plant biology

Plant Biology ISSN 1435-8603

REVIEW ARTICLE

What can next generation sequencing do for you? Next generation sequencing as a valuable tool in plant research

A. Bräutigam[1] & U. Gowik[2]

1 Institute of Plant Biochemistry, Heinrich-Heine University, Düsseldorf, Germany

2 Institute of Developmental and Molecular Biology of Plants, Heinrich-Heine University, Düsseldorf, Germany

Keywords
Application; non-model species; protocol; RNA-seq; transcriptomics.

Correspondence
Andrea Bräutigam, Heinrich-Heine Universität, Gebäude 26.03.01, Universitätsstraße 1, 40225 Düsseldorf, Germany.
E-mail: andrea.braeutigam@uni-duesseldorf.de

Editor
A. Weber

Received: 18 February 2010; Accepted: 30 April 2010

doi:10.1111/j.1438-8677.2010.00373.x

ABSTRACT

Next generation sequencing (NGS) technologies have opened fascinating opportunities for the analysis of plants with and without a sequenced genome on a genomic scale. During the last few years, NGS methods have become widely available and cost effective. They can be applied to a wide variety of biological questions, from the sequencing of complete eukaryotic genomes and transcriptomes, to the genome-scale analysis of DNA–protein interactions. In this review, we focus on the use of NGS for plant transcriptomics, including gene discovery, transcript quantification and marker discovery for non-model plants, as well as transcript annotation and quantification, small RNA discovery and antisense transcription analysis for model plants. We discuss the experimental design for analysis of plants with and without a sequenced genome, including considerations on sampling, RNA preparation, sequencing platforms and bioinformatics tools for data analysis. NGS technologies offer exciting new opportunities for the plant sciences, especially for work on plants without a sequenced genome, since large sequence resources can be generated at moderate cost.

INTRODUCTION

Abstractly speaking, next generation sequencing (NGS) technologies enable the quick, inexpensive and comprehensive analysis of complex nucleic acid populations (Metzker 2010). In other words, they produce DNA sequence reads, and a lot of them. The production, assembly and analysis of these sequence reads requires different experimental approaches from sequencing library generation to new bioinformatics tools for post-sequencing procedures.

During Sanger sequencing, the sequence of bases is read from DNA fragments of different length, which are generated by a DNA polymerase that breaks off whenever it encounters a terminator nucleotide. Since the terminator is labelled, the sequence can be read based on the different lengths of the fragments. The large-scale sequencing of complex nucleic acid populations with Sanger sequencing requires subcloning of the nucleic acids into vectors and their amplification in hosts. The initial sequencing of the yeast, *Arabidopsis* and human genomes, for example, and a host of EST sequencing projects (*e.g.* http://compbio.dfci.harvard.edu/tgi/tgipage.html) were accomplished by consortia of many laboratories and were very labour intensive. Even with optimised protocols, a megabase (Mb) of sequence costs about US$1330 (Wall *et al.* 2009). In contrast, NGS technologies produce huge amounts of DNA sequences at a much lower cost per sequence, from US$4–90 per megabase, depending on the technology (Wall *et al.* 2009). All NGS technologies avoid the subcloning step and directly sequence the DNA. The differences in input (the DNA) and output (the reads) are discussed in detail below.

To date, NGS has been successfully used to study genomes and transcriptomes of species with and without sequenced genomes. Many small prokaryotic genomes have been completely sequenced *de novo* by shotgun sequencing (*e.g.* Aury *et al.* 2008), and the first complex genome sequenced entirely with NGS is that of the panda bear. In this case, different sized paired end libraries were sequenced and assembled into the 20 pairs of autosomes and one sex chromosome pair, with 73-fold coverage (Li *et al.* 2010). Until April 2010, no plant genome sequence generated exclusively with NGS methods had been published, although genome snapshots had been generated for some species. For waterhemp, a weed of significance in North America, a genome snapshot included the near complete chloroplast genome (Lee *et al.* 2009), and the plastid genomes of two basal eudicot plants have also been generated with NGS technology (Moore *et al.* 2006). For barley, a genome snapshot yielded about 1% of the haploid genome equivalent and showed that the barley genome contains about 60% of transposable elements and 9% of novel repetitive sequences (Wicker *et al.* 2009). The technology is also suitable for analysing, on a genomic scale, epigenetic modifications such as DNA methylation and histone modification or DNA–protein interactions (Cokus *et al.* 2008; Jothi *et al.* 2008; Lister *et al.* 2008). NGS can also be used for mapping of mutants. The mutant pool of F2 plants of a mapping cross will be enriched in SNPs of the parental

mutant line in the vicinity of the mutation and also reveal the mutation itself (Schneeberger *et al.* 2009; Laitinen *et al.* 2010). Several recent reviews have focused on NGS technologies in the context of genome (re-)sequencing, epigenetics and the analysis of model plants (*e.g.* Mardis 2008; Shendure & Ji 2008; Lister *et al.* 2009).

Transcriptome sequencing yields information about the genes of an organism at lower cost compared to genome sequencing since only transcribed regions are investigated. In this review, we focus mainly on the use of NGS for plant transcriptomics. We discuss the output of different sequencing platforms and describe how NGS is used to analyse transcriptomes qualitatively as well as quantitatively. We distinguish between the analyses of species with a complete, preferably well annotated, genome sequence available and species without a sequenced genome, since the experimental design as well as the appropriate NGS method differs between these scenarios.

NGS TECHNIQUES

The different NGS technologies vary in their input requirements and their sequence output in terms of total bases sequenced and length of each sequence read, as well as the price per megabase sequence information. Depending on the application, one particular type of NGS may be more suitable than another. Currently, there are three NGS platforms commercially available, the Genome Sequencer FLX from 454 Life Sciences/Roche, the Illumina Genome Analyser from Solexa and Applied Biosystems' SOLiD (acronym for: 'Sequencing by Oligo Ligation and Detection'). Furthermore, there are two methods that depend on single molecule sequencing in advanced development but not yet widely available, the true Single Molecule Sequencing (tSMS) of Helicos Bioscience and the Single Molecule Real Time sequencing (SMRT) of Pacific Biosciences. The technical principles of these sequencers and the chemistries used have recently been reviewed in great detail (Holt & Jones 2008; Mardis 2008; Metzker 2010). In this review, we will focus on practical aspects and briefly describe what material is needed, as well as the quantity and condition of the sequence data produced by the different NGS platforms.

The Genome Sequencer FLX from 454 Life Sciences is capable of producing over a million reads of about 250 or 400 (Titanium chemistry) bases per run, leading to a total yield of 250 or 400 megabases, respectively, at a price of approximately US$90 per megabase (Wall *et al.* 2009). We refer to this technology as a *long read technology*. It is based on emulsion PCR and pyrosequencing (Ronaghi *et al.* 1998). As template, one can use fragmented (300–800 bp) double-stranded DNA, such as genomic DNA or cDNA. For paired end sequencing, libraries from fragments of 3–20 kb can be used. The 454 sequencing produces the longest reads but the total sequence output per run is low compared to the other platforms (Table 1). The sequence accuracy is 99%, with most of the errors occurring in homopolymer stretches (http://www.454.com/index.asp; Metzker 2010).

The Illumina Genome Analyser produces over 100 million short reads (35–76 bases, depending on the sequencing chemistry) leading to 3–6 gigabases of sequencing data in one run. A megabase costs about US$4. We refer to this technology as a *short read technology*. It is based on solid-phase bridge PCR and uses a 'sequencing by synthesis' approach, with fluorescent dye-labelled reversible terminator nucleotides. It uses fragmented double-stranded DNA as template. Fragments of up to 10 kb can be used for the construction of paired end sequencing libraries. The technology is also referred to as Solexa sequencing. The accuracy of the produced sequence data is greater 98.5% (http://www.solexa.com/; Metzker 2010).

The Applied Biosystems SOLiD system is based on emulsion PCR in combination with sequencing by ligation with dye-labelled oligonucleotides (Shendure *et al.* 2005). It produces up to one billion short reads (up to 50 bases) per run, leading to a total sequence output of up to 30 gigabases per single read run. As templates it uses fragmented double-stranded DNA. Fragment sizes for the construction of paired end sequencing libraries can be up to 10 kb. The sequences produced are 99.94% accurate (http://www3.appliedbiosystems.com/AB_Home/; Metzker 2010). We refer to the SOLiD technology as *short read technology*.

