairs. Two chromosomes further duplicated, thus yielding a chromosome number of 12. This number was maintained in rice but decreased in other grasses such as sorghum, maize and wheat, which showed evidence of chromosome fusion [79–83].

Additional phenomena could be superimposed on this basic scheme. An apparently recent (~7.7 My) duplication was observed over 3 Mbp in the rice genome at the top of chromosomes 11 and 12. However, further sequencing of this region indicated that this duplication is present in all Oryz*a* species and beyond, in sorghum and *Brachypodium*, which suggests that the high degree of conservation of this region is more the result of concerted evolution [84,85].

The polyploid origin of most plants and its consequences have been extensively discussed, which clarifies that a major driving force in plant genome evolution is the doubling of chromosome numbers, followed by conservation, specialization or loss of duplicated genes associated with chromosome breakage or fusion phenomena [78,86,87]. As a result of this process, paralogous genes are often differentially expressed, and gene loss seems not random: for instance, there is a bias for conservation and specialization of transcription factors [88–89].

The availability of genome sequences for an increasing number of plants allows for determining how many genes are common to all species, how many are specific to a subset and which role these specific genes may play in the adaptation to a specific ecological niche. The most recent analyses compared the sorghum genome to that of rice, *Arabidopsis* and poplar [44] and the maize genome with those of rice, sorghum and *Arabidopsis* [46]. A total of ~9500 gene families are shared by the first four species, thus representing 58% of sorghum genes. Another 4000 are restricted to the 2 grasses, and 1150 are unique to sorghum. In the second study, similar results were obtained, with 465 genes specific to maize, but the availability of the maize data reduced the sorghum-specific genes to 265. These figures highly depend on the quality of the sequence and its annotation. The soybean sequence was also compared to the other available genomes: ~450 soybean genes seem to be legume specific. These comparisons revealed that distinctive features of the different crops are often reflected in their gene catalogue, some specific gene families being expanded in a specific genome. Typical examples are the terpene and polyphenol biosynthetic genes in grape [41] and cucumber [45], genes for C4 photosynthetic pathways in sorghum [90], and genes for lipid metabolism, signaling and nodulation in soybean [47]. Transcription factors have also undergone important expansion in several species: more than 5300 are predicted in soybean and maize, whereas only ~1600 are predicted in *Vitis vinifera* or *Ricinus communis* [91].

The above examples give just a hint of what we have learned about genome structure and gene repertoire because of the acceleration of sequencing programs. The knowledge curve will increase at a much faster speed with the development of NGSTs although limited human resource to analyse the data might be a bottleneck.

Illumina Exhibit 1058, Page 9 of 16
Illumina v. Molecular Loop Biosciences - IPR2024-00964

4.2. Transposon inventory and dynamics

The sequencing of plant genomes not only identified genes but also provided considerable insight into transposable elements. These elements can be organized in two large superfamilies: – long terminal repeat (LTR) retroelements (class I) and classical transposable elements (class II) – depending on whether they required a RNA-mediated copy-and-paste mechanism to create a new copy or a cut-and-paste mechanism to jump from one place to another in the genome [92]. Transposable elements play a major role in genome size variation. Analysis of BESs in cultivated and wild-type rice species revealed that the doubling in size of *Oryza australiensis* with respect to *O. sativa* is essentially due to three bursts of three distinct retroelements [93]. Additional examples were observed in other wild rice species and in cotton [94–95]. A direct comparison of the rice and sorghum genomes [44] revealed size expansion in the latter, although total gene number was about the same. Retroelements are essentially responsible for this increase, ~26% of the rice genome being composed of retrotransposons, whereas sorghum has 55%. A similar additional extension up to 75% was observed in maize [46,96]. Indeed, it contains 8.6% of its genome as class II elements and 75% as class I elements. Class II elements are represented by 855 families of which 82% have been discovered by sequencing and are new families. Class I elements are represented by 468 families. The most abundant and complex class II family is the *Mutator*. The distribution of these transposable elements is not random: most of the class II elements (with the exception of the CACTA transposons) are localized in gene-rich regions, whereas the retroelements are essentially in gene-poor regions and in the centromeric and pericentromeric areas. In rice, the proportion between the two classes is about the same as in maize, and more than 300 retroelement families were identified [58]. The soybean genome contains 42% class I elements, a percentage intermediate between rice and sorghum and maize, thus representing 510 families and ~17% class II elements [47]. Transposable elements represent less than 10% of the *Arabidopsis* genome. Most are incomplete and have undergone small deletions by illegitimate recombination. Homologous recombinations between LTRs result in loss of most of the element and chromosomal rearrangements and are marked by the presence of "solo LTRs". Therefore, LTR-retrotransposons contribute to the expansion of the genomes, but this driving force is counteracted by the mechanisms of silencing and deletion [94–97].

When moving from one place to another in the genome, transposons can capture other genomic sequences and, eventually, give rise to new genes. Such an important role is illustrated by the analysis of MUtator-Like transposable Element (MULE) sequences in the rice genome [98]. More than 3000 MULE sequences are observed in chimeric fragments, called Pack-MULE; they derived from more than 1000 genes. Many are transcribed, and some have given rise to new functional genes, thus illustrating their role in shaping the genome. A similar study of maize [46,96] showed that 385 Pack-MULE transposons acquired fragments from ~420 nuclear gene fragments. Another type of DNA-transposable elements that transposes via a rolling-circle mechanism, the *Helitron*, is also able to capture gene fragments [99–100]. These are more difficult to detect than are other transposable elements because of their high sequence diversity, but recently new software based on structural features has been used to screen some of the available sequences [99]: ~1240 *Helitron* elements were found in *Arabidopsis*, ~7000 in rice and ~4900 in sorghum. Depending on the species, they are organized in 10–23 families. A few have acquired gene fragments.

Obviously, transposable elements have played an important role in shaping plant genomes, particularly those of grasses, but important questions remain: How active are they? How are they silenced and eliminated after they have transposed? And how much do they contribute to the emergence of new genes and functions or to the extinction of genes? Indeed, a direct approach to determine whether transposable elements are moving is to directly compare the sequence of two lines, one control and one induced for transposition, and to describe new insertions. Such approaches have become possible with NGSTs. Transposable elements are normally silenced by methylation and the small interfering (siRNA) silencing pathway. Sequencing of *Arabidopsis* mutants in this pathway (*ddm1* and *met1*/+*nrpd2a*) showed that transposition was reactivated in these genetic backgrounds [101–102]. These two studies directly demonstrate that transposon dynamics are under epigenetic control.

The availability of a large number of sequences from different organisms now allows for direct sequence comparison and examination of sequence phylogenies. Surprisingly, several highly similar sequences of retroelements were observed in genomes such as rice and other grasses that have diverged for several million years. This observation suggests that these sequences arose by horizontal transfer [103].

Various classes of transposable elements accumulate more in centromeric regions. Sequencing of these regions is difficult because of their highly repeated nature. However, the sequence of a few plant centromeres could be elucidated [104–105]. Except for the tandem repeats, the sequences show little homology from one chromosome to another within the same species, and centromeric sequences from homologous chromosomes from two rice species have been shown to diverge rapidly [106]. Centromeres are also defined by epigenetic markers such as the centromere-specific histone variant CenH3 in their chromatin. This feature has been used to immunoprecipitate CenH3 chromatin, extract DNA and sequence centromeric DNA from maize chromosomes 2 and 5 by 454 sequencing [107].

