Part One
Tag-Based Nucleic Acid Analysis

Tag-based Next Generation Sequencing, First Edition. Edited by Matthias Harbers and Günter Kahl.

1
DeepSuperSAGE: High-Throughput Transcriptome Sequencing with Now- and Next-Generation Sequencing Technologies

Hideo Matsumura, Carlos Molina, Detlev H. Krüger, Ryohei Terauchi, and Günter Kahl

Abstract

SuperSAGE is a variant of the serial analysis of gene expression (SAGE) expression profiling technology, in which 26-bp tags are extracted from cDNA using the type III restriction endonuclease *Eco*P15I. The use of a longer tag size in SuperSAGE allows a secure tag-to-gene annotation by homology searches against genome, transcript, or expressed sequence tag sequences. For organisms without genomic information, the 26-bp tags can be used as polymerase chain reaction primers to recover the full-length cDNA by 5′- and 3′-rapid amplification of cDNA ends. Here, we present the combination of SuperSAGE and high-throughput sequencing technologies (now- or next-generation sequencing (NGS)). We coin this merger deepSuperSAGE. The direct sequencing of millions of tag fragments shortens time and reduces costs for the analysis enormously. Furthermore, the incorporation of an indexing system expands the potential of deepSuperSAGE to analyze multiple samples in a single NGS run. The most recent version of deepSuperSAGE (high-throughput SuperSAGE) at least equals or even outcompetes microarrays in throughput. These improvements allow the application of deepSuperSAGE in transcriptome analysis in any eukaryotic system.

1.1
Introduction

Technologies for gene expression analysis have dramatically been improved over the past years. Northern blot analysis and polymerase chain reaction in combination with a reverse transcription reaction (reverse transcription-polymerase chain reaction RT-PCR) still are, to some extent, standard tools for expression analysis of individual genes. However, these techniques by all their virtue cannot be expanded to measure gene expression genome-wide and therefore will instead be used to analyze expression on a gene-by-gene basis, although it is possible to expand the analysis to 384 genes or more by multiplexing in the case of quantitative PCR (quantitative polymerase chain reactionqPCR, which is then called high-throughput real-time RT-PCR). This variant of qPCR – if controlled properly – allows an ultra-sensitive measurement of transcription by using gene-specific primers and probes in a PCR-based assay [1]. Although guidelines for the proper design of qPCR experiments have been established [2–4], a further increase in the number of addressed genes still meets with difficulties.

The recent explosion of information from genome and transcriptome sequencing projects now encourages analysis of the expression of a large number, preferably all, genes at a given time. Traditionally, microarrays of various architectures already

Tag-based Next Generation Sequencing, First Edition. Edited by Matthias Harbers and Günter Kahl.

represented tools for this kind of high-throughput gene expression profiling [5]. Microarrays are microscale solid supports (e.g., nylon membranes, nitrocellulose, glass, quartz, silicon, or other synthetic material) onto which either DNA fragments, cDNAs, oligonucleotides, genes, open reading frames, peptides, or proteins (e.g., antibodies) are spotted in an ordered pattern ("array") at extremely high density. Microarray-based expression profiling ("transcript profiling") for the simultaneous detection of the expression of thousands or tens of thousands of genes (the so-called "expressome"), whose complementary DNA sequences are immobilized on the array, requires the hybridization of fluorophore-labeled cDNAs from target tissue(s). After hybridization and high-stringency washing, the hybridization patterns can be visualized by fluorescence detection. It is appreciated that microarrays played a pioneering role in transcriptomics. However, their role in transcriptomics is fading [6]. The reasons for this decline are manifold. By all their virtue, microarrays of whatever format suffer from a series of devaluating insufficiencies. In fact, the poor correlation between different microarray platforms stands out (relatively large differences in data obtained in different labs using the same platform), but – equally important – its closed architecture format allows us to detect only the transcription of genes that are spotted on the array. Therefore, microarrays cannot detect novel genes. They require large amounts of input RNA for robust answers, which are at the most semiquantitative and at their best with the more abundant mRNAs. Microarrays are also prone to cross-hybridization of a single probe to different target RNAs and the experimenter has no reliable predictor for on-chip hybridization efficiency. Ambiguity exists in data analysis and interpretation, and in some cases defective oligonucleotides prior to printing have been reported. The widely different fluorescence intensity signals generated by different probes targeting the same gene confuses the experimenter. All these many inadequacies and inconsistencies of the microarray platform, and more so the irreproducibility of microarray-based results [7,8], which persisted in spite of many improvements, catalyzed the development of substitute technologies. For example, expressed sequence tag (EST) analysis – the large-scale sequencing of partial cDNA fragments – generating sequence information of thousands of expressed genes was and is extensively used. The number of sequence reads from a particular gene represents the expression level of the gene in the sample. However, ESTs are also only a sample of the whole transcriptome, contain a high sequence error rate (up to 3%), are relatively short (average 400 bp), contain artifacts such as vector and bacterial sequence contaminations, only represent 5′ and/or 3′ ends of transcripts, suffer from a bias in the dbEST database (http://www.ncbi.nlm.nih.gov/dbEST; splice variants involving exons located in the center of long transcripts are under-represented), and due to high sequencing costs EST analysis has not always been a suitable method in terms of throughput.

In a seminal publication, Velculescu *et al.* [9] reported a method to count the transcripts in a high-throughput manner, which they named serial analysis of gene expression (SAGE). In SAGE, originally a short fragment of 13–15 bp in size is isolated as a tag from a defined position of each cDNA. Those tags are then concatenated and cloned into a plasmid vector for sequencing. The key to the SAGE technique is the use of the type IIS restriction endonuclease *Bsm*FI as the tagging enzyme that extracts tag fragments from transcripts. *Bsm*FI cuts 13–15 bases away from its recognition site, allowing the isolation of 13- to 15-bp tag sequences from cDNAs. Each transcript is uniquely represented by a tag fragment and the tag frequency in the sample (tag count) reflects the abundance of the corresponding transcript. The obtained 13- to 15-bp tag sequence can be used as query by BLAST search against EST databases of the species from which the tag sequence is derived (tag annotation). By listing the count and annotation of thousands of tags, one can obtain a comprehensive and quantitative profile of gene expression. In contrast to analog datasets generated by hybridization-based methods like microarrays, SAGE data are digital and easy to handle bioinformatically. SAGE is an open-architecture method whereby researchers can theoretically address all the expressed transcripts

simply by increasing the number of tags to be analyzed. All these advantages make SAGE superior to microarrays as a closed-architecture method.

However, the original SAGE method had a major problem of accuracy in tag-to-gene annotation, owing to the short size of the tag. To overcome this inadequacy, improved versions of SAGE were established, that obtained longer tag sequences from cDNAs. For example, LongSAGE [10] employed *Mme*I to isolate 21-bp tags and, more recently, SuperSAGE [11] has generated much longer tags (26 bp). Recent rapid advancements of DNA sequencing technologies dramatically improved the Super-SAGE protocol by increasing throughput and reducing analytical cost. The merger of SuperSAGE with one of the "now- or next-generation sequencing" (NGS) platforms [11] is known as deepSuperSAGE or also high-throughput SuperSAGE (HT-SuperSAGE) [12,23]. The potential of this technology for genome-wide and quantitative gene expression profiling has now been amply demonstrated and will be addressed in the present chapter.

1.2 Overview of the Protocols

1.2.1 Principle of the SuperSAGE Method

SuperSAGE is an improved version of the SAGE technology, whereby 26-bp tags are extracted from cDNA using the type III enzyme *Eco*P15I [13,14]. The distance between recognition and cleavage sites of *Eco*P15I is the longest for all the known restriction enzymes, which can cleave 25/27 bp away from its recognition site [15].