These well-established NGS methods depend on amplification of the target molecules by either emulsion or solid phase PCR, and perform the actual sequencing reaction on an amplified clonal template to enhance the detectable fluorescence signals. In contrast, tSMS of Helicos Bioscience and SMRT of Pacific Biosciences use single DNA molecules for the sequence reactions. The tSMS Helicos system uses a 'sequenc-

Table 1. Comparison of available NGS technologies.

platform	template preparation	sequencing chemistry	read length (bp)	total output per run (Gbp)[a]	reference; company homepage
Genome Sequencer FLX from 454 Life Sciences	Emulsion PCR	Pyrosequencing	400	0.4	(Metzker 2010) http://www.454.com/index.asp
Illumina Genome Analyzer	Solid Phase PCR	Sequencing by synthesis	76	6	(Metzker 2010) http://www.solexa.com/
Applied Biosystems SOLiD	Emulsion PCR	Sequencing by ligation	50	30	(Metzker 2010) http://www3.appliedbiosystems.com/AB_Home/
tSMS by Helicos Bioscience	Single molecule	Sequencing by synthesis	32	21	(Metzker 2010) http://www.helicosbio.com/
SMRT by Pacific Biosciences	Single molecule	Real time	>900	?	(Metzker 2010) http://www.pacificbiosciences.com/

[a]Total output for single read runs.

ing by synthesis' approach with fluorescent dye-labelled virtual terminator nucleotides. It can generate up to 800 million short reads (25 bp) and up to 21 gigabases per run, and the template size may vary between 25 and 5000 bases (Table 1). Paired end sequencing is also possible (Pushkarev *et al.* 2009). In contrast, the SMRT system of Pacific Biosciences works with sequencing by synthesis in real time, without reversible terminators. This fact, and the use of nucleotides with the fluorescent dye coupled to the phosphate group, allows the generation of very long reads of about 1000 bp (Table 1). Currently, one instrument is capable of producing a raw read throughput that is equivalent to one-fold coverage of a diploid human genome per day (Eid *et al.* 2009). The single molecule sequencing, especially the SMRT method, needs significantly less chemicals than current commercial NGS methods. It is expected that the costs for NGS will significantly decrease with broader commercial availability of these methods.

TRANSCRIPTOME ANALYSIS – GENERAL CONSIDERATIONS

Tissue sampling, RNA preparation and treatment, sequencing and analysis strategies differ for the individual applications of transcriptome analysis from plants with or without a sequenced genome. The following consideration should be included to maximise results of the sequence analysis (summarised for plants without a sequenced genome in Fig. 1).

Sampling

Sampling is an important consideration since different genetic backgrounds are available in plants. They can come from inbred or outbred species, and be either well defined or collected from the wild. This is especially critical if species are sequenced for which no reference genome or transcriptome is available. For traditional Sanger EST sequencing, multiple different accessions were frequently used by different researchers, resulting in many extremely similar unigenes (NCBI nomenclature; http://www.ncbi.nlm.nih.gov) or tentative consensus sequences (TIGR/DFCI nomenclature; http://compbio.dfci.harvard.edu/tgi/definitions.html), which may actually represent the same gene. Since traditional sequencing approaches yielded fairly long ESTs to be assembled into contigs, the different alleles have not significantly hampered the construction of EST contig databases (Wall *et al.* 2009). However, producing a transcriptome of an outbreeding species with multiple organisms sampled will inevitably lead to allelic variation being captured in the sequence tags (Novaes *et al.* 2008). From an assembly standpoint, a single plant, preferably inbred for several generations, should be used to capture RNA from a variety of tissues. In practice, high sequencing cost, outbreeding species or collection in the wild will hamper this ideal approach and require compromises between creating a transcriptome database while at the same time answering specific questions (*e.g.* Novaes *et al.* 2008; Alagna *et al.* 2009; Dassanayake *et al.* 2009).

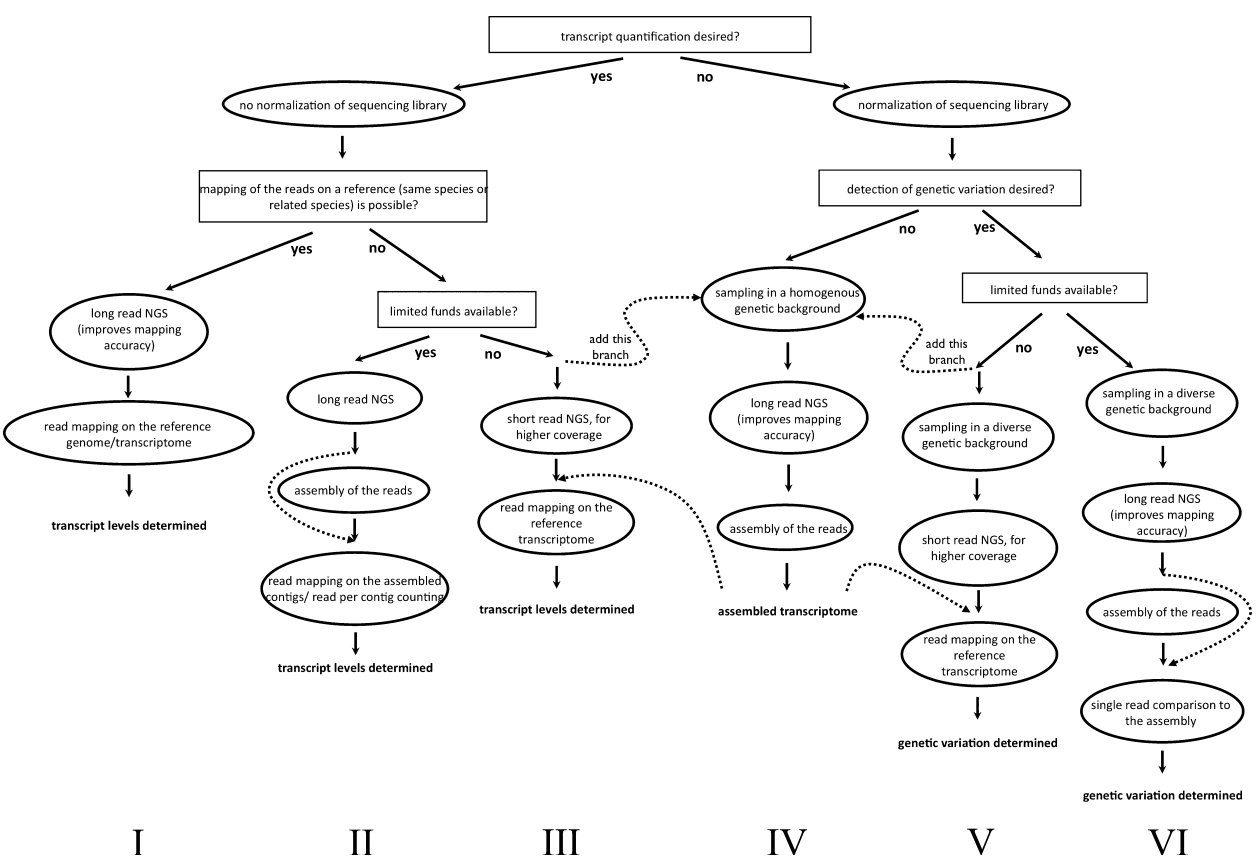

Fig. 1. Decision flowchart for the transcriptome analysis of plants without a sequenced genome.