4.3. Population structure, genetic diversity and genotyping

Major challenges in plant breeding are the discovery of genetic variation within a species, assessment of population structure and pattern of linkage disequilibrium. Linkage disequilibrium is a measurement of the non-random association of alleles at different loci. Genetic variation is due to three types of changes: (1) major rearrangements such as translocations and chromosome fusions, (2) insertions/deletions (InDels), (3) single nucleotide polymorphisms (SNPs). SNPs are the most abundant variants. Their discovery initially relied on classical sequencing of PCR fragments and re-sequencing by hybridization. A core collection of 20 *Arabidopsis* ecotypes was analysed by use of an Affymetrix (http://www.affymetrix.com) DNA chip representing the reference sequence. This approach was good enough to produce ~350,000 high-confidence SNPs with a false discovery rate of <2% and to demonstrate that linkage disequilibrium decays within 10 kbp [108]. The low value for linkage disequilibrium means that a collection of diverse *Arabidopsis* lines can provide sufficient resolution for gene discovery by use of genome-wide association mapping [109]. A recent study of flowering time by genome-wide association mapping in natural conditions identified genes regulating the circadian clock but failed to detect an association with the classical flowering-time genes that are important in greenhouse conditions [110]. The impact of this polymorphism is considerable because these SNPs result in >100,000 amino acids changes, of which ~1800 alter protein structure.

A similar approach has been reported for rice (http://www.OryzaSNP.org) [111] and maize [112]. About 100 Mbp of non-repeated sequences in 20 cultivated rice accessions have been analysed. More than 150,000 high-quality SNPs were identified with a false discovery rate estimated at ~8%. On average, between 26,700 and 57,700 SNPs are detected between Japonica/Indica pairs. This analysis revealed haplotype blocks that

Illumina Exhibit 1058, Page 10 of 16
Illumina v. Molecular Loop Biosciences - IPR2024-00964

are larger than that in *Arabidopsis* and of the order of 200 kbp. Approximately 2.7% of the rice genes contained SNPs that are expected to have a large effect because they affect the integrity of the corresponding proteins. Because rice is an important crop, the data set can be used to assess the degree of introgression among varietal groups with an unprecedented resolution.

Although important, these observations give an incomplete view of genetic variation because many SNPs are missed and because InDels and repeats cannot be investigated by hybridization techniques. They will soon be overwhelmed by new data resulting from NGSTs. Re-sequencing of three accessions of *Arabidopsis* (the reference Col-0 accession and two highly divergent accessions, Bur-0 and Tsu-1) by use of Illumina technology has been reported [113]. In fact, this re-sequencing is part of a much larger project, the 1001 genome project [114], which aims to identify all the genotypic and phenotypic variations present in natural accessions. This project consists of re-sequencing a small number of accessions with accuracy close to that of the reference Col-0 genome, then to sequence the accessions examined by the hybridization technique with the NGSTs at a low (8×) coverage. Finally, accessions from 10 geographical regions of Eurasia and one in North Africa will be selected to sample 10 populations in each region; 10 individuals per population will be sequenced ($10 \times 10 \times 10 + 1 = 1001$). Since recombinant inbred lines (RIL) populations of Col-0 × Bur-0 and Col-0 × Tsu-0 (almost identical to Tsu-1) already exist, the data can be immediately used to resolve and clone quantitative trait loci detected in theses crosses. For each accession, 120–173 million 36-bp reads were obtained, representing a 15- to 25-fold coverage. After quality control, 89% of the reads can be readily aligned on the reference sequence. Again, quality controls are important and the authors consider that 67% of the aligned sequences are error-free. The alignments so far reveal 823,000 unique SNPs and 80,000 unique 1- to 3-bp InDels between the divergent accessions and >2000 potential errors in the reference genome. A 3.4-Mbp region of Bur-0 and Tsu-1 is very different from the reference Col-0 genome and includes InDels as large as 641 bp. This re-sequencing strategy was recently used to evaluate the mutation rate in *Arabidopsis*: five lines, which were distant by 30 generations from the original reference Col-0 strain, were re-sequenced and compared [115]. Sensitivity of this approach is such that a spontaneous frameshift mutation in a non-reference ecotype could be identified and characterized [116]. Variation can be identified by sequencing individual genomes and directly from populations as was illustrated by sequencing several populations of *Arabidopsis lyrata*, a project that allowed for identifying putative genes responsible for adaptation to serpentine rich soils [117]. Similar projects are currently in progress with several other plants (Supplementary Table 2). A first-generation haplomap of maize based on re-sequencing 27 diverse maize lines has just been reported [118]. Genetic diversity in the genus *Vitis* was investigated by re-sequencing 10 varieties and 7 wild species, which uncovered a large number of SNPs and allowed for all samples to be distinguished [119]. With the availability of new reference genomes for major crops, no doubt that similar projects will blossom in the next few months and years and will change our approach to plant physiology and breeding. Such data will considerably facilitate important gene identification, by positional gene cloning or association mapping, and will have a considerable effect on plant breeding strategies as was illustrated by recent publications [120–124]. Nevertheless, identifying high-confidence SNPs and small InDels remains a major challenge until base-calling is improved and false discovery rate is reduced.

So far, genotyping has involved mostly microsatellites, a technology with limited resolution and requiring much time and human resources. The ultimate resolution is at the base-pair level and NGSTs now open the possibility to genotype a line by sequencing it, provided a reference genome is available. This strategy has recently been illustrated in rice [125]: in total 150 RILs derived from a cross between 93-11 × Nipponbare, (the two indica and japonica rice reference genomes) were re-sequenced with use of the Illumina machine and the barcode technology. With this multiplexing, the sequencing of the 150 lines could thus be completed in two runs. In all, 1,500,000 SNPs were detected and the authors could map >5000 recombination points (33.8 per line, on average). This resolution is ~35× higher than that available with use of classical markers. The time of acquisition of data is 20× faster than with previous methods. Such an approach can substitute for the genotyping microarray strategy because of its higher resolution.

NGSTs have opened the way to molecular ecology because the very high throughput allows for investigating genomes at the individual and population levels.

4.4. Targeted sequencing

For many applications, it is sufficient to re-sequence DNA fragments. So far, many targeted sequencing projects have addressed comparative sequencing of a locus in different, more or less closely related species. With the intensification of sequencing, more samples can be compared. As an example, the organisation of the *Adh1* locus across nine cultivated and wild rice genomes has been reported [126]. These species have differentiated during the last 15–20 My. A total of 46 genes and 4 pseudogenes were identified, with 38 belonging to 8 multigene families. Analysis of each of these families revealed lineage-specific gain and loss of gene members. Transposable elements had a clear role in mediating replacement of intergenic space, gene disruption and gene movement. Similar conclusions were drawn from a study of the agronomically important *MONOCULM1* locus, which controls tiller number [127].

Because many genomes contain large amounts of repetitive sequences and transposable elements, selecting relevant portions might be of interest. This selection is particularly important in medicine, in which sequencing the same locus in a large range of patients is an important target. This finding has led to the development of methods to capture the desired sequences [128]. These approaches will certainly be applied in plants to assess the genetic diversity associated with a number of genes contributing to important agronomic traits, one of the main goals of the international Generation Challenge Program, http://www.generationcp.org/subprogramme2.php) [129]. In maize, SNPs of 747 genes have already been compared between inbred lines and teosinte [130]. This study demonstrated that two classes of genes could be recognized: those consistent with a domestication bottleneck of moderate intensity and those presenting a strong reduction of genetic diversity due to selection. From the importance of this latter class, an estimated ~1200 genes have been the target of selection.