Basically, the experimental procedure of SuperSAGE is similar to that of the original SAGE, except for the tagging enzyme, oligo(dT) primers, and linkers. All the details from cDNA synthesis to tag extraction are described in the following protocol. For an efficient DNA digestion with *Eco*P15I, two copies of its recognition sequence 5′-CAGCAG-3′ should be located in head-to-head orientation within the target DNA molecule [13]. Therefore, one 5′-CAGCAG-3′ site is inserted into the adapter-oligo (dT) primer sequence and the other site is incorporated in the linkers, which are ligated to the digested cDNA. These linker-ligated cDNA fragments are cleaved by *Eco*P15I at a position 25/27 bp away from either of the recognition sites in the linkers and adapter-(dT) primers. Two "linker–tag" fragments are ligated in head-to-head orientation to generate "linker–ditag–linker" fragments and the resulting fragments are amplified by polymerase chain reactionPCR. After removal of linkers, ditags are concatenated and the concatemers are cloned into a plasmid vector for sequencing. From sequencing reads of plasmid inserts (concatemers), tag sequences are extracted. Although most of the tags are expected to be 27 bp in size, a considerable number of 26-bp tags was actually obtained, as also described in Section 1.2.5. Therefore, we defined a 26-bp sequence as the tag from concatemer sequences.

1.2.2 Power of the SuperSAGE Tag

With the increased tag length (26 or 27 bp), the efficiency of tag-to-gene annotation is considerably improved. In model organisms, 26-bp tags allow almost perfect gene annotation by a BLAST search against the genome or cDNA sequence databases [12]. BLAST analysis with tags of different sizes (15, 20, or 26 bp) convincingly demonstrated that 15- and 20-bp tags usually match DNA sequences of multiple species, whereas the 26-bp SuperSAGE tags matches DNA sequences of a single species in most of the cases [12]. Therefore, the sequence information of 26-bp tags can uniquely identify the gene and species from which the tag was derived. Using this high-specificity of a SuperSAGE tag, it allows simultaneous monitoring of gene expression

of two or more species synchronously that are in a tight interaction (e.g., a pathogen and its host cells) [16,17].

An additional advantage of the 26-bp SuperSAGE tags is that the full or partial sequence of the corresponding genes could easily be recovered by PCR. This allows the analysis of transcriptomes even in nonmodel organisms. For recovery of corresponding genes from tag sequences, rapid amplification of cDNA ends (RACE) is the most conventional method [18]. By a combination of 3′- and 5′-RACE, sequences of several full-length cDNAs were obtained easily starting from 26-bp tag sequences in *Nicotiana benthamiana* [19]. Alternatively, Coemans *et al.* [20] succeeded in amplifying genes corresponding to the tags by thermal asymmetric interlaced (TAIL)-PCR) using genomic *Musa accuminata* (banana) DNA as template. This method recovered corresponding genes including their promoter regions from the tag without preparing a high-quality cDNA template.

In summary, the high specificity of tag-to-gene annotation and its applicability to nonmodel organisms are the two major advantages of SuperSAGE.

1.2.3 Development of DeepSuperSAGE

Recent advances in DNA sequencing technologies – the NGS platforms – are dramatically changing the whole research strategy in biological studies. These technologies aim at reading huge amounts of DNA sequences in a short time at low cost. Currently available NGS technologies are based on massively parallel sequencing, which produces sequences of more than millions of DNA fragments at a time. The output of the NGS DNA sequencers is a huge number of short sequences, so-called "reads." This feature of NGS is extremely suitable for sequencing the 26-bp SuperSAGE tags. Thus, we have tried to combine SuperSAGE and an NGS technology to establish deepSuperSAGE, which greatly reinforces the traditional SuperSAGE technology.

The first NGS instrument released was the Genome Sequencer ("GS" series) from 454 Life Sciences in 2005 [21]. This sequencer employs pyrosequencing and the average read length spans from 100 (GS20) to 400 bp (GS FLX Titanium). We developed a protocol for direct sequencing of SuperSAGE ditags with linkers for the GS20 DNA sequencer (Figure 1.1) [22]. Afterwards, more powerful massively parallel sequencers continuously emerged. Since the read length of these machines is short (less than 35–50 bp), fragments containing single tags (not ditags) were applied to these sequencers (Figure 1.1) [23].

This deepSuperSAGE technology allows a high-throughput analysis of any transcriptome. The advantages of this method include:

i) Huge numbers of 26-bp tags (more than 1 million) are obtained in a single sequencing run.
ii) DNA fragments containing tags or ditags are directly sequenced without plasmid cloning.
iii) Tags from several independent samples can be pooled and analyzed together in a single run by employing index (barcode) sequences in the linker or adapter.

Since increasing the numbers of analyzed tags apparently contributes to improve accuracy of profiling data, it is promising that high-quality data can be obtained in deepSuperSAGE analysis. Additionally, analytical costs are reduced, owing to the lower sequencing costs per base in NGS. In the original SuperSAGE protocol, concatenation of ditags and plasmid cloning were necessary for sequencing [12]. Using this approach, it was not easy to optimize cloning efficiency and obtain clones with large inserts. Even after a high-quality library was constructed, several hundreds of clones or inserts had to be prepared. DeepSuperSAGE now avoids all these steps and, as a consequence, greatly contributes to reduction in effort, time, and costs. Previously, SuperSAGE was regarded as a gene expression profiling method for a limited number of samples, because the time, cost, and effort required proportionally

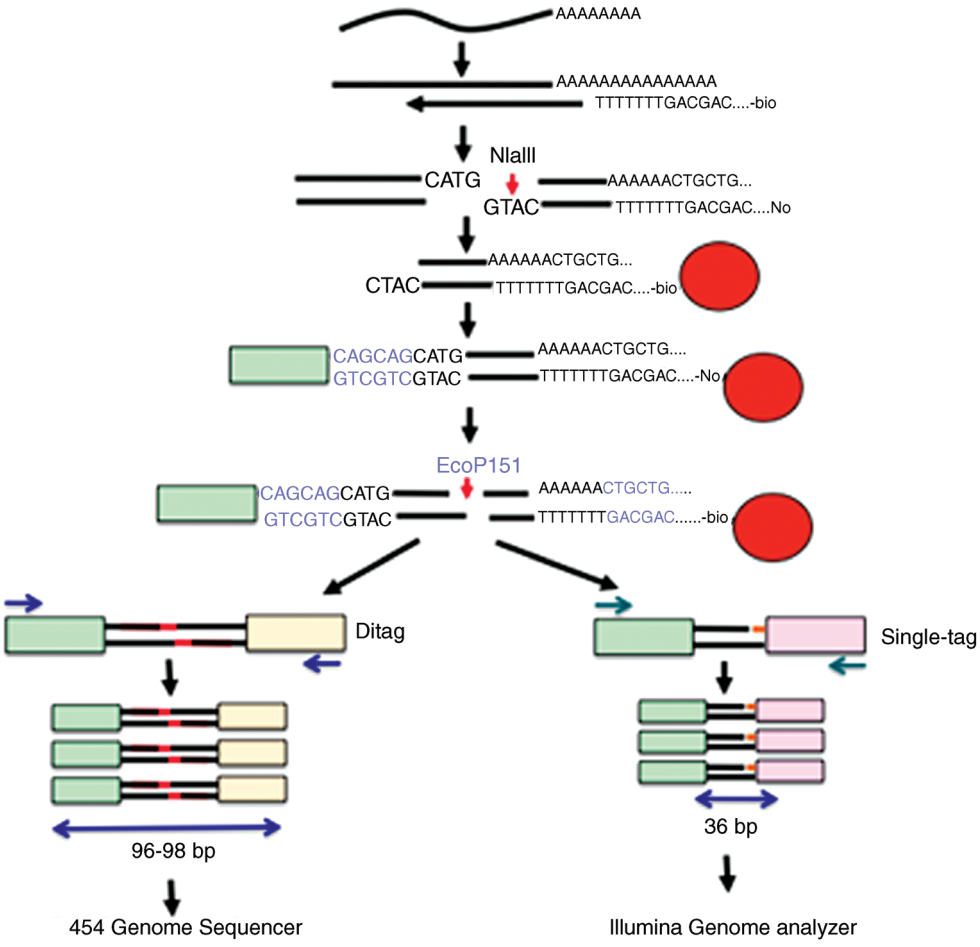

Fig. 1.1 Scheme of deepSuperSAGE. After *Eco*P15I digestion of linker (adapter)-ligated cDNA fragments immobilized on paramagnetic beads, ditags were formed for 454 pyrosequencing analysis (left) or another adapter was immediately ligated to the *Eco*P15I digestion end (single-tag) for Illumina GA analysis (right). Sizes of sequenced fragments were 96–98 bp in the ditag protocol and 36 bp in the single-tag protocol.