RNA preparation and treatment

After sampling the species RNA, preparation will require special attention. The next generation sequencing technologies critically depend on pure samples, without any degradation. High-quality RNA has been successfully isolated with commercially available kits and also from the single step guanidium isothiocyanate- or phenol-based methods (Bräutigam *et al.* 2008; Novaes *et al.* 2008; Alagna *et al.* 2009; Barakat *et al.* 2009; Dassanayake *et al.* 2009; Wang *et al.* 2009). It is critical to check the quality of the RNA, for example with the Agilent 2100 Bioanalyser RNA chip (Agilent Technologies, Santa Clara, CA, USA). mRNA purification reduces the amount of rRNA present in the sample. If the mRNA is enriched, RNA species without a poly (A) tail are largely lost, including ribosomal RNAs and small RNAs. If the small RNAs are to be sequenced alongside the mRNAs, a depletion of rRNA is probably more desirable (*e.g.* Lister *et al.* 2008). Two different ways to generate cDNA can be used, oligo d(T) priming and random priming. Random priming will increase the likelihood to sample sequence reads across the complete sequence and to include small RNAs, but increases the proportion of sequences representing rRNA even if mRNA was purified. Random priming without prior mRNA purification leads to a large portion of sequences representing rRNA (Wall *et al.* 2009). Since only a limited amount of random primers can be used to create long cDNA fragments (Sambrook & Russell 2001), a bias may be introduced by the depletion of primers matching highly abundant transcripts. Oligo d(T) priming reduces rRNA contributions but may introduce a bias by the loss of the 3′ portion of long transcripts (Wall *et al.* 2009). The quality of oligo d(T) primed cDNA can be estimated based on comparison to the mRNA on the Agilent 2100 Bioanalyser (Agilent Technologies).

After RNA preparation, it needs to be decided whether to normalise the cDNA library or whether to keep it in its original format. Normalisation will increase the number of genes represented by at least one sequence, as well as the coverage of sequencing reads, for low to medium abundance genes (Franssen S.U., Shrestha R.P., Bräutigam A., Bornberg-Bauer E., Weber A.P.M., unpublished results). In contrast, it will reduce the abundance of reads, especially for highly abundant transcripts such as RubisCO, and therefore the sequence coverage for these genes (Wall *et al.* 2009). With normalisation, the quantitative information will largely be lost. Since normalisation is rarely complete, however, it may still be possible to determine highly expressed genes compared to low-expressed genes (Franssen S.U., Shrestha R.P., Bräutigam A., Bornberg-Bauer E., Weber A.P.M., unpublished results).

TRANSCRIPTOME SEQUENCING OF SPECIES WITHOUT A SEQUENCED GENOME

Before the advent of NGS, the transcriptome analysis of species without a genome or transcriptome reference database was very costly in terms of time, labour and money. NGS technologies allow the qualitative and quantitative analysis of entire transcriptomes at reasonable cost.

Creating a transcriptome sequence database

Many NGS transcriptome projects aim to lay a foundation for future experiments and create a sequence resource (Bräutigam *et al.* 2008; Novaes *et al.* 2008; Alagna *et al.* 2009; Barakat *et al.* 2009; Dassanayake *et al.* 2009; Wang *et al.* 2009). To create this transcriptome database, many researchers have opted to use 454 NGS since it produces the longest reads of the currently available NGS technologies (Table 2) (Bräutigam *et al.* 2008; Novaes *et al.* 2008; Alagna *et al.* 2009; Barakat *et al.* 2009; Dassanayake *et al.* 2009; Wang *et al.* 2009). A short read-based technology such as Solexa has been used for re-sequencing in *Brassica napus* (Trick *et al.* 2009) but not for *de novo* sequencing. During the assembly of contigs, single reads are assessed for their overlapping sequence and the identity within the sequence overlap. The assembly becomes increasingly more difficult when the read length gets shorter and shorter (summarised in Pop & Salzberg 2008), which is the most compelling reason for choosing a long read technology. Indeed, despite using a long read technology, many of the EST contigs of *de novo* sequencing projects remain short (Bräutigam *et al.* 2008; Novaes *et al.* 2008; Alagna *et al.* 2009; Dassanayake *et al.* 2009; Wang *et al.* 2009). A number of assembly programs based on de Bruijn graphs have been spe-

Table 2. NGS transcriptome projects in non-model species.

species	technology	purpose	citation
Rhizophora mangle *Heritiera littoralis*	454/Roche GS FLX	Transcriptome database, pathway representation	Dassanayake *et al.* 2009
Castanea dentata *Castanea mollissima*	454/Roche GS FLX	Transcriptome database, pathway representation, comparative analysis	Barakat *et al.* 2009
Brassica napus	Illumina	SNP discovery	Trick *et al.* 2009
Eschscholzia californica	454/Roche GS 20	sRNA characterization	Barakat *et al.* 2007
Artemisia annua	454/Roche GS FLX	Transcriptome database, pathway representation	Wang *et al.* 2009
Olea europaea	454/Roche GS FLX	Transcriptome database, pathway representation, comparative analysis	Alagna *et al.* 2009
Arachis hypogaea	Illumina	sRNA characterization	Zhao *et al.* 2010
Eschscholzia californica *Persea americana*	454/Roche GS 20	Transcriptome database	Wall *et al.* 2009
Eucalyptus grandis	454/Roche GS 20 and GS FLX	Transcriptome database, SNP discovery	Novaes *et al.* 2008
Pisum sativum	454/Roche GS 20 and GS FLX	Transcriptome database, proteomics as a follow up application	Bräutigam *et al.* 2008

cifically designed to handle short reads from NGS technology: for example, 454/Roches Newbler, SHARCGS, VCAKE, VELVET, EULER-SR, EDENA, ABySS and ALLPATHS (summarised in Flicek & Birney 2009). These new algorithms are designed for genome assemblies, however, and have not yet been used for a transcriptome assembly, except for the Newbler assembler supplied by Roche/454. Programs like MIRA (Chevreux *et al.* 2004) phrap (Gordon *et al.* 1998) and CAP3 (Huang & Madan 1999) use a more traditional overlap-based approach to assembly. The transcriptome of mangroves was assembled with phrap and with the Newbler assembler provided with the 454 sequencer (Dassanayake *et al.* 2009), and those of *Arabidopsis* and pea with Newbler followed by CAP3 (Weber *et al.* 2007; Bräutigam *et al.* 2008). The assembly parameters between programs vary. For example, the mangrove transcriptomes were assembled using an overlap window of 40 bases with >90% identity (Dassanayake *et al.* 2009), whereas a program like Mira recovers different transcripts separately instead of assembling a consensus transcript from different alleles (http://www.chevreux.org/mira_ex_est.html). Relaxed parameters will allow assembly of sequences representing different alleles of a gene, which are likely captured when sampling from multiple organisms in the wild, at the cost of possibly merging recently duplicated genes (Dassanayake *et al.* 2009), while a conservative assembly will retain many unigenes that likely represent different alleles at the cost of shorter contigs (Franssen S.U., Shrestha R.P., Bräutigam A., Bornberg-Bauer E., Weber A.P.M., unpublished results). Many transcriptome databases generated by NGS retain small contig sizes despite a high number of bases sequenced (Bräutigam *et al.* 2008; Novaes *et al.* 2008; Alagna *et al.* 2009; Barakat *et al.* 2009; Dassanayake *et al.* 2009; Wang *et al.* 2009). In one of the earliest plant transcriptome re-sequencing efforts, the *de novo* sequence assembly returned less contigs than could have been generated based on the mapping of sequences onto the reference genome (Weber *et al.* 2007). This indicates that the assembly retains a large optimisation potential. Depending on the follow-up experiments, a database of less quality, compared to EST databases generated by traditional sequencing methods, may be sufficient for subsequent applications such as proteomics (Bräutigam *et al.* 2008). The decision flow chart for creating a transcriptome database solely as a sequence resource is depicted in Fig. 1 column IV. Following assembly of the contigs, the transcriptome databases are annotated with different databases, such as the proteins from the current TAIR release, from NCBIs nr or from SWISS-PROT (Novaes *et al.* 2008; Alagna *et al.* 2009; Barakat *et al.* 2009; Dassanayake *et al.* 2009; Wang *et al.* 2009). In the different *de novo* EST sequencing projects, between 20% and 40% of sequences were recovered that could not be mapped onto a plant sequence with BLAST. This was not due to bacterial or viral contamination, as the EST contigs could not be mapped onto NCBIs nr database. These sequences may represent 'garbage sequences' generated by errors within the amplification and/or sequencing technology, or may be genes that have diverged to the point that they are no longer recognised by BLAST. The transcriptome databases are rarely created on their own, but rather to address a specific biological question, such marker discovery, comparative transcriptomics or pathway distribution within plants, as described below.