Specific regions of the genome can be selected by immunoprecipitating chromatin with specific antibodies and analysed by extracting and sequencing the corresponding DNA. NGSTs, with their high coverage, now allow for characterizing discrete fractions of a genome with high confidence. This chromatin immunoprecipitation sequencing (ChIP-seq) [131] has been used for the analysis of maize centromeres [107] and has been used recently to identify the target sequences of the *Arabidopsis* transcription factors SEPALLATA 3 and APETALA 1 [132–133].

Another application of targeted sequencing is for systematics. The sequences of one (or a limited number of) short DNA fragment(s) are used as a "barcode" to identify and distinguish species from each others. Although this concept has rapidly progressed in the animal systematics community, debates within the botanical community concern the fragments to be chosen. However, recent progress indicates that a plant barcode might become a reality fairly soon [134]. After testing a limited number of candidate loci, an

Illumina Exhibit 1058, Page 11 of 16
Illumina v. Molecular Loop Biosciences - IPR2024-00964

international consortium (COBL: Consortium for the Barcode of Life, http://www.barcoding.si.edu) proposed to amplify and sequence two chloroplast DNA loci, *rbcL* and *matK* with universal primers. In total, >900 samples corresponding to 550 species representing the major phyla of higher plants were screened, with 72% success in discrimination. The remaining species were identical to closely related species. Although this number is still far from the ~380,000 named plant species, *rbcL* sequence information has already been obtained for >13,500 plant taxa. Such a barcoding system would have many applications in identifying species in dry or juvenile samples or in mixtures where morphological identification is difficult or impossible.

4.5. Sequencing the epigenome

Genomic sequences can be modified and temporarily silenced by methylation. These modifications can be stable for several generations and can be reversible. They are named epigenetic modifications and they profoundly influence gene expression [135]. Until recently, the epigenetic modifications have been analysed at specific loci or genome-wide, by use of methylation-sensitive and -insensitive restriction enzyme digestions or a combination of immunoprecipitation of DNA with antibodies against methyl-cytosine and hybridization to genome-wide tiled microarrays. However, all these techniques have a limited resolution. DNA methylation can be directly analysed by the Sanger technique, after treating DNA with sodium bisulfite under denaturing conditions. This treatment converts cytosine into uracil but leaves methyl-cytosine intact. On sequencing, unmodified cytosines are read as thymidines and methylated cytosines as cytosines. This method has now been adapted to NGSTs and two groups have characterized the *Arabidopsis* DNA "methylome" at a single base-pair resolution [136–138]. The two groups used slightly different strategies and different tissues (rosette leaves and flower buds), so direct comparison of the two results is limited. Both studies revealed that most methylation sites were previously undetected. The methylation pattern approximately matches the distribution of repeated sequences, but CG methylation is observed almost exclusively in genes. Distribution of methyl-cytosines between CG, CHG and CHH contexts (where H represents any nucleotide) was established and represents 24% of CG, 6.7% of CHG and 1.7% of CHH. Methylation of cytosines in these different contexts is the result of the activities of specific DNA-methyltransferases MET1, CMT3 and DRM1/DRM2 as established by sequencing genomes from corresponding mutants [136]. In the other study, similar global results were obtained. The authors also sequenced several mutants affected in DNA-methyltransferases (*met1* and *drm1*, *drm2 cmt3*), as well as the triple mutant *ros1 dml2 dml3*, which almost completely abolishes demethylation. In this triple mutant, methylation was globally similar to the wild-type Col-0 methylation but some hypermethylated sites were observed in the pericentromeric regions. Most of these mutants do not show an obvious phenotype but have reduced fitness in subsequent generations. In parallel to sequencing the methylome, the small RNAs and the overall transcriptome were deep sequenced. Good correlation was observed between the distribution of methylated sites and the sequence localization of the siRNA, thus pointing to their role in guiding methylation of DNA sequences, particularly in repeated DNA sequences and retrotransposons. In mutants affected in CG methylation, the biogenesis of siRNA is altered. Changing the DNA methylation status in the mutants also altered the distribution of numerous mRNA and transposable element-derived transcripts [137–138]. These data have been further analysed to examine the interactions between epigenetic and evolutionary forces [139]. The results suggest that silencing transposable elements near genes has long-term deleterious effects on neighbouring gene expression and results in preferential loss of methylated transposable elements from gene-rich regions. More recently, the methylome of several eukaryotic genomes, including *Arabidopsis*, rice, poplar, *Physcomitrella*, *Selaginela*, *Chlamydomonas* and *Volvox*, was analysed, thus confirming previous observations in *Arabidopsis* and establishing that higher plant, moss, fern, fungi and animal genomes have distinct methylation patterns [140,141].

Histone modifications and changes in chromatin condensation are also major mechanisms in epigenetic regulation that can also be analysed by NGSTs and the ChIP-seq method. First, DNA and its associated nucleosomes are cross-linked. After fragmentation, specific fractions are immunoprecipitated with antibodies against modified histones [107] and DNA is isolated and prepared so that it can be sequenced by NGSTs.

These are the first genome-wide studies, but many more are expected to follow to examine epigenetic changes in relation to developmental processes or in response to environmental cues. However, handling short reads after bisulfite treatment remains a computational challenge. This situation might rapidly change because the PacBio SMRT™ technology can distinguish 5-methylated cytosine, 5-hydroxymethylcytosine and N6-methyladenine from the corresponding non-methylated base without any treatment [142].

4.6. Sequencing the transcriptome

Sequencing the transcriptome is another major application of high throughput sequencing. Early methods, such as SAGE and MPSS, were limited by the short size of the reads, their cost and labour requirement [143,144]. During the last few years, new protocols with NGSTs have been developed to sequence cDNAs corresponding to mRNAs and a variety of small RNAs and allowing for deep coverage. A direct measurement of the abundance of any RNA is obtained just by counting the number of hits of its sequence: this "digitalized transcriptome" substitutes for classical northern blot or microarray transcriptome evaluations. Most of these advances have recently been reviewed [144] and will therefore not be examined in details here.

The first applications were the characterization of small RNAs, namely microRNA (miRNA), which are 21–22 nt single-stranded RNAs, and small siRNAs, which are usually slighly longer (24 nt) and double-stranded. These approaches identified new, rare miRNAs and particularly those that are species or tissue specific, as well as the miRNAs* (the complementary sequence of a miRNA) and their precursors. Since the last review, a few papers reporting new inventories of small RNAs have been published [145–148]. Small RNAs have been deep sequenced to evaluate their role during genomic shocks caused by interspecific hybridization between *A. thaliana* and *Arabidopsis arenosa* [149]. Related studies have investigated the evolution of miRNAs in the genus *Arabidopsis* [150]. Deep sequencing by NGSTs, combined with ChIP-Seq with specific anti-AGO1 antibodies allowed for partial functional specialization of three of the four rice AGO1 homologs to be demonstrated [151] (AGO1 is the main protein in the complex that recognizes the mRNA target of an miRNA and cleaves it). Finally, by use of graft experiments and *Arabidopsis* mutants in the silencing pathways, it was demonstrated that siRNAs are mobile from the source tissue and direct epigenetic modification in the recipient tissues [152].