increased with sample numbers. By employing the multiplexing protocol in deepSuperSAGE, the SuperSAGE technology is now applicable to many samples without increasing the time, cost, and effort. In combination with NGS, digital gene expression (DGE) and RNA-seq are used commonly for high-throughput transcript profiling [24,25]. The DGE protocol for the Illumina Genome Analyzer (GA) platform was based on LongSAGE. However, deepSuperSAGE turned out to be superior, due to the longer size of the obtained tags. RNA-seq, on the other hand, is suitable to understand the structure of transcripts rather than quantifying amounts of transcripts. Consequently, we suggest that deepSuperSAGE is still the best method of tag-based quantitative transcriptome analysis employing NGS.

The method and some of its applications will be described, separately for (i) ditag- and (ii) single-tag-based deepSuperSAGE.

1.2.4
Ditag-Based DeepSuperSAGE (for 454 Pyrosequencing)

The first version of the released 454 pyrosequencer (GS20) produced reliable sequence reads of 100 bp from each fragment. Coincidentally, the size of a SuperSAGE "linker–ditag–linker" fragment generated after PCR amplification is 96–98 bp, which perfectly fits the size of a single sequence path of GS20 sequencing. Therefore, amplified fragments directly served as sequencing templates without concatenation and plasmid cloning. A single sequencing run produces sequences of 200 000–1 000 000 ditags on average, indicating that a total of 400 000–2 000 000 tags could well be obtained. Since in this procedure linker regions are also sequenced together with ditags, we considered using different linker fragments with unique sequences to

generate individual SuperSAGE libraries and separating sequencing data from independent samples based on the linker sequences. Introduction of this improvement allowed a multiplexed SuperSAGE analysis of different samples. Thereby, the scale of multiplexing and tag count for each sample can be flexibly changed and adapted depending on research objectives.

Experimental steps from RNA to ditag amplification and purification were almost identical to the original SuperSAGE protocol as described later. Generally, starting from several hundred microgram total RNA (1–3 μg poly(A)$^+$ RNA), 1 μg of amplified ditags is obtained from 40 PCR reactions. Successful sequencing will provide more than 200 000 sequence reads. For tag extraction from sequence data, several processes are required, including elimination of incomplete (short) sequence reads, sorting libraries by linker sequences (if multiplexed), and exclusion of duplicated ditag sequences. For this purpose, we developed our own programs, such as SuperSAGE_tag_extract_pipe [22] or GXP-Tag sorter (GenXPro) [26].

1.2.5 Single-Tag-Based DeepSuperSAGE (HT-SuperSAGE)

After the release of the 454 pyrosequencer, other NGS technologies became available, as, for example, the Illumina GA, based on "sequencing-by-synthesis" (SBS), and the Applied Biosystems SOLiD system, based on "sequencing-by-ligation" methods. These DNA sequencers enabled 100 000 000 reads in a single run. It was expected that a complete transcript profile would be obtained when these sequencers are used for deepSuperSAGE. In the early version of these sequencers (GA or SOLiD), the size of a sequence read was typically 35 or 36 bp, shorter than the ditag length (52 bp). Therefore, for the adaptation of deepSuperSAGE to GA or SOLiD sequencers, a single-tag sequencing protocol was designed. It basically follows the original SuperSAGE or deepSuperSAGE workflow for 454 pyrosequencing up to the step of 26-bp tag extraction. However, after this step, no "ditags" are formed. Instead, two adapter fragments are ligated to each end of the single tag. At this step, we skip the purification from a polyacrylamide gel electrophoresis (PAGE) gel and the fill-in reaction of *Eco*P15I-digested fragments, which are necessary in the original and 454 pyrosequencing SuperSAGE protocols. This measure reduces the time for experiments and avoids loss of DNA fragments. Single tags flanked by the adapters are amplified by PCR. Finally, PCR products of the expected size (accurately containing adapters and tag) are purified and applied to direct sequencing.

In this protocol, two additional improvements were included:

i) The number of PCR amplification cycles of adapter–tag fragments was reduced.
ii) For the analysis of multiple samples in a single sequencing run, a systematic indexing (barcoding) was employed.

By incorporating these improvements, we developed the protocol of HT-SuperSAGE [23].

We were concerned that removal of duplicated ditags could not be integrated in the single-tag protocol and therefore expected distortion of transcript profiles due to PCR amplification biases [9]. To avoid this potential problem, PCR cycles were reduced to 5–10 cycles. By comparing tag profiling among different PCR cycles (3, 5, and 10 cycles), we could assure that an increase in PCR cycle numbers up to 10 did not cause any significant distortion in the expression profiles [23]. Since the required amount of template DNA for sequencing on Illumina GA platforms is about 10 ng, sufficient template DNA can be prepared by 10 PCR cycles.

In our sequence data, tags with various sizes were observed. If sorted by length, we found that 27-bp tags made up 66% and 26-bp tags made up 25% of all tags. Tags with other sizes were under-represented [23]. Therefore, tags can be recovered from more than 90% of all sequence reads by extracting 26-bp tag sequences.

The strategy for sample multiplexing was already employed in deepSuperSAGE using 454 pyrosequencing. The Illumina GA has larger sequencing capacities and

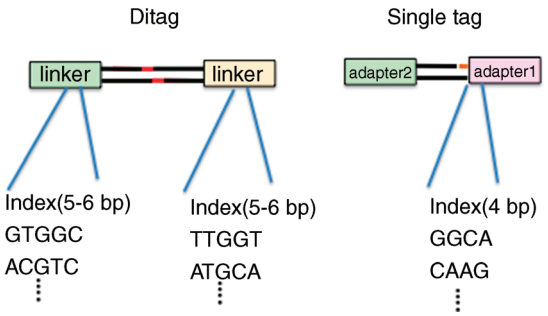

Fig. 1.2 Position of index for multiplexing. Index sequences were located in the linker or adapter sequences. For ditag sequence analysis (left), a 5- or 6-bp index sequence was incorporated within 10 bp upstream of the *Eco*P15I site in the linkers. For single-tag (HT-SuperSAGE) analysis (right), 4-bp index sequences were located adjacent to the sequencing primer.

therefore allows pooling much larger numbers of samples. For this purpose, a systematic indexing protocol should be developed. For the design of this index it is recommended that it should be located close to the tag sequence, due to limitations in read length. Yamaguchi *et al.* [27] combined SuperSAGE with the Illumina GA and employed a 2-base index upstream of the *Eco*P15I site in the adapter. In our established protocol, we have designed a 4-base index just downstream of the sequencing primer site (Figure 1.2) [23]. Adapter fragments with different index sequences are separately ligated to 26-bp tag fragments derived from different samples. Adapter–tag fragments from different libraries are pooled and sequenced together. The sequence reads are separated *in silico* according to their index sequences. By positioning the index in the first 4 bases of the sequence read, the frequency of sequencing errors is minimized.