Marker discovery

Identifying markers serves to assess the variation in a wild population or create breeding resources for crops without sequenced genomes or transcriptomes. Unlike SNP discovery in species with sequenced genomes, such as the 1001 genomes project in *Arabidopsis* (Weigel & Mott 2009), for species without a sequenced genome different considerations are important. In principle, there are two possible approaches to discovering markers: based on genomic sequence and based on transcriptome sequences. In species with a sequenced genome, reads of any length can be mapped onto the reference genome, and several algorithms have been developed for SNPs (http://seqanswers.com/wiki/Special:BrowseData/Bioinformatics%20application?Biological_domain=SNP_discovery).

A transcriptome sequencing approach may be more suitable if it is desirable to create a comprehensive sequence resource at the same time and with a moderate investment (Novaes *et al.* 2008) (Fig. 1, line VI), or if third base bias is to be used as SNP candidate criterion. If there are sequence resources available in the public databases, sequencing with a short read system such as Illumina, SOLiD or HeliScope yields more sequences in terms of number and coverage compared to 454/Roche technology. For *Brassica napus,* for example, about 40,000 SNPs were discovered in a Solexa-based transcript sequencing of two cultivars. Reads were analysed by mapping them onto a reference database of unigenes retrieved from public archives (Trick *et al.* 2009) (Fig. 1, line V with the reference database already available). For *Eucalyptus grandis*, SNPs were identified in a combined approach of database generation and SNP discovery with a long read technology (Fig. 1, line VI). Initially, the sequences were assembled with Newbler and Paracel transcript assembler into consensus sequences. SNPs were identified with GS Reference Mapper (454 Life Science, Branford, CT, USA) (Novaes *et al.* 2008). The transcriptome database projects from mangroves, in which samples were gathered from multiple individuals in a diverse population, expressly report a high number of unigenes mapping to the same *Arabidopsis* reference gene (Dassanayake *et al.* 2009). Although untested, the high frequency of unigenes for one reference may represent allelic variation.

For a marker discovery project in a species without any sequence resources at the time of the experiment, a dual strategy may be most advantageous (Fig. 1, line V). Initially, a transcriptome reference database should be created from a single, preferably inbred, individual with a long read technology. Under these conditions, the assembly will profit from long read length and it will not be hampered by allelic differences. Once this transcriptome is available, SNP discovery can be carried out cost effectively with short read sequencing technologies followed by read mapping onto the reference.

Marker discovery is also possible based on genomic DNA. In species with a sequenced genome, short read technology will yield reads that can be mapped onto the reference (Weigel & Mott 2009). If limited genome information is available, SNP markers can be generated from PCR amplicons of different genotypes that are mixed and sequenced by 454/Roche NGS (Bundock *et al.* 2009). If little to no sequence information is available, microsatellite markers can be derived from a genome snapshot (Tangphatsornruang *et al.* 2009). If genomic DNA is used for sequencing, large contributions from

organellar DNA to the total (*e.g.* Lee *et al.* 2009) can be avoided by purifying the organellar DNA away from the sample (Willmitzer & Wagner 1981; Steinmüller & Apel 1986), whereas one should consider that the isolation of nuclei might be difficult or even impossible for certain species or tissues. A recent dedicated review provides details on the possibilities and challenges of NGS for polymorphism detection and includes a wealth of examples (Imelfort *et al.* 2009).

Comparative transcriptomics in species without a sequenced genome

Transcriptome comparisons using microarray analyses in model plants have become a successful tool to gain more comprehensive understanding of organs, developmental stages (Schmid *et al.* 2005), responses to external stimuli (Kilian *et al.* 2007; Goda *et al.* 2008) and a multitude of other processes that involve changes in transcript expression. With NGS technology it is possible to analyse the transcriptional profile of non-model plants on a genomic scale (Alagna *et al.* 2009).

The expression profile of thousands of genes was assessed in two different cultivars at two different stages of olive ripening. In this case, sequencing cover was sacrificed for quantitative information, as the libraries were not normalised prior to sequencing (Alagna *et al.* 2009) (Fig. 1, line II). To analyse transcript abundance, the number of sequence reads clustering into one tentative consensus with >90% sequence identity over a 100-bp window was used as the measure. This method will discriminate against low-expressed genes for which contigs of the same gene will remain disjointed because not enough sequence can be recovered (Wall *et al.* 2009). It can also be shown that a non-normalised library discriminates against pathways of lower expression compared to a normalised library from the same tissue (Franssen S.U., Shrestha R.P., Bräutigam A., Bornberg-Bauer E., Weber A.P.M., unpublished results), which limits the potential to analyse transcripts of genes expressed at a low level, such as genes involved in signalling and regulation or channel genes. If the expression level is based on transcriptome assembly, the analyses are limited to transcripts with sufficient coverage for assembly and large expression differences (Alagna *et al.* 2009).

For a more detailed analysis, two different strategies are possible. If the non-model species is closely related to a species with a sequenced genome, the sequence reads can be mapped onto the reference (Palmieri & Schlotterer 2009) (Fig. 1, line I). The choice of mapping software will influence the results, with the commercial program CLC NGS cell being least affected by differences in reads compared to the reference (Palmieri & Schlotterer 2009). The free software BLAT (Kent 2002) is also barely affected by alterations if the sequence reads are at least 100-bp long (Palmieri & Schlotterer 2009). The second strategy is a dual strategy similar to that proposed for marker discovery (Fig. 1, line III). Initially, normalised cDNA libraries from all conditions to be analysed later and non-normalised libraries from these conditions should be assembled into a reference transcriptome. Using a long read sequencing technology for at least the normalised cDNA library will facilitate the assembly (Pop & Salzberg 2008). This analysis will profit from using a well-defined genetic background, either an inbred line or vegetatively propagated individuals with identical genomes, as this will ease the assembly. The non-normalised libraries could be analysed either with long read technology, in which case they aid in providing a good coverage reference database, or with short read technology, with which more reads can be produced at the same cost compared to long read technology. For aligning the reads to the reference, a range of programs is available (Flicek & Birney 2009; Pepke *et al.* 2009) (http://seqanswers.com/wiki/Special:BrowseData/Bioinformatics_application?Bioinformatics_method=Alignments).

Comparative transcriptome analysis with NGS not only enables intra-species comparison but also the comparison of two different non-model species with each other. This approach has already been explored using microarray technology. However, the unpredictable, imprecise matches to the probes of the array hamper analyses. A comparative analysis of non-models is especially desirable when traits that are not present in a model or in a model and a close relative are to be analysed. Theoretically, one could assume that there are also two possibilities to compare the transcriptomes: (i) creating transcriptome databases and mapping the reads of each species to the databases, and (ii) mapping the reads to a common reference genome. The first method is fraught with several problems. If only one reference transcriptome is generated, reads from one species will map perfectly while the reads from the other species will map imperfectly and thus bias will be injected, especially for fast-evolving genes. If two reference transcriptomes are generated, this issue is resolved. However, since the genomes and the history of the genomes remain unknown, it will be impossible to analyse read mappings to closely related genes when two references are used. For example, if gene A undergoes duplication in the other species to gene A1 and A2, A will appear to be expressed twice as high in a comparative approach, since either A1 or A2, but not both, will be assigned as the closest homologue and thus the comparison partner to A. Mapping the reads to a closely related reference genome resolves these problems. Several software tools, *e.g.* BLAT and CLC, are available to map imperfectly matching reads (Palmieri & Schlotterer 2009).

ANALYSING THE TRANSCRIPTOME OF SPECIES WITH A SEQUENCED GENOME

When working with model species, a genome and/or transcriptome database is available, and genome scale analyses such as microarrays and tiling arrays are enabled. However, these are inherently biased by the chip design at the current state of the art. Frequently, genes are missing from the arrays, alternative and antisense transcripts are not well represented and RNA species other than mRNA are frequently not present. Since NGS remains more expensive than array-based analyses, it needs to be carefully weighed whether the advantages of NGS justify the higher costs.