The other target for deep sequencing is mRNA. A highly efficient NGST method for sequencing mRNA, RNA-Seq, was reported. The principle is to fragment mRNA or cDNA molecules, add adaptors and sequence. Short reads are then aligned against the genomic reference sequence, classified as exonic reads, junction reads and poly(A)-end reads, and the sequence of mRNA is reconstructed [153]. The strategy has been validated by investigating yeast, *Ara-*

Illumina Exhibit 1058, Page 12 of 16
Illumina v. Molecular Loop Biosciences - IPR2024-00964

bidopsis and rice transcriptomes [63,64,137]. It appears to be more sensitive than microarrays and useful in detecting splicing variants.

At this stage, all transcriptome studies rely on transforming the RNA molecules into cDNA libraries, which induces a number of inconveniences and artefacts such as chimera, incomplete molecules and representation bias. However, this situation might become obsolete because the Helicos technology was recently reported to allow for directly sequencing RNA [154]. Poly(A)-RNA, blocked at its 3′ end, is hybridized to a poly(dT)-coated slide. Poly(dT) is used as a primer for synthesis of a complementary strand to the RNA. Highly fluorescent modified nucleotides are sequentially added as usual for DNA sequencing. Although the technique is still error prone (∼4%), one can obtain reasonably good and reliable tags in the 3′ end of RNA molecules because of its high throughput and sensitivity. Depending on the sample, a poly(A) tail may need to be added before sequencing. The proof of concept of this revolutionary approach has been demonstrated by sequencing a yeast mRNA library starting from as little as 2 ng of total RNA [155]. Such a technology represents a key breakthrough because it avoids most of the drawbacks of sequencing cDNA and allows for preparing RNA from minute amounts of material, such as single cells.

5. Conclusion and perspectives

This review has illustrated some of the fascinating progress in DNA sequencing during the last 5 years. In addition to the current commercialized technologies, others in development, make use of new advances in nanotechnologies and imaging to further improve the power of the technology. Sequencing now costs almost nothing; however, machines are expensive and preparing samples and analysing the data still require a very significant effort as well as highly qualified human resources. In fact, the bottlenecks probably no longer reside in sequencing but rather in archiving, annotating, analysing and interpreting the data. Despite tremendous developments in bioinformatics, with much new software, browsers and graphical interfaces, most classical biologists have not yet realized the tremendous opportunities that are now available and they do not have access to sufficiently powerful computing facilities.

Within the next few years, most, if not all, important crop genomes will be sequenced, and we will have access to a greater portion of their genetic diversity, thus allowing breeders to associate this diversity with phenotypic traits and to continue to engineer new varieties better adapted to a changing environment and to animal, human or industry requirements. Currently, two factors limit the full exploitation of genomic sequences by breeders. The sequences must be anchored to genetic maps to allow breeders to use them in marker-assisted selection, and, most importantly, genes (and their alleles) underlying desired traits still have to be recognized and experimentally validated. The emerging picture is that genomes within a species can be highly variable and the genome of a species is best described by the reference sequence of an individual and an extensive survey of the genetic diversity of the population. Lowering sequencing costs means that SNP and InDel markers will be affordable for most crops, thus allowing breeders to use marker-assisted selection more extensively with unprecedented resolution. Most of the sequencing data are almost immediately freely available through web interfaces and this already completely changes the way plant physiologists, botanists and breeders plan their research and publication strategies.

Sequencing by itself is of limited interest but, as shown in this review, brings much new information of biological interest on genome structure, gene repertoire, genome evolution, and gene expression, as well as population structures and crop species domestication. Thus, the most challenging issue is certainly annotation of the genomes and functional characterization of genes and other DNA sequences. It is already clear that this is just the beginning of a new era: only a few species have been explored and only a limited number of cultivars for each one. High throughput sequencing for transcriptome studies is also just starting: although only a limited number of samples have been examined this has already profoundly changed our vision of the RNA world, with the discovery of many small RNAs and a few other RNA types. For many transcriptome studies, the microarray technology will be progressively substituted by deep sequencing because it is already cheaper, more sensitive and more reliable when a reference genome is available. In the future, cloning and amplification steps could be avoided. Questions such as expression of parental alleles in hybrid individual can now be approached with a previously unattainable resolution. Combining this approach with studies of epigenomes and small RNA expression studies should provide new insights in the heterosis phenomenon. Meanwhile, techniques have been developed to amplify RNA from just a few cells after laser capture microdissection of specific tissues or after cell sorting. Many developments are also expected in the field of evolutionary and comparative genomics, such as discovery of conserved regulatory sequences, as well as in environmental genomics and molecular ecology. Clearly a new era of plant biology is now open because of this sequencing revolution.

Acknowledgements

The authors wish to thank Richard Cooke, François Sabot and Jordi Garcia-Mas for their fruitful comments. They thank Richard Cooke and Laura Small for Revising the English language.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.plantsci.2010.07.019.

References

[1] F. Sanger, G.M. Air, B.G. Barrell, N.L. Brown, A.R. Coulson, J.C. Fiddes, C.A. Hutchison III, P.M. Slocombe, M. Smith, Nucleotide sequence of bacteriophage phi X174 DNA, Nature 265 (1977) 687–695.
[2] The Arabidopsis Genome Initiative, Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*, Nature 408 (2000) 796–815.
[3] International Human Genome Sequencing Consortium, Finishing the euchromatic sequence of the human genome, Nature 431 (2004) 931–945.
[4] International Rice Genome Sequencing Project, The map-based sequence of the rice genome, Nature 436 (2005) 793–800.
[5] J. Shendure, H. Li, Next generation DNA sequencing, Nat. Biotechnol. 26 (2008) 1135–1145.
[6] D. Edwards, J. Batley, Plant genome sequencing: applications for crop improvement, Plant Biotechnol. J. 7 (2009) 1–8.
[7] R.K. Varshney, S.N. Nayak, G.D. May, S.A. Jackson, Next generation sequencing technologies and their implications for crop genetics and breeding, Trends Biotechnol. 27 (2009) 522–530.
[8] M.L. Metzker, Sequencing technologies. The next generation, Nat. Rev. Genet. 11 (2010) 31–46.
[9] D.J. Munroe, T.J.R. Harris, Third-generation sequencing fireworks at Marco Island, Nat. Biotechnol. 28 (2010) 426–428.
[10] J.M. Rothberg, J.H. Leamon, The development and impact of 454 sequencing, Nat. Biotechnol. 26 (2008) 1117–1124.
[11] J. Shendure, G.J. Porreca, N.B. Reppas, et al., Accurate multiplex polony sequencing of an evolved bacterial genome, Science 309 (2005) 1728–1732.
[12] R. Drmanac, A.B. Sparks, M.J. Callow, et al., Human genome sequencing using unchained base reads of self-assembling DNA nanoarrays, Science 327 (2010) 72–74.
[13] T.D. Harris, P.R. Buzby, H. Babcock, et al., Single-molecule DNA sequencing of a viral genome, Science 320 (2008) 106–109.
[14] J. Eid, A. Fehr, J. Gray, et al., Real-time DNA sequencing from single polymerase molecules, Science 323 (2009) 133–138.
[15] Nature editorial, Human genome at ten: the sequence explosion, Nature 464 (2010) 967.
[16] L.W. Hillier, G.T. Marth, A.R. Quinlan, et al., Whole genome sequencing and variant discovery in *C. elegans*, Nat. Methods 5 (2008) 183–188.