1.3 Methods and Protocols

1.3.1 Linker or Adapter Preparation

454 Pyrosequencing

Linker DNAs for SuperSAGE are prepared by annealing the two complementary oligonucleotides, as shown in Table 1.1 (Linker-1A, -1B, -2A, -2B). Linker DNAs have cohesive ends, which are compatible with the end generated by *Nla*III digestion (5′-CATG-3′). An *Eco*P15I recognition site (5′-CAGCAG-3′) is present adjacent to the 5′-CATG-3′ site. The 3′ ends of the Linker-*X*Bs should be amino-modified to prevent ligation to the cDNA or another linker molecule at this site. We can synthesize several different pairs of linker DNAs (Linker-1, -2, -3, -4, etc.) for the preparation of multiple SuperSAGE libraries. In these linkers, sequence variation of 5–6 bp is incorporated within the 10-bp region upstream of the *Eco*P15I recognition site as an index

Table 1.1 Oligonucleotide sequences for linkers or adapters in deepSuperSAGE.

Name	Sequences
Linker-1A	5′-TTTGGATTTGCTGGTGCAGTACAACTAGGCTTAATACAGCAGCATG
Linker-1B	5′-CTGCTGTATTAAGCCTAGTTGTACTGCACCAGCAAATCCAAA-amino
Linker-2A	5′-TTTCTGCTCGAATTCAAGCTTCTAACGATGTACGCAGCAGCATG
Linker-2B	5′-CTGCTGCGTACATCGTTAGAAGCTTGAATTCGAGCAGAAA-amino
Adapter-1A	5′-ACAGGTTCAGAGTTCTACAGTCCGACGATCXXXX[a)]
Adapter-1B	5′-NNYYYYGATCGTCGGACTGTAGAACTCTGAACCTGT-amino[a)]
Adapter-2A	5′-CAAGCAGAAGACGGCATACGATCTAACGATGTACGCAGCAGCATG
Adapter-2B	5′-CTGCTGCGTACATCGTTAGATCGTATGCCGTCTTCTGCTTG

a) XXXX and YYYY indicate arbitrary index sequences. Each of them should be complementary.

(Figure 1.2). In this protocol, we show the sequences of only Linker-1 and Linker-2 (Table 1.1).

1.1 Dissolve the synthetic linker oligonucleotides (Linker-1A, -1B, -2A, -2B) in LoTE buffer (3 mM Tris–HCl, pH7.5; 0.2 mM EDTA), so that their concentration is 1 μg/μl.
1.2 Mix 1 μl Linker-1B (or Linker-2B), 1 μl 10× polynucleotide kinase buffer, 1 μl 10 mM ATP, 7 μl H_2O, and 1 μl T4 polynucleotide kinase, and incubate at 37 °C for 30 min to phosphorylate the 5′ ends.
1.3 Add 1 μl Linker-1A or -2A to the 5′-phosphorylated Linker-1B or -2B solution from the previous step, respectively.
1.4 After mixing, denature by incubating at 95 °C for 2 min and cool down to 20 °C for annealing.
1.5 The annealed double-stranded DNAs (200 ng/μl) are designated as Linker-1 and Linker-2, respectively.

HT-SuperSAGE

The procedure for HT-SuperSAGE adapter preparation basically follows the linker preparation for 454 sequencing libraries described above. Sequences of adapter oligonucleotides were changed for Illumina GA sequencing (Table 1.1). Adapter-1 has a 4-bp index sequence ("XXXX" in Table 1.1) and a 2-base cohesive end of "NN" for ligating to *Eco*P15I-digested tag ends. The annealed double-stranded DNAs (200 ng/μl) are designated as Adapter-1 and Adapter-2, respectively. Adapter-2 carries a cohesive end for the *Nla*III site (CATG) and an *Eco*P15I recognition site (5′-CAGCAG-3′) adjacent to the *Nla*III site.

1.3.2 RNA Samples

About 20–30 μg of total RNA as starting material for 454 sequencing allows a successful deepSuperSAGE experiment. For HT-SuperSAGE, 1–10 μg total RNA is sufficient.

1.3.3 cDNA Synthesis and NlaIII Digestion

The protocols for cDNA synthesis and *Nla*III digestion do not depend on any sequencing technology. Any cDNA synthesis protocol is applicable to SuperSAGE, but biotinylated adapter-oligo(dT) primer (5′-biotin- CTGATGTAGAGGTACCGGA TGCCAGCAGTTTTTTTTTTTTTTTTTT-3′) should be used for reverse transcription. We employ the SuperScript II double-strand cDNA synthesis kit (Invitrogen), following the experimental procedures given in its instruction manual.

3.1 After second-strand cDNA synthesis, double-stranded cDNA is purified by passing it through a column (QIAquick PCR purification kit; Qiagen) instead of phenol/chloroform extraction and ethanol precipitation.
3.2 Purified cDNA (50 μl eluted DNA from a column) is completely digested with *Nla*III, by adding 20 μl *Nla*III digestion buffer (NEBuffer 4), 2 μl bovine serum albumin (BSA), 123 μl LoTE, 5 μl *Nla*III (10 U/μl; NEB).
3.3 Incubate at 37 °C for 1.5 h.

1.3.4 Tag Extraction from cDNA

454 Pyrosequencing

4.1 Digested cDNA solution (without purification) is divided into the two tubes, tube A and tube B (each 100 μl).

4.2 Tube A and tube B both contain cDNA to be ligated with Linker-1 and Linker-2, respectively, as described above.
4.3 An equal volume of 2× B&W buffer (10 mM Tris–HCl, pH 7.5; 1 mM EDTA; 2 M NaCl) is added to each of the tubes A and B.
4.4 Contents of tubes A and B are separately added to the washed streptavidin-coated paramagnetic beads (Dynabeads M-270).
4.5 Biotinylated cDNA fragments are associated with streptavidin-coated magnetic beads by incubation at room temperature for 30 min.
4.6 After washing the beads 3 times with 1× B&W buffer and once with LoTE buffer, Linker-1 and Linker-2, respectively, are ligated to the ends of the cDNAs bound to the magnetic beads in the two tubes.
4.7 For ligation, 200 ng linker DNA is usually added to a tube.
4.8 To ligate linkers to digested cDNAs bound to the magnetic beads, add 21 µl LoTE, 6 µl 5× T4 DNA ligase buffer, and either 1 µl Linker-1 or -2 solution (~100–200 ng), respectively.
4.9 The bead suspension is incubated at 50 °C for 2 min for the dissociation of linker dimers and kept at room temperature for 15 min.
4.10 T4 DNA ligase (10 U) is added and the tubes are incubated at 16 °C for 2 h.
4.11 After ligating the linkers, the bead suspension from the two tubes is mixed.
4.12 The beads are washed 4 times with 1× B&W buffer, followed by washing with LoTE buffer for 3 times.
4.13 The resulting linker-cDNA fragments on the beads are digested with *Eco*P15I to release "linker–tag" fragments.
4.14 For *Eco*P15I digestion, 10 µl 10× *Eco*P15I digestion buffer (100 mM Tris–HCl, pH 8.0; 100 mM KCl; 100 mM $MgCl_2$; 1 mM EDTA; 1 mM dithiothreitol; 50 µg/ml BSA), 2 µl 100 mM ATP, 83 µl sterile water, and 5 µl *Eco*P15I (2 U/µl; NEB) are added to the washed paramagnetic beads.
4.15 Tubes are incubated at 37 °C for 2 h.