Sequencing the transcriptome

In 2008, several groups demonstrated that the Genome Sequencer FLX from 454 Life Sciences/Roche can be used for transcriptome profiling (Shin *et al.* 2008; Torres *et al.* 2008). Torres and co-workers point out the influence of the method

used for sequence library production on the final library. They suggest fragmentation of the cDNA by nebulisation to avoid under-representation of long (>300 bp) or short transcripts (<80 bp) in the sequencing libraries and the final 454 sequence reads (Torres *et al.* 2008). A transcriptome analysis of species with a fully sequenced genome identifies novel transcripts, checks and optimises transcript predictions and identifies splicing isoforms (Cheung *et al.* 2006; Emrich *et al.* 2007; Weber *et al.* 2007). 454 pyrosequencing of cDNA from maize shoot apical meristem cells sequenced over 260,000 reads that map to over 25,000 maize genomic sequences. 30% of the reads could not be aligned to the 640,000 maize ESTs known at the time indicating a large number of new transcripts were discovered (Emrich *et al.* 2007). More than 17,400 expressed genes could be identified through sequencing the cDNA from aboveground organs of *Arabidopsis* seedlings. This is equivalent to more than 90% of the transcripts expressed in these tissues. The expressed genes included previously unannotated transcripts as well as genes with no prior EST support (Weber *et al.* 2007).

A much higher sequencing depth at comparable cost can be achieved using short read technology such as the Illumina Genome Analyser or the Applied Biosystems SOLiD system, producing over 100 million sequencing reads. These reads are directly mapped to the genome sequence (Cloonan *et al.* 2008; Lister *et al.* 2008; Mortazavi *et al.* 2008). Splice isoforms can be identified by reads reaching over predicted exon boundaries (Mortazavi *et al.* 2008). Novel genes and incorrectly annotated 5′ or 3′ untranslated regions are discovered if reads map to genomic regions for which no elements were annotated. The abundance of a transcript can be measured simply by counting how many reads map onto a given gene. In contrast to microarray experiments, which report a ratio of fluorescence in arbitrary units, NGS measurements are absolute. To compare the abundance of transcripts within a sequence library these read counts are often normalised to the transcript length, *e.g.* reads per kilobase (RPK) of transcript (Wilhelm & Landry 2009). To compare the abundance of transcripts in different libraries representing, for example, different tissues, cell types or states of an organism, the read counts are additionally normalised to one million reads. According to this method, the abundance of a certain transcript in a certain cDNA population/sequence library, obtained by NGS, is generally given as reads per kilobase and million (RPKM), meaning reads counted per 1000 bp of this transcript and per one million total reads from the sequence library (Wilhelm & Landry 2009). This way, not only relative but absolute abundance values are determined for a given condition.

In complex eukaryotic genomes there is widespread antisense transcription (Katayama *et al.* 2005); therefore it is worthwhile to maintain strand-specific information for the RNA molecules used for preparation of the sequencing libraries. Some studies describe methods to achieve this and demonstrate the feasibility of these methods (Cloonan *et al.* 2008; Lister *et al.* 2008). One possibility to create such a sequencing library is to fragment the RNA by metal hydrolysis, dephosphorylate the 5′ ends, ligate a specific single stranded adaptor to the 3′ end using T4 RNA ligase, phosphorylate the 5′ end and ligate a specific 5′adaptor. These RNAs can then be used for first and second strand cDNA synthesis, leading to cDNAs providing information about the original 5′ to 3′ orientation of the RNA template by means of the specific 3′and 5′adaptor sequences (Lister *et al.* 2008).

Mapping the huge amounts of short read sequences produced by NGS to a given reference sequence is challenging. A traditional and well established sequence alignment tool like BLAST (Altschul *et al.* 1997) can be used for mapping these short reads, but BLAST is not optimised to cope with high numbers of reads; therefore such mappings are very time consuming. BLAT was developed to perform alignment tasks much faster (Kent 2002). To map about 10 million reads of 32 bp on 5 megabases of human genomic sequence, BLAT needed 6 h whereas BLAST needed 42 h (Li *et al.* 2008b). BLAT is suitable to map reads from the 454 platform, but short read sequencing technologies can produce over ten times more data within a single run, thus new bioinformatics tools capable of dealing with such huge amounts of data have been developed (Flicek & Birney 2009). Such programs are able to map the 10 million reads onto the 5 Mb of human sequence within <10 min (Li *et al.* 2008b); examples of such specialised programs are BOWTIE (Langmead *et al.* 2009), SOAP (Li *et al.* 2009), MAQ (Li *et al.* 2008a), ELAND (http://www.illumina.com/), SSAHA2 (Ning *et al.* 2001), ZOOM! (Lin *et al.* 2008) and SHORE (Ossowski *et al.* 2008). Currently, there is much progress in the development of such software, leading to the publication of several new programs within the last 2 years. Since experience with these programs and also comparative investigations is limited at the moment, it is difficult to predict if and which of these tools will become accepted as the standard. Perhaps it will turn out that, depending on the amount of reads, read length and genome complexity of the organism investigated, a different program is favoured (Palmieri & Schlotterer 2009).

The high throughput, short read NGS systems have been successfully used in several studies for quantitative and qualitative transcriptome analysis in animal, plant and microbial model systems (Cloonan *et al.* 2008; Mortazavi *et al.* 2008; Nagalakshmi *et al.* 2008; Sultan *et al.* 2008; Wilhelm *et al.* 2008). An example of a particularly comprehensive study comes from Lister *et al.* (2008). By combining different techniques, they assessed the strand-specific transcriptome, small RNAs and cytosine methylation in *Arabidopsis* on the genome scale, using short read sequencing with the Illumina Genome Analyser. The comparison of wild-type plants, DNA methyltransferase and DNA demethylase mutants allowed analysis of the interactions between DNA methylation, small RNA function and effects on transcriptional regulation within the experiment.

Combining NGS and SAGE

In cases where one is solely interested in quantitative data, thus in measuring transcript levels, it is possible to combine NGS with serial analysis of gene expression (SAGE). SAGE is characterised by the fact that each transcript within an RNA population is represented by a certain tag, a DNA fragment of typically 20–26 bp. In former times, these tags were ligated to longer fragments and sequenced using Sanger sequencing (Velculescu *et al.* 1995). With the availability of short read NGS sequencers like the Illumina Genome Analyser and Applied Biosystems SOLiD system, these tags are an ideal

template for direct sequencing (Meyers *et al.* 2004). To generate the tags, the mRNA is converted to double stranded cDNA, which is bound to a matrix by the polyA tails. The cDNA is restricted using an enzyme with a four-base recognition site like NlaIII or DpnII. After removal of the 5′moiety of the cDNAs, an adaptor containing the recognition motif of a type II restriction endonuclease like MmeI or EcoP15I is ligated. These enzymes cut 21, in the case of MmeI, or 26 nucleotides, in the case of EcoP15I, downstream of the recognition site (Matsumura *et al.* 2008). Following the restriction with such an enzyme, the DNA fragments are recovered and, after addition of a 3′adapter, they can be directly used for short read NGS. The abundance of a given tag, *i.e.* how often this tag was sequenced, within the collection of tags from a certain mRNA population determines the expression level of the corresponding gene (Matsumura *et al.* 2003; Meyers *et al.* 2004; Molina *et al.* 2008).

To assign the short sequence tags to mRNAs and genes, the complete annotated genome sequence, or at least the complete transcriptome sequence, of the species must be known. Even the short 21-nt tags generated by an MmeI digest from cDNAs match mostly once to complex eukaryotic genomes (Simon *et al.* 2009), allowing the unequivocal relation of tags and genes.

Since for each gene, only the short tag instead of the whole mRNA is sequenced, SAGE leads to a much deeper sampling and broader coverage of the sequenced transcriptome than simple RNA-seq, for the same sequencing effort. While it is possible to detect currently unknown or unpredicted transcripts if the tags are mapped to the whole genome sequence, of course, one loses information about differential mRNA processing and splice variants included in the datasets from simple RNA-seq. If deep coverage transcript profiling is the focus, SAGE is a good and cost-effective alternative to simple RNA-seq.

Assessing small RNAs at the genome level

The size of small RNAs (sRNA) makes them the ideal target for NGS, and the introduction of NGS technologies has been followed by significant advances in sRNA discovery and analysis. Two predominant forms of plant sRNAs have been observed. The 21-nt microRNAs (miRNAs) mainly act post-transcriptionally by direct cleavage of a specific target mRNA (Bartel 2004; Jones-Rhoades *et al.* 2006). The 24-nt short interfering RNAs (siRNAs) typically direct *de novo* DNA methylation and regulate gene expression at the transcriptional level (Vaucheret 2006).