Illumina Exhibit 1058, Page 13 of 16
Illumina v. Molecular Loop Biosciences - IPR2024-00964

[17] C.P. Van Tassell, T.P.L. Smith, L. Matukumalli, et al., SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries, Nat. Methods 5 (2008) 247–252.
[18] S. Deschamps, M.A. Campbell, Utilisation of next generation sequencing platforms in plant genomics and genetic variant discovery, Mol. Breed. 25 (2010) 553–570.
[19] V. Bansal, O. Harismendy, R. Tewhey, et al., Accurate detection and genotyping of SNP utilising population sequencing data, Genome Res. 20 (2010) 537–545.
[20] J.T. Simpson, K. Wong, S.D. Jackman, J.E. Schein, S.J. Jones, I. Birol, AbySS: a parallel assembler for short read sequence data, Genome Res. 19 (2009) 1117–1123.
[21] R. Li, H. Zhu, J. Ruan, et al., *De novo* assembly of human genomes with massively parallel short read sequencing, Genome Res. 20 (2010) 265–272.
[22] A.L. Young, H.O. Abaan, D. Zerbino, J. Millikin, E. Birney, E.H. Margulies, A new strategy for genome assembling short sequence reads and reduced representation libraries, Genome Res. 20 (2010) 249–256.
[23] M.C. Schatz, A.L. Delcher, S.L. Salzberg, Assembly of large genomes using second generation sequencing, Genome Res. 20 (2010), in press (doi:10.1101/gr.101360.109).
[24] L.D. Stein, The case for cloud computing in genome informatics, Genome Biol. 11 (2010) 207, doi:10.1186/gb-2010-11-5-207.
[25] M.A. Quail, I. Kozarewa, F. Smith, A. Scally, P.J. Stephens, R. Durbin, H. Swerdlow, D.J. Turner, A large genome center's improvements to the Illumina sequencing system, Nat. Methods 5 (2008) 1005–1010.
[26] J.C. Venter, H.O. Smith, L. Hood, A new strategy for genome sequencing, Nature 381 (1996) 364–366.
[27] J.O. Korbel, A. Eckehart-Urban, J.P. Affourit, et al., Paired-end mapping reveals extensive structural variation in the human genome, Science 318 (2007) 420–426.
[28] M.J. Fullwood, C.L. Wei, E.T. Liu, et al., Next generation DNA sequencing of paired end tags (PET) for transcriptome and genome analyses, Genome Res. 19 (2009) 521–532.
[29] D.W. Craig, J.V. Pearson, S. Szelinger, et al., Identification of genetic variants using bar-coded multiplexed sequencing, Nat. Methods 5 (2008) 887–893.
[30] Y. Erlich, K. Chang, A. Gordon, et al., DNA Sudoku-harnessing high throughput sequencing for multiplexed specimen analysis, Genome Res. 19 (2009) 1243–1253.
[31] S. Prabhu, I. Pe'er, Overlapping pools for high-throughput targeted resequencing, Genome Res. 19 (2009) 1254–1261.
[32] J. Quackenbush, J. Cho, D. Lee, et al., The TIGR gene indices: analysis of gene transcript sequences in highly sampled eukaryotic species, Nucleic Acids Res. 29 (2001) 159–164.
[33] M. Trick, Y. Long, J. Meng, I. Bancroft, Single nucleotide polymorphism (SNP) discovery in the polyploid *Brassica napus* using Solexa transcriptome sequencing, Plant Biotechnol. J. 7 (2009) 334–3464.
[34] T. Umezawa, T. Sakurai, Y. Totoki, et al., Sequencing and analysis of approximately 40,000 soybean cDNA clones from a full length-enriched cDNA library, DNA Res. 15 (2008) 333–346.
[35] C. Soderlund, A. Descour, D. Kudnra, et al., Sequencing, mapping and analysis of 27,455 maize full length cDNA, PLoS Genet. 5 (2009) e1000740.
[36] J.S.S. Ammiraju, M. Luo, J.L. Goicochea, et al., The *Oryza* bacterial artificial chromosome library resource: construction and analysis of 12 deep-coverage large-insert BAC libraries that represent 10 genome types of the genus *Oryza*, Genome Res. 16 (2006) 140–147.
[37] N. Huo, G.R. Lazo, J.P. Vogel, et al., The nuclear genome of *Brachypodium distachyon*: analysis of BAC end sequences, Funct. Integr. Genomics 8 (2007) 135–147.
[38] J. Yu, S. Hu, J. Wang, et al., A draft sequence of the rice genome (*Oryza sativa* L ssp *Indica*), Science 296 (2002) 79–92.
[39] S.A. Goff, D. Ricke, T.H. Lan, et al., A draft sequence of the rice genome (*Oryza sativa* L *Japonica*), Science 296 (2002) 92–100.
[40] G.A. Tuskan, D. DiFazio, S. Janssson, et al., The genome of black cottonwood *Populus trichocarpa* (Torr. & Gray), Science 313 (2006) 1596–1616.
[41] O. Jaillon, J.M. Aury, B. Noel, et al., The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla, Nature 449 (2007) 463–467.
[42] A. Zharkikh, M. Troggio, D. Pruss, et al., Sequencing and assembly of highly heterozygous genome of *Vitis vinifera L.* cv Pinot Noir: problems and solutions, J. Biotechnol. 136 (2008) 38–43.
[43] R. Ming, S. Hou, Y. Feng, et al., The draft genome of the transgenic tropical fruit tree Papaya (*Carica papaya* Linnaeus), Nature 452 (2008) 991–996.
[44] A.H. Paterson, J.E. Bowers, R. Bruggmann, et al., The *Sorghum bicolor* genome and the diversification of grasses, Nature 457 (2009) 551–556.
[45] S. Huang, R. Li, Z. Zhang, et al., The genome of the cucumber, *Cucumis sativus* L., Nat. Genet. 41 (2009) 1275–1281.
[46] P.S. Schnable, D. Ware, R.S. Fulton, et al., The B73 maize genome: complexity, diversity and dynamics, Science 326 (2009) 1112–1115.
[47] J. Schmutz, S.B. Cannon, J. Schlueter, et al., Genome sequence of the palaeopolyploid soybean, Nature 463 (2010) 178–183.
[48] The International *Brachypodium* Initiative, Genome sequence analysis of the model grass *Brachypodium distachyon*: insights into grass genome evolution, Nature 463 (2010) 763–768.
[49] P.S.G. Chain, D.V. Grafham, R.S. Fulton, et al., Genome project standards in a new era of sequencing, Science 326 (2009) 236–237.
[50] K.F.X. Mayer, S. Taudien, M. Martis, et al., Gene content and virtual gene order of Barley chromosome 1H, Plant Physiol. 151 (2009) 496–505.
[51] S. Rounsley, P.R. Marri, Y. Yu, et al., *De novo* next generation sequencing of plant genomes, Rice 2 (2009) 35–43.
[52] D.E. Soltis, V.A. Albert, J. Leebens-Mack, et al., The Amborella genome: an evolutionary reference for plant biology, Genome Biol. 9 (2008) 402, doi: 10.1186:gb-2008-9-3-402.
[53] C.N. Stewart (Ed.), Weedy and Invasive Plant Genomics, Wiley–Blackwell, Ames, USA, 2009, p. 272.
[54] M.R. Brent, Steady progress and recent breakthroughs in the accuracy of automated genome annotation, Nat. Rev. Genet. 9 (2008) 62–73.
[55] C. Liang, L. Mao, D. Ware, L. Stein, Evidence-based gene predictions in plant genomes, Genome Res. 19 (2009) 1912–1923.
[56] D. Swarbreck, C. Wilks, P. Lamessch, et al., The Arabidopsis Information Resource (TAIR): gene structure and function annotation, Nucleic Acids Res. 36 (2008) D1009–D1014.
[57] S. Ouyang, C.R. Buell, The TIGR Plant Repeat Databases: a collective resource for the identification of repetitive sequences in plants, Nucleic Acids Res. 32 (2004) D360–D363.
[58] C. Chaparro, R. Guyot, A. Zuccolo, B. Piegu, O. Panaud, RetrOryza: a database of the rice LTR-retrotransposons, Nucleic Acids Res. 35 (2007) D66–D70.
[59] B.C. Meyers, M.J. Atxell, B. Bartel, et al., Criteria for annotation of plant microRNA, Plant Cell 20 (2008) 3186–3190.
[60] Rice Annotation Project, Curated genome annotation of *Oryza sativa* ssp japonica and comparative genome analysis with *Arabidopsis thaliana*, Genome Res. 17 (2007) 175–183.
[61] W. Zhu, C.R. Buell, Improvement of whole genome annotation of cereals through comparative analyses, Genome Res. 17 (2008) 299–310.
[62] S. Proos, M. Van Bel, L. Sterck, K. Billau, T. Van Parys, Y. Van de Peer, K. Vandepoele, PLAZA: a comparative genomics resource to study gene and genome evolution, Plant Cell 21 (2009) 3718–3731.
[63] S.A. Filichkin, H.D. Priest, S.A. Givan, et al., Genome wide mapping of alternative splicing in *Arabidopsis thaliana*, Genome Res. 20 (2010) 45–58.
[64] G. Zhang, G. Guo, X. Hu, et al., Deep RNA sequencing at single base-pair resolution reveals high complexity of the rice transcriptome, Genome Res. 20 (2010) 646–654.
[65] N.E. Castellana, S.H. Payne, Z. Shen, M. Stanke, V. Bafna, S.P. Briggs, Discovery and revision of Arabidopsis genes by proteogenomics, Proc. Natl. Acad. Sci. USA 105 (2008) 21034–21038.
[66] M.A. Grobei, E. Qeli, E. Brunner, et al., Deterministic protein inference for shotgun proteomic data provides new insights into *Arabidopsis* pollen development and function, Genome Res. 19 (2009) 1786–1800.
[67] C. Somerville, J. Dangl, Plant biology in 2010, Science 290 (2000) 2077–2078.
[68] Q. Zhang, J. Li, X. Xue, B. Han, X.W. Deng, Rice 2020: a call for an international coordinated effort in rice functional genomics, Mol. Plant 1 (2008) 715–719.
[69] Y. Lu, R.L. Last, Web-based Arabidopsis functional and structural genomics resources, The Arabidopsis Book (2008), doi:10.1199/tab0118.
[70] S. Ouyang, W. Zhu, J. Hamilton, et al., The TIGR rice genome annotation resource: improvements and new features, Nucleic Acids Res. 35 (2007) D883–D887.
[71] A. Kahvejian, J. Quackenbush, J.F. Thompson, What would you do if you could sequence everything? Nat. Biotechnol. 26 (2008) 1125–1133.
[72] R. Lister, B.D. Gregory, J.R. Ecker, Next is now: new technologies for sequencing of genomes, transcriptomes and beyond, Curr. Opin. Plant Biol. 12 (2009) 107–118.
[73] G. Blanc, A. Barakat, R. Guyot, R. Cooke, M. Delseny, Extensive duplication and reshuffling in the Arabidopsis genome, Plant Cell 12 (2000) 1093–1101.
[74] G. Blanc, K. Hokamp, K.H. Wolfe, A recent polyploidy superimposed on older large scale duplications in the Arabidopsis genome, Genome Res. 13 (2003) 137–144.
[75] H. Tang, X. Wang, J.E. Bowers, R. Ming, M. Alam, A.H. Paterson, Unravelling ancient hexaploidy through multiply-aligned angiosperm gene maps, Genome Res. 18 (2008) 1944–1954.
[76] M. Delseny, Re-evaluating the relevance of ancestral shared synteny as a tool for crop improvement, Curr. Opin. Plant Biol. 7 (2004) 126–131.
[77] Y. Van de Peer, J.A. Fawcett, S. Proost, L. Sterck, K. Vandepoele, The flowering world: a tale of duplications, Trends Plant Sci. 14 (2009) 680–688.
[78] A.H. Paterson, J.E. Bowers, B.A. Chapman, Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics, Proc. Natl. Acad. Sci. USA 101 (2004) 9903–9908.
[79] J. Salse, S. Belot, M. Troude, et al., Identification and characterization of shared duplications between rice and wheat provide new insight into grass genome evolution, Plant Cell 20 (2007) 11–24.
[80] F. Wei, E. Coe, W. Nelson, et al., Physical and genetic structure of the maize genome reflects its complex evolutionary history, PLoS Genet. 3 (2007) e123.
[81] H. Tang, J.E. Bowers, X. Wang, A.H. Paterson, Angiosperm genome comparison reveal early polyploidy in the monocot lineage, Proc. Natl. Acad. Sci. USA 107 (2010) 472–477.
[82] J. Salse, M. Abrouk, S. Bolot, et al., Reconstruction of monocotyledoneous proto-chromosomes reveals faster evolution in plants than in animals, Proc. Natl. Acad. Sci. USA 106 (2009) 14908–14913.
[83] S. Bolot, M. Abrouk, U. Masood-Quraishi, N. Stein, J. Messing, C. Feuillet, J. Salse, The "inner circle" of the cereal genomes, Curr. Opin. Plant Biol. 12 (2009) 1–7.
[84] X. Wang, H. Tang, J.E. Bowers, F.A. Feltus, A.H. Paterson, Extensive concerted evolution of rice paralogs and the road to regaining independence, Genetics 177 (2007) 1753–1763.