1.3.5
Tag Extraction from cDNA

HT-SuperSAGE

5.1 Prepare a 100-µl suspension of streptavidin-coated magnetic beads (Dynabeads M-270) in a siliconized 1.5-ml microtube. Beads are washed once with 1× B&W buffer.
5.2 To the washed magnetic beads, 200 µl 2× B&W solution and 200 µl digested cDNA solution are added and suspended well.
5.3 After the digested cDNAs are associated with the beads for 30 min incubation, the tube is placed on the magnetic stand, and the supernatant is discarded.
5.4 Magnetic beads are washed 3 times with 200 µl 1× B&W and once with 200 µl LoTE.
5.5 For Adapter-2 ligation to the digested cDNAs, 21 µl LoTE, 6 µl 5× T4 DNA ligase buffer, and 1 µl Adapter-2 solution are added to the magnetic beads.
5.6 After mixing with pipettes, the bead suspension is incubated at 50 °C for 2 min for the dissociation of adapter dimers.
5.7 Tubes are kept at room temperature for 15 min.
5.8 After the tubes cooled down, 2 µl T4 DNA ligase (10 U) is added and incubated at 16 °C for 2 h with occasional mixing.
5.9 After ligation reaction, beads are washed 4 times with 1× B&W and 3 times with LoTE.
5.10 The beads are suspended in 75 µl LoTE.
5.11 For *Eco*P15I digestion, 10 µl 10× NEBuffer 3, 10 µl 10× ATP solution (1 mM), 1 µl 100× BSA (100 µg/ml), and 4 µl *Eco*P15I are added to the suspended magnetic beads.
5.12 Incubate the tube at 37 °C for 2 h.

1.3.6
Purification of Linker–Tag Fragments

6.1 In both sequencing methods, the DNA released from the beads after *Eco*P15I digestion is extracted by phenol/chloroform.
6.2 Precipitate DNA by adding 100 µl 10 M ammonium acetate, 3 µl glycogen, and 950 µl cold ethanol.
6.3 The tube is kept at −80 °C for 1 h.
6.4 The DNA is precipitated by centrifugation at 15 000 × *g* for 40 min at 4 °C and the resulting pellet is washed once with 70% ethanol.
6.5 After drying, the pellet is dissolved in 10 µl LoTE buffer.

454 Pyrosequencing

6.6 Dissolved DNA solution is loaded onto an 8% PAGE gel, which is prepared by mixing 3.5 ml 40% acrylamide/bis solution, 13.5 ml dH_2O, 350 µl 50× TAE (Tris–acetate–EDTA) buffer, 175 µl 10% ammonium persulfate, and 15 µl TEMED.
6.7 The polyacrylamide gel is run at 75 V for 10 min and then at 150 V for around 30 min.
6.8 The gel is stained with SYBR Green (Molecular Probes) and the DNA visualized on a UV trans-illuminator.
6.9 The "linker–tag" fragments of expected size (around 70 bp) are cut out and put into a 0.5-ml tube.
6.10 Holes are made at the top and the bottom of the tube with a needle and it is placed in a 2-ml tube.
6.11 The tube is centrifuged at the maximum speed for 2–3 min (table centrifuge).
6.12 Polyacrylamide gel pieces are collected at the bottom of the 2-ml tube and 300 µl LoTE is added to the gel pieces for resuspension.
6.13 After incubation at 37 °C for 2 h, the gel suspension is transferred to a Spin-X column (Corning) and centrifuged at maximum speed for 2 min.
6.14 Collected solution at the bottom of the tube is extracted by phenol/chloroform and precipitated as described above.
6.15 After once washing with 70% ethanol, the dried linker–tag DNA is dissolved in 8 µl LoTE buffer.

HT-SuperSAGE

Further purification of *Eco*P15I-digested fragments was not necessary.

1.3.7
Ditag or Adapter–Tag Formation and Amplification

454 Pyrosequencing

7.1 Purified "linker–tag" fragments (a mixture of Linker-1–tag and Linker-2–tag fragments) are blunt-ended by fill-in reaction using Blunting High Kit (Toyobo).
7.2 To the linker–tag solution (8 µl), 1 µl 10× blunting buffer and 1 µl KOD DNA polymerase (Toyobo) are added.
7.3 The tube is incubated at 72 °C for 2 min and immediately transferred onto ice.
7.4 For ditag formation, 30 µl LoTE and 40 µl Ligation High (Toyobo) are added to the 10 µl blunt-ended reaction.
7.5 After incubation of the ligation reaction mixture at 16 °C for 4 h to overnight, a small aliquot of the ligation product is removed and diluted (1/5 and 1/10) with LoTE buffer.
7.6 These diluents are used as templates for the PCR amplification of the "linker–ditag–linker" fragments.
7.7 For Linker-1 and Linker-2, we use PCR primers with the sequence 5′- CAAC-TAGGCTTAATACAGCAGCA-3′ and 5′- CTAACGATGTACGCAGCAGCA-3′, respectively.

7.8 If other linkers with different indexes are employed, PCR primers should be changed, according to used linker sequences.
7.9 Hot-start PCR is not always necessary for amplifying "linker–ditag–linker" fragments.
7.10 We amplify "linker–ditag–linker" in a reaction mixture containing 5 µl 10× PCR buffer, 5 µl 2 mM dNTP, each 0.2 µl primer (350 ng/µl), 38.34 µl dH_2O, 1µl diluted template solution, and 0.26 µl *Taq* DNA polymerase (5 U/µl).
7.11 We amplify "linker–ditag–linker" with the following reaction cycle: 94 °C for 2 min, then 25 cycles each at 94 °C for 40 s and 60 °C for 40 s.
7.12 With the pilot PCR experiment, we determine which of the 1/5 and 1/10 template dilutions gives the better amplification of the "linker–ditag–linker."
7.13 PCR products (96–98 bp) are observed in a SYBR Green-stained acrylamide gel.
7.14 A bulk PCR is carried out under the same conditions for 40–48 tubes, each containing 50 µl, using diluted template (either of 1/5 or 1/10 dilutions) that gave the better amplification in the pilot PCR (see above).
7.15 All PCR products are collected in a tube and purified with QIAquick PCR purification kit (Qiagen).
7.16 For purification, six to eight columns are used and eluted DNAs from all the columns are collected in a single tube.
7.17 This DNA solution (180–240 µl) is loaded onto an 8% polyacrylamide gel.
7.18 After running the gel and staining with SYBR Green as described above, the separated DNA fragments of the expected size (96–98 bp) are cut out from the gel.
7.19 DNA is eluted from the polyacrylamide gel and purified by ethanol precipitation after phenol/chloroform extraction, as described above. Around 1 µg of purified "linker–ditag–linker" fragments can be obtained from 40–48 PCR reaction tubes.

HT-SuperSAGE

7.20 Prepare Adapter-1 with defined index sequences assigned to individual samples (Adapter-1a, -1b, -1c, etc.).
7.21 For ligation of Adapter-1, 3 µl 5× T4 DNA ligase buffer and 0.5 µl Adapter-1 solution are added to the solution of the Adapter-2-ligated tags.
7.22 Incubate the tube at 50 °C for 2 min and keep it at room temperature for 15 min.
7.23 After the tubes cooled down, 1.5 µl T4 DNA ligase (7.5 U) is added and incubated at 16 °C for 2 h.
7.24 For PCR amplification of adapter-ligated tag fragments, PCR reaction mixture, containing 3 µl 5× Phusion HF buffer, 0.3 µl 2.5 mM dNTP, 0.1 µl 50 mM $MgCl_2$, 0.15 µl Adapter-1 primer, 0.15 µl Adapter-2 primer, 10.1 µl dH_2O, 1 µl ligation solution, and 0.2 µl Phusion Hot Start DNA polymerase, is prepared in a tube.
7.25 PCR reaction proceeds under the following conditions: 98 °C for 2 min, then 5–10 cycles each at 98 °C for 30 s and 60 °C for 30 s.
7.26 Prepare an 8% PAGE gel by mixing 3.5 ml 40% acrylamide/bis solution, 13.5 ml dH_2O, 350 µl 50× TAE buffer, 175 µl 10% ammonium persulfate, and 15 µl TEMED.
7.27 Running buffer (1× TAE) is prepared and added to the upper and lower electrophoresis chambers.
7.28 Then 3 µl 6× loading dye is added to 15 µl of the PCR solution and loaded into the well.
7.29 An aliquot of 2 µl of a 20-bp marker ladder is also loaded as molecular size marker. Run the gel at 75 V for 10 min and then at 150 V for around 30 min.
7.30 After staining the gel with SYBR Green, it was visualized on a UV illuminator. The size of the expected amplified fragment (tags sandwiched with two adapters) is 123–125 bp.
7.31 Repeat PCR reactions under the same condition in 8–14 tubes.