To use sRNAs for NGS, sRNA molecules are isolated, usually by purification of low-molecular weight RNA from total RNA, followed by size selection of RNA in the range of 20–30 nt *via* a polyacrylamide gel. After addition of 5′ and 3′ adapters, reverse transcription and, if necessary, linear amplification with a few PCR cycles to obtain the template amount necessary for NGS, the sRNAs are sequenced. While for the first large-scale sRNA studies 454 pyrosequencing was used (Henderson *et al.* 2006; Lu *et al.* 2006), analysis of small RNAs is currently performed mainly with the Illumina Genome Analyser (Lister *et al.* 2008). But also the Applied Biosystems SOLiD system should be ideally suited for studying sRNAs. Both systems deliver reads that are long enough to cover the complete sRNA sequences. Since more than 100 million reads are produced, the analyses are very deep and also detect very low abundance sRNAs.

The number of plant species with sRNAs sequenced by the NGS method is ever increasing (Rajagopalan *et al.* 2006; Barakat *et al.* 2007; Yao *et al.* 2007; Dolgosheina *et al.* 2008; Nobuta *et al.* 2008; Sunkar *et al.* 2008; Zhu *et al.* 2008). One interesting outcome of this collection of sequence data is that sRNAs are not strictly conserved between all plant species. Even the distribution of sRNAs amongst various size classes has been found to differ between plants. It can be assumed that differences in maintenance of genomic organisation between plant species that have genomes of drastically different sizes leading to this differential distribution of sRNA lengths (Morin *et al.* 2008).

CONCLUSION

NGS technologies open up new ways to pursue research. When microarray technology appeared, suddenly the expression not only of one gene or a small group of genes but of all (or nearly all) genes could be tested at the same time. Microarrays were used successfully both to generate and to test hypotheses, and to generate community resources of expression profiles for model species. In model plants, NGS added the capacity to analyse gene expression in an unbiased way, to detect more expressed sequences, to define genes with their alternative splice form and to analyse DNA methylation and histone modifications on a genome scale. Creative ways to use the new technology continue to be published. For plants without a sequenced genome, NGS produces completely new ‘playgrounds’. It is now possible to build a custom resource for your plant or even plants and projects of interest, basically making – at moderate cost – any plant a ‘model plant’ with sequence resources. The decrease in costs, with five competing technology platforms available, and the increase of computational tools for NGS will only extend the possibilities for non-models, which already include gene and marker discovery and genome-wide quantification of gene expression.

ACKNOWLEDGEMENTS

We thank Dr Sigrun Wegner-Feldbrügge and Dr Marc Linka for helpful comments on the manuscript.

REFERENCES

Alagna F., D’Agostino N., Torchia L., Servili M., Rao R., Pietrella M., Giuliano G., Chiusano M.L., Baldoni L., Perrotta G. (2009) Comparative 454 pyrosequencing of transcripts from two olive genotypes during fruit development. *BMC genomics*, **10**, 15.

Altschul S.F., Madden T.L., Schaffer A.A., Zhang J., Zhang Z., Miller W., Lipman D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, **25**, 3389–3402.

Aury J.M., Cruaud C., Barbe V., Rogier O., Mangenot S., Samson G., Poulain J., Anthouard V., Scarpelli C., Artiguenave F., Wincker P. (2008) High quality draft sequences for prokaryotic genomes using a mix of new sequencing technologies. *BMC Genomics*, **9**, 603.

Barakat A., Wall K., Leebens-Mack J., Wang Y.J., Carlson J.E., dePamphilis C.W. (2007) Large-scale identification of microRNAs from a basal eudicot (*Eschscholzia californica*) and conservation in flowering plants. *Plant Journal*, **51**, 991–1003.

Barakat A., DiLoreto D.S., Zhang Y., Smith C., Baier K., Powell W.A., Wheeler N., Sederoff R., Carlson J.E. (2009) Comparison of the transcriptomes of American chestnut (*Castanea dentata*) and Chinese chestnut (*Castanea mollissima*) in response to the chestnut blight infection. *BMC Plant Biology*, **9**, 11.

Bartel D.P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, **116**, 281–297.

Bräutigam A., Shrestha R.P., Whitten D., Wilkerson C.G., Carr K.M., Froehlich J.E., Weber A.P.M. (2008) Comparison of the use of a species-specific database generated by pyrosequencing with databases from related species for proteome analysis of pea chloroplast envelopes. *Journal of Biotechnology*, **136**, 44–53.

Bundock P.C., Eliott F.G., Ablett G., Benson A.D., Casu R.E., Aitken K.S., Henry R.J. (2009) Targeted single nucleotide polymorphism (SNP) discovery in a highly polyploid plant species using 454 sequencing. *Plant Biotechnology Journal*, **7**, 347–354.

Cheung F., Haas B.J., Goldberg S.M.D., May G.D., Xiao Y.L., Town C.D. (2006) Sequencing *Medicago truncatula* expressed sequenced tags using 454 Life Sciences technology. *BMC Genomics*, **7**, 272.

Chevreux B., Pfisterer T., Drescher B., Driesel A.J., Muller W.E.G., Wetter T., Suhai S. (2004) Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. *Genome Research*, **14**, 1147–1159.

Cloonan N., Forrest A.R., Kolle G., Gardiner B.B., Faulkner G.J., Brown M.K., Taylor D.F., Steptoe A.L., Wani S., Bethel G., Robertson A.J., Perkins A.C., Bruce S.J., Lee C.C., Ranade S.S., Peckham H.E., Manning J.M., McKernan K.J., Grimmond S.M. (2008) Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nature Methods*, **5**(7), 613–619.

Cokus S.J., Feng S.H., Zhang X.Y., Chen Z.G., Merriman B., Haudenschild C.D., Pradhan S., Nelson S.F., Pellegrini M., Jacobsen S.E. (2008) Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature*, **452**, 215–219.

Dassanayake M., Haas J.S., Bohnert H.J., Cheeseman J.M. (2009) Shedding light on an extremophile lifestyle through transcriptomics. *New Phytologist*, **183**, 764–775.

Dolgosheina E.V., Morin R.D., Aksay G., Sahinalp S.C., Magrini V., Mardis E.R., Mattsson J., Unrau P.J. (2008) Conifers have a unique small RNA silencing signature. *RNA-A Publication of the RNA Society*, **14**, 1508–1515.

Eid J., Fehr A., Gray J., Luong K., Lyle J., Otto G., Peluso P., Rank D., Baybayan P., Bettman B., Bibillo A., Bjornson K., Chaudhuri B., Christians F., Cicero R., Clark S., Dalal R., Dewinter A., Dixon J., Foquet M., Gaertner A., Hardenbol P., Heiner C., Hester K., Holden D., Kearns G., Kong X.X., Kuse R., Lacroix Y., Lin S., Lundquist P., Ma C.C., Marks P., Maxham M., Murphy D., Park I., Pham T., Phillips M., Roy J., Sebra R., Shen G., Sorenson J., Tomaney A., Travers K., Trulson M., Vieceli J., Wegener J., Wu D., Yang A., Zaccarin D., Zhao P., Zhong F., Korlach J., Turner S. (2009) Real-time DNA sequencing from single polymerase molecules. *Science*, **323**, 133–138.

Emrich S.J., Barbazuk W.B., Li L., Schnable P.S. (2007) Gene discovery and annotation using LCM-454 transcriptome sequencing. *Genome Research*, **17**, 69–73.

Flicek P., Birney E. (2009) Sense from sequence reads: methods for alignment and assembly. *Nature Methods*, **6**, S6–S12.

Goda H., Sasaki E., Akiyama K., Maruyama-Nakashita A., Nakabayashi K., Li W.Q., Ogawa M., Yamauchi Y., Preston J., Aoki K., Kiba T., Takatsuto S., Fujioka S., Asami T., Nakano T., Kato H., Mizuno T., Sakakibara H., Yamaguchi S., Nambara E., Kamiya Y., Takahashi H., Hirai M.Y., Sakurai T., Shinozaki K., Saito K., Yoshida S., Shimada Y. (2008) The AtGenExpress hormone and chemical treatment data set: experimental design, data evaluation, model data analysis and data access. *Plant Journal*, **55**, 526–542.