Illumina Exhibit 1058, Page 14 of 16
Illumina v. Molecular Loop Biosciences - IPR2024-00964

[85] J. Jacquemin, M. Laudié, R. Cooke, A recent duplication revisited: phylogenetic analysis reveals an ancestral duplication highly conserved throughout the *Oryza* genus and beyond, BMC Plant Biol. 9 (2009) 146.
[86] J.J. Doyle, L.E. Flagel, A.H. Paterson, R.A. Rapp, D.E. Soltis, P.S. Soltis, J.F. Wendel, Evolutionary genetics of genome merger and doubling in plants, Annu. Rev. Genet. 42 (2008) 443–461.
[87] M. Freeling, Bias in plant gene content following different sorts of duplication: tandem, whole genome, segmental, or by transposition, Annu. Rev. Plant Biol. 60 (2009) 433–453.
[88] T. Casneuf, S. De Bodt, J. Raes, S. Maere, Y. van de Peer, Non random divergence of gene expression following gene and genome duplications in the flowering plant *Arabidopsis thaliana*, Genome Biol. 7 (2006) R13, doi:10.1186/gb-2006-7-2-r13.
[89] M. Throude, S. Bolot, M. Bosio, et al., Structure and expression analysis of rice paleoduplications, Nucleic Acids Res. 37 (2009) 1248–1259.
[90] X. Wang, U. Gowik, H. Tang, J.E. Bowers, P. Westhoff, A.H. Paterson, Comparative genomic analysis of C4 photosynthesis pathway evolution in grass, Genome Biol. 10 (2009) R68.
[91] M. Libault, T. Joshi, V.A. Benedito, D. Xu, M.K. Udvardy, G. Stacey, Legume transcription factor genes: what makes legume so special, Plant Physiol. 151 (2009) 991–1001.
[92] T. Wicker, F. Sabot, A. Hua-van, et al., A unified classification system for eukaryotic transposable elements, Nat. Rev. Genet. 8 (2007) 973–982.
[93] B. Piegu, R. Guyot, N. Picault, et al., Doubling genome size without polyploidization: dynamics of retrotransposon driven genomic expansions in *O. australiensis*, a wild relative of rice, Genome Res. 16 (2006) 1262–1269.
[94] A. Zuccolo, A. Sebastian, J. Talag, et al., Transposable element distribution, abundance and role in genome size variation in the genus Oryza, BMC Evol. Biol. 7 (2007) 152.
[95] J.S. Hawkins, S.R. Proulx, R.A. Rapp, J.F. Wendel, Rapid DNA loss as a counterbalance to genome expansion through retrotransposon proliferation in plants, Proc. Natl. Acad. Sci. USA 106 (2009) 17811–17816.
[96] R.S. Baucom, J.C. Estill, C. Chaparro, N. Upshaw, A. Jog, J.M. Deragon, R.P. Westerman, P.J. San Miguel, J.L. Bennetzen, Exceptional diversity of non random distribution and rapid evolution of retroelements in the B73 maize genome, PloS Genet. 5 (2009) e1000732.
[97] C. Vitte, O. Panaud, H. Quesneville, LTR-retrotransposon in rice (*Oryza sativa* L): recent burst amplifications followed by rapid DNA loss, BMC Genomics 8 (2007) 218.
[98] K. Hanada, V. Vallejo, K. Nobuta, R.K. Slotkin, D. Lisch, B.C. Meyers, S.H. Shiu, N. Jiang, The functional role of Pack-MULEs in rice inferred from purifying selection and expression profile, Plant Cell 21 (2009) 25–28.
[99] L. Yang, J.L. Bennetzen, Distribution, diversity evolution and survival of helitrons in the maize genome, Proc. Natl. Acad. Sci. USA 106 (2009) 19922–19927.
[100] S.M. Oetjens, L.C. Hannah, Helitrons: enigmatic abductors and mobilizers of host genome sequences, Plant Sci. 176 (2009) 181–186.
[101] S. Tsukahara, A. Kobayashi, A. Kawabe, O. Mathieu, A. Miura, T. Kakutami, Bursts of retrotransposition reproduced in Arabidopsis, Nature 461 (2009) 423–426.
[102] M. Mirouze, J. Reinders, E. Bucher, et al., Selective epigenetic control of the retrotransposition in Arabidopsis, Nature 461 (2009) 427–431.
[103] A. Roulin, B. Piégu, P.M. Fortuné, F. Sabot, A. D'Hont, D. Manicacci, O. Panaud, Whole genome surveys of rice, maize and sorghum reveal multiple horizontal transfers of the LTR-retrotransposon Route66 in Poaceae, BMC Evol. Biol. 9 (2009) 58, doi:10.1186/1471-2148-9-58.
[104] J. Ma, R.A. Wing, J.L. Bennetzen, S.A. Jackson, Plant centromere organization: a dynamic structure with conserved functions, Trends Genet. 23 (2007) 134–139.
[105] J.A. Birchler, F. Han, Maize centromeres, structure, function, epigenetics, Annu. Rev. Genet. 43 (2009) 287–303.
[106] J. Wu, M. Fujizawa, Z. Tian, et al., Comparative analysis of complete orthologous centromeres from two subspecies of rice reveals rapid variation of centromere organisation and structure, Plant J. 60 (2009) 805–819.
[107] T.K. Wolfgruber, A. Sharma, K.L. Schneider, et al., Maize centromere structure and evolution sequence analysis of centromeres 2 and 5 reveals dynamic loci shaped primarily by retrotransposons, PloS Genet. 5 (2009) e1000743.
[108] G. Zeller, R.M. Clark, K. Schneeberger, et al., Detecting polymorphic regions in *Arabidopsis thaliana* with resequencing microarrays, Genome Res. 18 (2008) 918–929.
[109] S. Atwell, Y.S. Huang, B.J. Vilhjalmsson, et al., Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines, Nature 465 (2010) 627–631.
[110] B. Brachi, N. Faure, M. Horton, et al., Linkage and association mapping of *Arabidopsis thaliana* flowering time in nature, PLoS Genet. 6 (2010) e1000940.
[111] K.L. McNally, K.L. Childs, R. Bohnert, et al., Genome-wide SNP variation reveals relationships among landraces and modern varieties of rice, Proc. Natl. Acad. Sci. USA 106 (2009) 12273–12278.
[112] N.M. Springer, K. Ying, Y. Fu, et al., Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content, PloS Genet. 5 (2009) e1000734.
[113] S. Ossowski, K. Schneeberger, R.M. Clark, C. Lanz, N. Warthmann, D. Weigel, Sequencing of natural strains of *Arabidopsis thaliana* with short reads, Genome Res. 18 (2008) 2024–2033.
[114] D. Weigel, R. Mott, The 1001 genomes project for *Arabidopsis thaliana*, Genome Biol. 10 (2009) 107, doi:10.1186/gb-2009-10-5-107.