7.32 After the PCR reaction, solutions from all the tubes are collected in a 1.5-ml tube and purified by MinElute Reaction Cleanup kit or by ethanol precipitation.
7.33 Prepare 8% polyacrylamide gel as described in Step 5.6. Add 3 µl 6× loading buffer to purified PCR product and load it in the well.
7.34 After running the gel as described above, the gel is stained with SYBR Green and bands are visualized under UV light.
7.35 Only the 123- to 125-bp band (Adapter-1 and Adapter-2 ligated 26- to 27-bp tag) is cut out from the gel and transferred to a 0.5-ml microtube.
7.36 Elution and purification of DNA in the gel was done as described above.
7.37 Finally, the resulting pellet after ethanol precipitation is dissolved in 10–15 µl LoTE.

1.3.8 Preparation of Templates for Sequencing

454 Pyrosequencing

Purified DNA is ready for sequencing analysis after adapter ligation for 454 pyrosequencing analysis instructed by manufacturer's protocol.

HT-SuperSAGE

The purified PCR product from each sample is quantified by an Agilent Bioanalyzer system.

8.1 A DNA chip from Agilent DNA 1000 kit is prepared and filled with Gel-Dye Mix supplied with the kit.
8.2 Load 1 µl purified PCR product in the well of the chip and run the chip in the Agilent 2100 Bioanalyzer.
8.3 The DNA concentration of the 123- to 125-bp fragment is measured using 2100 Expert software (Agilent Technologies).
8.4 Based on this quantification, an equal amount of DNA (PCR product) from each sample is mixed and the mixture sequenced on an Illumina GA.
8.5 For the sequencing reaction, GEX sequencing primer (5′-CGACAGGTTCAGAGTTCTACAGTCCGACGATC) should be employed.

1.4 Applications

DeepSuperSAGE recommends itself for whole-genome transcriptome studies of any eukaryotic organism. It has already been employed as a transcriptome analysis tool in various studies, particularly of nonmodel organisms without sequenced genomes (banana, chickpea, pea, lentil, *Boechera*, etc.). The high quality of data produced, the relatively simple procedure in combination with one of the NGS platforms, and the lower costs for a transcriptome analysis as compared to, for example, a complete microarray experiment will promote its applications in future.

1.4.1 Applications of DeepSuperSAGE in Combination with 454 Pyrosequencing

DeepSuperSAGE reveals many facets of the transcriptome reacting upon abiotic or biotic stresses or deciphers the changing involvement of transcription and transcripts during development of any organism (Table 1.2).

Particularly in higher plants, deepSuperSAGE has shown its resolving power as a transcriptome analysis tool. However, genome sequences of most plants are either incomplete or untouched, regardless of their economic (mostly agricultural) importance. As described above, genes can be recovered from deepSuperSAGE tag sequences by RACE without searching databases. However, the most recent advances of NGS technologies now allow us to construct a substantial EST database by just

Table 1.2 Various published applications of deepSuperSAGE.

Author	Species	Sequencing technology
Molina *et al.* [26]	*Cicer arietinum*	454
Sharbel *et al.* [28]	*Boechera* spp.	454
Sharbel *et al.* [29]	*Boechera* spp.	454
Gilardoni *et al.* [30]	*Nicotiana attenuata*	454
Yamaguchi *et al.* [27]	*Solanum tovum*	Illumina
Pinto *et al.* [31]	*Tetradon nigroviridis*	454
Matsumura *et al.* [23]	*Oryza sativa, Danio renio, Arabidopsis thaliana, Magnaporthe oryzae*	Illumina

sequencing cDNA fragments from the experimenter's own materials. In chickpea (*Cicer arietinum* L.) or *Boechera* species, for example, deepSuperSAGE tag sequences were BLASTed against public or newly sequenced cDNA databases for the identification of the corresponding genes [26,28,29]. Without preparing one's own cDNA databases, EST sequences from related species are also applicable as reference sequences to BLAST searches of the tags. To give only one example, tags from chickpea were BLASTed against *Medicago truncatula* ESTs [26]. Similarly, for annotation of *Nicotiana attenuata* and *Solanum torvum* tags, DNA sequences of *Nicotiana* species, *Solanum* species, or egg plant Unigenes were employed as databases for retrieval [27,30]. It is still an open question whether and to what extent sequences from genetically distant species are acceptable for tag-to-gene annotation via sequence similarity. Practically, however, the few examples described above demonstrate that corresponding cDNAs (genes) could be successfully identified this way.

DeepSuperSAGE additionally identifies unique classes of transcripts, which cannot be detected by microarrays, for example. In differentially expressed tags of drought-exposed chickpea roots, 170 tags matched EST sequences in the antisense polarity [26]. Therefore, the detection of antisense transcripts is a rewarding advantage of deepSuperSAGE. Although a further (functional and/or structural) analysis is still required for each tag (or transcript), deepSuperSAGE nevertheless discovers novel transcripts. Sharbel *et al.* [29] could identify allelic variation of transcripts from the same locus by analyzing deepSuperSAGE tags from apomictic and sexual ovules of *Boechera* species. The window of a SuperSAGE tag expands over only 26 bases and therefore identified transcript variants might be limited in numbers. However, the tag likely localizes to the 3′-untranslated region of cDNAs, which increases the chances to identify sequence variations. Combining information of alleles and their expression patterns has helped to better understand complex events in living organisms like apomixis [28,29].

One of the best examples of the power of deepSuperSAGE as a transcriptome profiling technology is the identification of rapidly up- and downregulated genes, the quantification of their transcripts, the discovery of many sense and antisense transcripts, the multitude of alternatively spliced transcript isoforms, and their contribution to the various salt stress-induced metabolic pathways, to name a few benefits of the technique. Within the focus of the corresponding experiments, two deepSuperSAGE libraries were developed from roots and nodules of the salt-tolerant chickpea variety INRAT-93. A moderate salt stress of 25 mM NaCl was chosen and the deepSuperSAGE transcript profiles established after only 2 h of salt stress. Sequencing of the tags was done by the 454 platform. Among the various results and insights into the first wave of salt stress-compensatory measures of chickpea roots, a compilation of the 40 top upregulated transcripts and their annotations is shown in Table 1.3. In parallel, the 40 top upregulated transcripts from nodules of the same plants are shown in Table 1.4.

These 40 transcripts were chosen among thousands of upregulated transcripts in both organs that were significantly, but less activated after onset of the salt stress (86 919 transcripts representing 17 918 unique 26-bp deepSuperSAGE tags, so-called UniTags,

Table 1.3 Top 40 annotatable and upregulated UniTags of roots from the salt-tolerant chickpea variety INRAT-93 under salt stress.