Gordon D., Abajian C., Green P. (1998) Consed: a graphical tool for sequence finishing. *Genome Research*, **8**, 195–202.

Henderson I.R., Zhang X.Y., Lu C., Johnson L., Meyers B.C., Green P.J., Jacobsen S.E. (2006) Dissecting *Arabidopsis thaliana* DICER function in small RNA processing, gene silencing and DNA methylation patterning. *Nature Genetics*, **38**, 721–725.

Holt R.A., Jones S.J.M. (2008) The new paradigm of flow cell sequencing. *Genome Research*, **18**, 839–846.

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

Imelfort M., Duran C., Batley J., Edwards D. (2009) Discovering genetic polymorphisms in next-generation sequencing data. *Plant Biotechnology Journal*, **7**, 312–317.

Jones-Rhoades M.W., Bartel D.P., Bartel B. (2006) MicroRNAs and their regulatory roles in plants. *Annual Review of Plant Biology*, **57**, 19–53.

Jothi R., Cuddapah S., Barski A., Cui K., Zhao K. (2008) Genome-wide identification of in vivo protein–DNA binding sites from ChIP-Seq data. *Nucleic Acids Research*, **36**, 5221–5231.

Katayama S., Tomaru Y., Kasukawa T., Waki K., Nakanishi M., Nakamura M., Nishida H., Yap C.C., Suzuki M., Kawai J., Suzuki H., Carninci P., Hayashizaki Y., Wells C., Frith M., Ravasi T., Pang K.C., Hallinan J., Mattick J., Hume D.A., Lipovich L., Batalov S., Engstrom P.G., Mizuno Y., Faghihi M.A., Sandelin A., Chalk A.M., Mottagui-Tabar S., Liang Z., Lenhard B., Wahlestedt C. (2005) Antisense transcription in the mammalian transcriptome. *Science*, **309**, 1564–1566.

Kent W.J. (2002) BLAT – the BLAST-like alignment tool. *Genome Research*, **12**, 656–664.

Kilian J., Whitehead D., Horak J., Wanke D., Weinl S., Batistic O., D'Angelo C., Bornberg-Bauer E., Kudla J., Harter K. (2007) The AtGenExpress global stress expression data set: protocols, evaluation and model data analysis of UV-B light, drought and cold stress responses. *Plant Journal*, **50**, 347–363.

Laitinen R.A.E., Schneeberger K., Jelly N.S., Ossowski S., Weigel D. (2010) Identification of a spontaneous frame shift mutation in a non-reference *Arabidopsis thaliana* accession using whole genome sequencing. *Plant Physiology*, doi: 10.1104/pp.110.156448.

Langmead B., Trapnell C., Pop M., Salzberg S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, **10**, R25.

Lee R.M., Thimmapuram J., Thinglum K.A., Gong G., Hernandez A.G., Wright C.L., Kim R.W., Mikel M.A., Tranel P.J. (2009)

Sampling the Waterhemp (*Amaranthus tuberculatus*) Genome Using Pyrosequencing Technology. *Weed Science*, **57**, 463–469.

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

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

Li R., Yu C., Li Y., Lam T.W., Yiu S.M., Kristiansen K., Wang J. (2009) SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics*, **25**, 1966–1967.

Li R.Q., Fan W., Tian G., Zhu H.M., He L., Cai J., Huang Q.F., Cai Q.L., Li B., Bai Y.Q., Zhang Z.H., Zhang Y.P., Wang W., Li J., Wei F.W., Li H., Jian M., Li J.W., Zhang Z.L., Nielsen R., Li D.W., Gu W.J., Yang Z.T., Xuan Z.L., Ryder O.A., Leung F.C.C., Zhou Y., Cao J.J., Sun X., Fu Y.G., Fang X.D., Guo X.S., Wang B., Hou R., Shen F.J., Mu B., Ni P.X., Lin R.M., Qian W.B., Wang G.D., Yu C., Nie W.H., Wang J.H., Wu Z.G., Liang H.Q., Min J.M., Wu Q., Cheng S.F., Ruan J., Wang M.W., Shi Z.B., Wen M., Liu B.H., Ren X.L., Zheng H.S., Dong D., Cook K., Shan G., Zhang H., Kosiol C., Xie X.Y., Lu Z.H., Zheng H.C., Li Y.R., Steiner C.C., Lam T.T.Y., Lin S.Y., Zhang Q.H., Li G.Q., Tian J., Gong T.M., Liu H.D., Zhang D.J., Fang L., Ye C., Zhang J.B., Hu W.B., Xu A.L., Ren Y.Y., Zhang G.J., Bruford M.W., Li Q.B., Ma L.J., Guo Y.R., An N., Hu Y.J., Zheng Y., Shi Y.Y., Li Z.Q., Liu Q., Chen Y.L., Zhao J., Qu N., Zhao S.C., Tian F., Wang X.L., Wang H.Y., Xu L.Z., Liu X., Vinar T., Wang Y.J., Lam T.W., Yiu S.M., Liu S.P., Zhang H.M., Li D.S., Huang Y., Wang X., Yang G.H., Jiang Z., Wang J.Y., Qin N., Li L., Li J.X., Bolund L., Kristiansen K., Wong G.K.S., Olson M., Zhang X.Q., Li S.G., Yang H.M., Wang J., Wang J. (2010) The sequence and de novo assembly of the giant panda genome. *Nature*, **463**, 311–317.

Lin H., Zhang Z., Zhang M.Q., Ma B., Li M. (2008) ZOOM! Zillions of oligos mapped. *Bioinformatics*, **24**, 2431–2437.

Lister R., O'Malley R.C., Tonti-Filippini J., Gregory B.D., Berry C.C., Millar A.H., Ecker J.R. (2008) Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell*, **133**, 523–536.

Lister R., Gregory B.D., Ecker J.R. (2009) Next is now: new technologies for sequencing of genomes, transcriptomes, and beyond. *Current Opinion in Plant Biology*, **12**, 107–118.

Lu C., Kulkarni K., Souret F.F., MuthuValliappan R., Tej S.S., Poethig R.S., Henderson I.R., Jacobsen S.E., Wang W.Z., Green P.J., Meyers B.C. (2006) MicroRNAs and other small RNAs enriched in the Arabidopsis RNA-dependent RNA polymerase-2 mutant. *Genome Research*, **16**, 1276–1288.

Mardis E.R. (2008) Next-generation DNA sequencing methods. *Annual Review of Genomics and Human Genetics*, **9**, 387–402.

Matsumura H., Reich S., Ito A., Saitoh H., Kamoun S., Winter P., Kahl G., Reuter M., Kruger D.H., Terauchi R. (2003) Gene expression analysis of plant host–pathogen interactions by SuperSAGE. *Proceedings of the National Academy of Sciences USA*, **100**, 15718–15723.

Matsumura H., Kruger D.H., Kahl G., Terauchi R. (2008) SuperSAGE: a modern platform for genome-wide quantitative transcript profiling. *Current Pharmaceutical Biotechnology*, **9**, 368–374.

Metzker M.L. (2010) Sequencing technologies – the next generation. *Nature Reviews Genetics*, **11**, 31–46.

Meyers B.C., Tej S.S., Vu T.H., Haudenschild C.D., Agrawal V., Edberg S.B., Ghazal H., Decola S. (2004) The use of MPSS for whole-genome transcriptional analysis in Arabidopsis. *Genome Research*, **14**, 1641–1653.

Molina C., Rotter B., Horres R., Udupa S.M., Besser B., Bellarmino L., Baum M., Matsumura H., Terauchi R., Kahl G., Winter P. (2008) SuperSAGE: the drought stress-responsive transcriptome of chickpea roots. *BMC Genomics*, **9**, 553.

Moore M.J., Dhingra A., Soltis P.S., Shaw R., Farmerie W.G., Folta K.M., Soltis D.E. (2006) Rapid and accurate pyrosequencing of angiosperm plastid genomes. *BMC Plant Biology*, **6**, 13.