[115] S. Ossowski, K. Schneeberger, J.I. Lucas-Lledo, N. Warthmann, R.M. Clark, R.G. Shaw, D. Weigel, M. Lynch, The rate and molecular spectrum of spontaneous mutation in *Arabidopsis thaliana*, Science 327 (2010) 92–94.
[116] R.A. Laitinen, K. Schneeberger, N.S. Jelly, S. Ossowski, D. Weigel, Identification of a spontaneous frameshift mutation in a non-reference *Arabidopsis thaliana* accession using whole genome sequencing, Plant Physiol. 153 (2010) 652–654.
[117] T.L. Turner, E.C. Bourne, E.J. Von Wettberg, T.T. Hu, S.V. Nuzhdin, Population resequencing reveals local adaptation of *Arabidopsis lyrata* to serpentine soils, Nat. Genet. 42 (2010) 260–263.
[118] M.A. Gore, J.M. Chia, R.J. Elshire, et al., A first generation haploid map of maize, Science 326 (2009) 1115–1117.
[119] S. Myles, J.M. Chia, B. Hurwitz, C. Simon, G.Y. Zhang, E. Buckler, D. Ware, Rapid characterization of the genus *Vitis*, PLoS One 5 (2010) e8219, doi:10.1371/journal.pone0008219.
[120] S. Myles, J. Peiffer, P.J. Brown, E.S. Ersoz, Z. Zhang, D.E. Costich, E.S. Buckler, Association mapping critical considerations shift from genotyping to experimental design, Plant Cell 21 (2009) 2194–2202.
[121] S. Liu, H.D. Chen, I. Makarevitch, R. Shirmer, S.J. Emrich, C.R. Dietrich, W.B. Barbazuk, N.M. Springer, P.S. Schnable, High throughput genetic mapping of mutants via quantitative single nucleotide polymorphism typing, Genetics 184 (2010) 19–26.
[122] S.A. Flint-Garcia, E.S. Buckler, P. Tiffin, E. Ersoz, N.M. Springer, Heterosis is prevalent for multiple traits in diverse maize germplasm, PloS One 4 (2009) e7433.
[123] R.A. Swanson-Wagner, R. DeCock, Y. Jia, T. Bancroft, T. Ji, X. Zhao, D. Nettleton, P.S. Schnable, Paternal dominance of trans e-QTL influences gene expression patterns in maize hybrids, Science 326 (2009) 1118–1120.
[124] J.A. Rafalski, Association genetics for crop improvement, Curr. Opin. Plant Biol. 13 (2010) 174–180.
[125] X. Huang, Q. Feng, Q. Qian, et al., High-throughput genotyping by whole genome sequencing, Genome Res. 19 (2009) 1068–1076.
[126] J.S.S. Ammiraju, F. Lu, A. Sanyal, et al., Dynamic evolution of Oryza genomes is revealed by comparative genomic analysis of a genus-wide vertical data set, Plant Cell 20 (2008) 3191–3209.
[127] F. Lu, J.S.S. Ammiraju, A. Sanyal, et al., Comparative sequence analysis of *MONOCULM1*-orthologous regions in 14 *Oryza* genomes, Proc. Natl. Acad. Sci. USA 106 (2009) 2071–2076.
[128] G.J. Porreca, K. Zhang, J.B. Li, et al., Multiplex amplification of large sets of human exons, Nat. Methods 4 (2007) 931–936.
[129] J.C. Glaszmann, B. Killian, H.D. Upadhyaya, R.K. Varshney, Accessing genetic diversity for crops, Curr. Opin. Plant Biol. 13 (2010) 167–173.
[130] B.I. Wright, I.V. Bi, S.G. Schroeder, M. Yamasaki, J.F. Doebley, M.D. McMullen, B.S. Gaut, The effect of artificial selection on the maize genome, Science 308 (2005) 1310–1314.
[131] P.J. Park, ChIP-seq: advantages and challenges of a maturing technology, Nat. Rev. Genet. 10 (2009) 669–680.
[132] K. Kaufmann, J.M. Muino, R. Jauregui, C.A. Airoldi, C. Smaczniak, P. Krajewski, G. Angenent, Target genes of the MADS transcription factor SEPALLATA3: integration of developmental and hormonal pathways in the *Arabidopsis* flower, PLoS Biol. 7 (2009) e1000090, doi:10.1371/journal.pbio.1000090.
[133] K. Kaufmann, F. Wellmer, J.M. Muino, et al., Orchestration of floral initiation by APETALA 1, Science 328 (2010) 85–89.
[134] CBOL Plant Working Group, A DNA barcode for land plants, Proc. Natl. Acad. Sci. USA 106 (2009) 12794–12797.
[135] J.A. Law, S.E. Jacobsen, Establishing, maintaining and modifying DNA methylation patterns in plants and animals, Nat. Rev. Genet. 11 (2010) 204–220.
[136] S.J. Cokus, S. Feng, X. Zhang, Z. Chen, B. Merriman, C.D. Haudenschild, S. Pradhan, S.F. Nelson, M. Pellegrini, S.E. Jacobsen, Shotgun bisulfite sequencing of the *Arabidopsis* genome reveals DNA methylation patterning, Nature 452 (2008) 215–219.
[137] R. Lister, R.C. O'Malley, J. Tonti-Philippini, B.D. Gregory, C.C. Berry, A.H. Millar, J.R. Ecker, Highly integrated single base resolution maps of the epigenome in *Arabidopsis*, Cell 133 (2008) 523–536.
[138] R. Lister, J.R. Ecker, Finding the fifth base: genome wide sequencing of cytosine methylation, Genome Res. 19 (2009) 959–966.
[139] J.D. Hollister, B.S. Gaut, Epigenetic silencing of transposable elements: a trade-off between reduced transposition and deleterious effects on neighboring gene expression, Genome Res. 19 (2009) 1419–1428.
[140] S. Feng, S.J. Cokus, X. Zhang, et al., Conservation and divergence of methylation patterning in plants and animals, Proc. Natl. Acad. Sci. USA 107 (2010) 8689–8694.
[141] A. Zemach, I.E. McDaniel, P. Silva, D. Zilberman, Genome-wide evolutionary analysis of eukaryotic DNA methylation, Science 328 (2010) 916–919.
[142] B.A. Flusberg, D.R. Webster, J.H. Lee, K.J. Travers, E.C. Olivares, T.A. Clark, J. Korlach, S.W. Turner, Direct detection of DNA methylation during single-molecule, real-time sequencing, Nat. Methods 7 (2010) 461–465.
[143] M.E. Vega- Sanchez, M. Gowda, G.L. Wang, Tag-based approaches to deep transcriptome analysis in plants, Plant Sci. 173 (2007) 373–380.
[144] S.A. Simon, J. Zhai, R.S. Nandety, K.P. McCormick, J. Zeng, D. Melia, B.C. Meyers, Short read sequencing technologies for transcriptional analyses, Annu. Rev. Plant Biol. 60 (2009) 305–333.
[145] X. Zhou, R. Sunkar, H. Jin, et al., Genome-wide identification and analysis of small RNAs originated from natural antisense transcripts in *Oryza sativa*, Genome Res. 19 (2009) 70–78.