Tag ID	Associated gene annotation	R_{ln}	Associated process
STCa-18884	early nodulin 40	5.69	nodulation
STCa-7896	superoxide dismutase	3.70	ROS scavenging
STCa-318	trypsin protein inhibitor 3	3.59	endopeptidase inhibitor
STCa-19021	extensin	3.40	cell wall organization
STCa-17087	dormancy-associated protein	3.38	no associated process
STCa-7166	NADP-dependent isocitrate dehydrogenase I	3.25	metabolism
STCa-1381	acetyl-CoA synthetase	3.19	metabolism
STCa-2982	cysteine synthase	3.15	protein metabolism
STCa-15648	mitochondrial 24S mt-RNL ribosomal gene	3.10	no associated process
STCa-20215	putative extracellular dermal glycoprotein	3.08	proteolysis
STCa-20066	14-3-3-like protein A	3.03	protein domain-specific binding
STCa-15159	disease resistance protein DRRG49-C	2.98	response to stress
STCa-17434	AAD20160.1 protein	2.92	no associated process
STCa-22427	fiber protein Fb19	2.88	response to stress
STCa-4531	isoflavone 3′-hydroxylase	2.88	no associated process
STCa-14437	60S acidic ribosomal protein P1	2.83	protein biosynthesis
STCa-1385	1-aminocylopropane-1-carboxylate oxidase	2.83	metabolism
STCa-12309	ankyrin-like protein	2.83	no associated process
STCa-23197	hypothetical protein	2.78	response to stress
STCa-8459	UDP-glucose pyrophosphorylase	2.78	metabolism
STCa-12035	cytochrome P450 monooxygenase	2.73	electron transport/metal ion binding
STCa-11051	retinoblastoma-related protein	2.68	no associated process
STCa-7975	T5A14.10 protein	2.68	no associated process
STCa-14984	40S ribosomal protein S4	2.68	protein biosynthesis
STCa-21666	low-temperature salt-responsive protein LTI6B	2.68	Integral to membrane
STCa-1958	gibberellin-stimulated protein	2.68	Hormone response
STCa-17272	10-kDa photosystem II polypeptide	2.68	Oxygen evolving complex
STCa-24178	phosphoglycerate mutase	2.62	Metabolism/metal ion binding
STCa-13313	Chalcone isomerase	2.62	flavonoid biosynthesis
STCa-23978	inorganic pyrophosphatase-like protein	2.62	phosphate metabolism
STCa-10123	synaptobrevin-like protein	2.62	transport/integral to membrane
STCa-11172	caffeic acid 3-*O*-methyltransferase	2.56	lignin biosynthesis
STCa-181	myoinositol-1-phosphate synthase	2.56	inositol 3P biosynthesis/Ca^{2+} release
STCa-15340	alfin-1	2.56	regulation of transcription
STCa-24453	tonoplast intrinsic protein	2.56	transport
STCa-4528	cytochrome P450 monooxygenase	2.56	electron transport/metal ion binding
STCa-5543	ε-subunit of mitochondrial F1-ATPase	2.56	ATP-coupled proton transport
STCa-11309	60S ribosomal protein L18a	2.49	protein biosynthesis
STCa-16808	histone H2B	2.49	response to DNA damage stimulus
STCa-22470	glutathione S-transferase	2.49	ROS scavenging

Two deepSuperSAGE libraries derived from salt stressed- and nontreated chickpea roots, respectively, of the salt-tolerant variety INRAT-93 were developed. All 26-bp tags per library were grouped in classes sharing the same sequence (UniTags) and their counts were normalized to counts per million. After normalization, counts were compared between libraries and expression ratios were calculated for each UniTag (R_{ln}). Here, the 40 UniTags showing the largest expression ratios after salt stress induction (2 h 25 mM NaCl) are listed.

from roots, and 57 281 transcripts representing 13 115 UniTags from nodules of the same plants). The thousands of downregulated genes, the antisense transcripts, and their corresponding sense counterparts as well as the Gene Ontology (GO) terms for all of these various messages and their response to salt stress are completely ignored here. However, from a more detailed GO analysis we can infer that (i) transcripts associated with the generation and scavenging of reactive oxygen species (ROS), and (ii) transcripts involved in Na^+ homeostasis were over-represented in GO categories, to give only two examples. Both pathways undergo strong global transcriptome changes in chickpea roots and nodules already 2 h after onset of moderate salt stress. Additionally, a set of more than 15 candidate transcripts react as potential components of the salt-overly-sensitive (SOS) pathway in chickpea (Figure 1.3).

Some of the major insights into the first steps of salt stress response in chickpea are that (i) normal nodules already have elevated levels of transcripts encoding ROS

Table 1.4 Top 40 annotatable and up-regulated UniTags of nodules from the salt-tolerant chickpea variety INRAT-93 under salt stress.

Tag ID	Associated gene annotation	R_{ln}	Associated process
STCa-18884	early nodulin 40	4.11	nodulation
STCa-15648	24S mitochondrial ribosomal mt-RNL gene	3.17	translation
STCa-11090	40S ribosomal protein SA	2.73	protein biosynthesis
STCa-17434	AAD20160.1 protein	2.61	no associated term
STCa-1958	gibberellin-stimulated protein	2.61	no associated term
STCa-3760	cysteine proteinase inhibitor	2.48	inhibition of proteolysis
STCa-89	drought-induced protein	2.48	response to stress
STCa-16482	40S ribosomal protein S9-2	2.48	protein biosynthesis
STCa-10316	NtEIG-E80 protein	2.33	no associated term
STCa-3321	leghemoglobin	2.33	oxygen transport
STCa-1263	benzoyltransferase-like protein	2.33	no associated term
STCa-13055	nonspecific lipid-transfer protein precursor	2.33	transport (lipids)
STCa-22149	acyl carrier protein	2.33	lipid biosynthesis
STCa-10862	F6N18.8 protein	2.33	no associated term
STCa-21007	two-component response regulator PRR37	2.33	regulation of transcription
STCa-4833	T13M11_21 protein	2.14	regulation of transcription
STCa-8434	fiber protein Fb2	2.14	no associated term
STCa-23572	F7K24_140 protein	2.14	signal transduction
STCa-7572	protein phosphatase 2A	2.14	signal transduction
STCa-1895	GDP-mannose 3,5-epimerase	2.14	ascorbic acid biosynthesis
STCa-16007	aquaporin PIP-type 7a	1.92	transport (trans-membrane)
STCa-2175	glutathione S-transferase	1.92	ROS scavenging
STCa-12406	coatomer subunit β′-2	1.92	protein transport
STCa-12523	T23K23_9 protein	1.92	no associated term
STCa-269	phytochrome B	1.92	signal transduction
STCa-1589	β-galactosidase	1.92	metabolism (carbohydrates)
STCa-19649	vacuolar ATPase subunit A	1.92	ion transport
STCa-22041	root nodule extensin	1.92	cell wall organization
STCa-199	nodulin-like protein	1.92	transport (transmembrane)
STCa-542	prolyl 4-hydroxylase	1.92	ROS scavenging
STCa-13688	*O*-methyltransferase	1.92	lignin biosynthesis
STCa-15530	NADH ubiquinone oxidoreductase	1.92	electron transport
STCa-16514	NADH dehydrogenase	1.92	electron transport
STCa-22816	F17F16.27 protein	1.92	no associated term
STCa-4167	syringolide-induced protein	1.92	metabolism (carbohydrates)
STCa-2241	putative extensin	1.92	cell wall organization
STCa-319	trypsin protein inhibitor 3	1.92	inhibition of proteolysis
STCa-9781	eukaryotic translation initiation factor 3	1.92	protein biosynthesis
STCa-1461	HMG1 protein	1.92	regulation of transcription
STCa-13993	F8K7.2 protein	1.92	no associated term

Two SuperSAGE libraries derived from salt-stressed and nontreated chickpea nodules, respectively, of the salt-tolerant variety INRAT-93 were developed. All 26-bp tags per library were grouped in classes sharing the same sequence (UniTags) and their counts were normalized to counts per million. After normalization, counts were compared between libraries and expression ratios were calculated for each UniTag (R_{ln}). Here, the 40 UniTags showing the largest expression ratios after salt stress induction (2 h 25 mM NaCl) are listed.

scavengers prior to any salt treatment (i.e., are in a state of increased stress by ROS), and (ii) both nodules and roots rapidly (already 2 h after addition of 25 mM NaCl) respond to salt stress by transcription of genes encoding ROS scavengers. This rapid activation of genes in response to salt stress was unknown in nodulating legumes. We conclude that deepSuperSAGE expression profiling enriched our previously very limited knowledge of first reactions of a chickpea plant upon salt stress. We would like to point out that most of the data of the salt stress SuperSAGE experiments have not been evaluated yet. However, the two examples (although only superficially) presented here already show the potential of this next-generation transcriptome sequencing technology.