Morin R.D., Aksay G., Dolgosheina E., Ebhardt H.A., Magrini V., Mardis E.R., Sahinalp S.C., Unrau P.J. (2008) Comparative analysis of the small RNA transcriptomes of *Pinus contorta* and *Oryza sativa*. *Genome Research*, **18**, 571–584.

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

Nagalakshmi U., Wang Z., Waern K., Shou C., Raha D., Gerstein M., Snyder M. (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, **320**, 1344–1349.

Ning Z., Cox A.J., Mullikin J.C. (2001) SSAHA: a fast search method for large DNA databases. *Genome Research*, **11**, 1725–1729.

Nobuta K., Lu C., Shrivastava R., Pillay M., De Paoli E., Accerbi M., Arteaga-Vazquez M., Sidorenko L., Jeong D.H., Yen Y., Green P.J., Chandler V.L., Meyers B.C. (2008) Distinct size distribution of endogeneous siRNAs in maize: evidence from deep sequencing in the mop1-1 mutant. *Proceedings of the National Academy of Sciences USA*, **105**, 14958–14963.

Novaes E., Drost D.R., Farmerie W.G., Pappas G.J., Grattapaglia D., Sederoff R.R., Kirst M. (2008) High-throughput gene and SNP discovery in *Eucalyptus grandis*, an uncharacterized genome. *BMC Genomics*, **9**, 14.

Ossowski S., Schneeberger K., Clark R.M., Lanz C., Warthmann N., Weigel D. (2008) Sequencing of natural strains of *Arabidopsis thaliana* with short reads. *Genome Research*, **18**, 2024–2033.

Palmieri N., Schlotterer C. (2009) Mapping accuracy of short reads from massively parallel sequencing and the implications for quantitative expression profiling. *PLoS ONE*, **4**, 10.

Pepke S., Wold B., Mortazavi A. (2009) Computation for ChIP-seq and RNA-seq studies. *Nature Methods*, **6**, S22–S32.

Pop M., Salzberg S.L. (2008) Bioinformatics challenges of new sequencing technology. *Trends in Genetics*, **24**, 142–149.

Pushkarev D., Neff N.F., Quake S.R. (2009) Single-molecule sequencing of an individual human genome. *Nature Biotechnology*, **27**, 847–852.

Rajagopalan R., Vaucheret H., Trejo J., Bartel D.P. (2006) A diverse and evolutionarily fluid set of microRNAs in *Arabidopsis thaliana*. *Genes and Development*, **20**, 3407–3425.

Ronaghi M., Uhlén M., Nyrén P. (1998) DNA SEQUENCING: a sequencing method based on real-time pyrophosphate. *Science*, **281**, 363–365.

Sambrook J., Russell D.W. (2001) *Molecular Cloning. A Laboratory Manual*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor Laboratory.

Schmid M., Davison T.S., Henz S.R., Pape U.J., Demar M., Vingron M., Scholkopf B., Weigel D., Lohmann J.U. (2005) A gene expression map of *Arabidopsis thaliana* development. *Nature Genetics*, **37**, 501–506.

Schneeberger K., Ossowski S., Lanz C., Juul T., Petersen A.H., Nielsen K.L., Jorgensen J.E., Weigel D., Andersen S.U. (2009) SHOREmap: simultaneous mapping and mutation identification by deep sequencing. *Nature Methods*, **6**, 550–551.

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

Shendure J., Porreca G.J., Reppas N.B., Lin X.X., McCutcheon J.P., Rosenbaum A.M., Wang M.D., Zhang K., Mitra R.D., Church G.M. (2005) Accurate multiplex polony sequencing of an evolved bacterial genome. *Science*, **309**, 1728–1732.

Shin H., Hirst M., Bainbridge M.N., Magrini V., Mardis E., Moerman D.G., Marra M.A., Baillie D.L., Jones S.J.M. (2008) Transcriptome analysis for *Caenorhabditis elegans* based on novel expressed sequence tags. *BMC Biology*, **6**, 30.

Simon S.A., Zhai J., Nandety R.S., McCormick K.P., Zeng J., Mejia D., Meyers B.C. (2009) Short-read sequencing technologies for transcriptional analyses. *Annual Review of Plant Biology*, **60**, 305–333.

Steinmüller K., Apel K. (1986) A simple and efficient procedure for isolating plant chromatin which is suitable for studies of DNase I-sensitive domains and hypersensitive sites. *Plant Molecular Biology*, **7**, 87–94.

Sultan M., Schulz M.H., Richard H., Magen A., Klingenhoff A., Scherf M., Seifert M., Borodina T., Soldatov A., Parkhomchuk D., Schmidt D., O'Keeffe S., Haas S., Vingron M., Lehrach H., Yaspo M.L. (2008) A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science*, **321**, 956–960.

Sunkar R., Zhou X.F., Zheng Y., Zhang W.X., Zhu J.K. (2008) Identification of novel and candidate miRNAs in rice by high throughput sequencing. *BMC Plant Biology*, **8**, 25.

Tangphatsornruang S., Somta P., Uthaipaisanwong P., Chanprasert J., Sangsrakru D., Seehalak W., Sommanas W., Tragoonrung S., Srinives P. (2009) Characterization of microsatellites and gene contents from genome shotgun sequences of mungbean (*Vigna radiata* (L.) Wilczek). *BMC Plant Biology*, **9**, 137.

Torres T.T., Metta M., Ottenwalder B., Schlotterer C. (2008) Gene expression profiling by massively parallel sequencing. *Genome Research*, **18**, 172–177.

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

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

Velculescu V.E., Zhang L., Vogelstein B., Kinzler K.W. (1995) Serial analysis of gene expression. *Science*, **270**, 484–487.

Wall P.K., Leebens-Mack J., Chanderbali A.S., Barakat A., Wolcott E., Liang H.Y., Landherr L., Tomsho L.P., Hu Y., Carlson J.E., Ma H., Schuster S.C., Soltis D.E., Soltis P.S., Altman N., dePamphilis C.W. (2009) Comparison of next generation sequencing technologies for transcriptome characterization. *BMC Genomics*, **10**, 347.

Wang W., Wang Y.J., Zhang Q., Qi Y., Guo D.J. (2009) Global characterization of *Artemisia annua* glandular trichome transcriptome using 454 pyrosequencing. *BMC Genomics*, **10**, 10.

Weber A.P., Weber K.L., Carr K., Wilkerson C., Ohlrogge J.B. (2007) Sampling the Arabidopsis transcriptome with massively parallel pyrosequencing. *Plant Physiology*, **144**, 32–42.

Weigel D., Mott R. (2009) The 1001 Genomes Project for *Arabidopsis thaliana*. *Genome Biology*, **10**, 107.

Wicker T., Taudien S., Houben A., Keller B., Graner A., Platzer M., Stein N. (2009) A whole-genome snapshot of 454 sequences exposes the composition of the barley genome and provides evidence for parallel evolution of genome size in wheat and barley. *Plant Journal*, **59**, 712–722.

Wilhelm B.T., Landry J.R. (2009) RNA-Seq-quantitative measurement of expression through massively parallel RNA-sequencing. *Methods*, **48**, 249–257.

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

Willmitzer L., Wagner K.G. (1981) The isolation of nuclei from tissue-cultured plant cells. *Experimental Cell Research*, **135**, 69–77.

Yao Y.Y., Guo G.G., Ni Z.F., Sunkar R., Du J.K., Zhu J.K., Sun Q.X. (2007) Cloning and characterization of microRNAs from wheat (*Triticum aestivum* L.). *Genome Biology*, **8**, R96.

Zhao C.-Z., Xia H., Frazier T.P., Yao Y.-Y., Bi Y.-P., Li A.-Q., Li M.-J., Li C.-S., Zhang B.-H., Wang X.-J. (2010) Deep sequencing identifies novel and conserved microRNAs in peanut (*Arachis hypogaea* L.). *BMC Plant Biology*, **10**, 3.

Zhu Q.H., Spriggs A., Matthew L., Fan L.J., Kennedy G., Gubler F., Helliwell C. (2008) A diverse set of microRNAs and microRNA-like small RNAs in developing rice grains. *Genome Research*, **18**, 1456–1465.