Illumina Exhibit 1058, Page 15 of 16
Illumina v. Molecular Loop Biosciences - IPR2024-00964

[146] C. Johnson, A. Kasprzewska, K. Tennessen, et al., Cluster and superclusters of phased small RNAs in the developing inflorescence of rice, Genome Res. 19 (2009) 1429–1440.
[147] C.Z. Zhao, H. Xia, T.P. Frazier, et al., Deep sequencing identifies novel and conserved microRNAs in peanut (*Arachis hypogaea* L), BMC Plant Biol. 10 (2010) 3, doi:10.1186/1471-2229-10-3.
[148] D. Klevebring, N.R. Street, N. Fahlgren, K.D. Kasschau, J.C. Carrington, J. Lundeberg, S. Jansson, Genome-wide profiling of Populus small RNA, BMC Genomics 10 (2009) 620.
[149] M. Ha, J. Lu, L. Tian, et al., Small RNAs serve as a genetic buffer against genomic shock in Arabidopsis interspecific hybrids and allotetraploids, Proc. Natl. Acad. Sci. USA 106 (2009) 17835–17840.
[150] N. Fahlgren, S. Jogdeo, K.D. Kasschau, et al., MicroRNA gene evolution in *Arabidopsis lyrata* and *Arabidopsis thaliana*, Plant Cell 22 (2010) 1074–1089.
[151] L. Wu, Q. Zhang, H. Zhou, F. Ni, X. Wu, Y. Qi, Rice microRNA effector complexes and targets, Plant Cell 21 (2009) 3421–3435.
[152] A. Molnar, C.W. Melnyk, A. Bassett, T.J. Hardcastle, R. Dunn, D.C. Baulcombe, Small silencing RNAs in plants are mobile and direct epigenetic modification in recipient cells, Science 328 (2010) 872–875.
[153] Z. Wang, M. Gerstein, M. Snyder, RNA-Seq: a revolutionary tool for transcriptomics, Nat. Rev. Genet. 10 (2009) 57–63.
[154] F. Ozsolak, A.R. Platt, D.R. Jones, et al., Direct RNA sequencing, Nature 461 (2009) 814–818.
[155] D. Lipson, T. Raz, A. Kieu, et al., Quantification of the yeast transcriptome by single molecule sequencing, Nat. Biotechnol. 27 (2009) 652–658.