DeepSuperSAGE is the technique of choice for the identification of differentially expressed genes in any eukaryotic organism. Gilardoni *et al.* [30] systematically employed deepSuperSAGE from gene discovery to functional analysis of identified

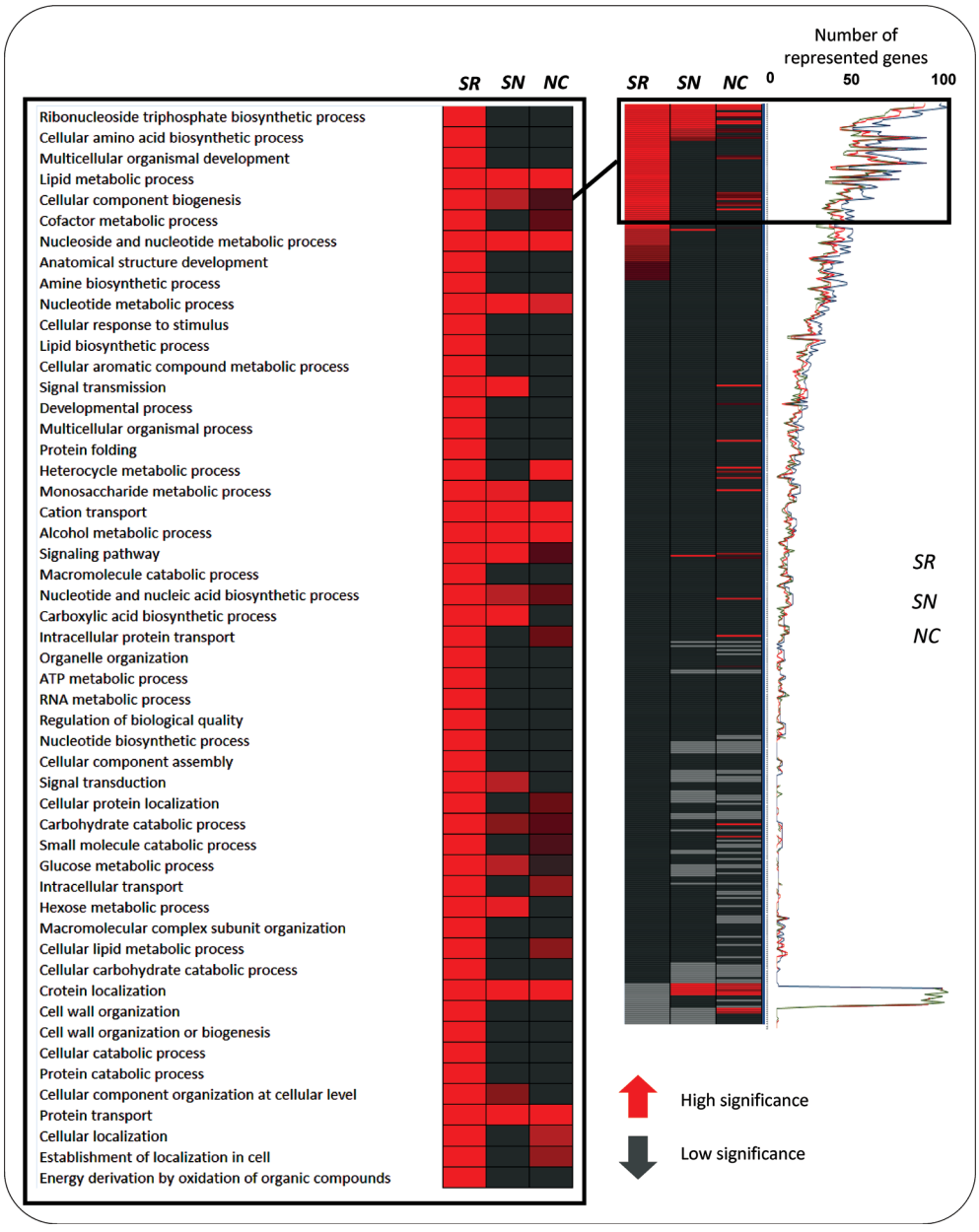

Fig. 1.3 Over-representation of more than 390 GO biological processes after salt stress induction, as calculated for chickpea roots and nodules using the software package ErmineJ. (Left panel) Heatmap of over-representation of GO biological processes in salt-stressed roots (*SR*) depicted in parallel with their over-representation levels in stressed nodules (*SN*) and nonstressed nodules in relation to roots of the same plants (*NC*). Numbers of represented genes per GO category for each case are shown by the curves right to the heatmap. (*Right panel*) Amplification of the heatmap region containing 52 high significance ($P < 1e\text{-}10$) over-represented GO terms in salt-stressed chickpea roots (*SR*). In parallel, the dynamics of the same processes in stressed nodules (*SN*) and nontreated nodules (*NC*) is shown.

genes in *N. attenuata*. Tools or resources for functional genomics are well developed for model species, like human, *Caenorhabditis elegans*, *Drosophila melanogaster*, or *Arabidopsis thaliana*. We exploited such tools by combining deepSuperSAGE and virus-induced gene silencing (VIGS), the later being a highly efficient tool for knocking-down target genes and measuring the resulting phenotype. Although VIGS is not applicable to any plant species, it nevertheless aided in the linking of a phenotype to a gene identified by deepSuperSAGE. It can be expected that the recent progress in RNA interference technology will support its application to a wide spectrum of species. Combining deepSuperSAGE and gene silencing technologies will, in our view, enrich our knowledge of the relationship between sequence and function.

1.4.2 Practical Analysis of HT-SuperSAGE

For the development of the described HT-SuperSAGE protocol, we designed 27 independently indexed adapters [23]. Additionally, cDNAs from two tissue samples

were digested with three different 4-bp cutter restriction endonucleases (anchoring enzymes) and tags were prepared. Amplified adapter–tag fragments from all 31 samples (27 indexed samples and additional four samples employing different anchoring enzymes) in total were pooled and sequenced in three lanes of a flow cell in an Illumina GAIIx sequencer (16 057 777 sequence reads of 35 bases) [23]. For tag extraction from sequence reads of pooled samples, our own programs were written in Perl script. Tag profiling data from all the applied samples was successfully separated and retrieved. As expected, contamination of tags from different samples was only less than 0.2% of the analyzed independent tags, even among index sequences with single-base differences.

Three benefits can be expected by pooling many samples in HT-SuperSAGE: (i) expansion of deepSuperSAGE applications, (ii) reduction of analytical cost per sample, and (iii) savings of starting material (RNA) from each sample. The analyses of biological replicates and expression kinetics were easy in HT-SuperSAGE and, additionally, a sufficient amount of tags can be prepared from 1 µg total RNA. Currently, with all the advances made, the performance and potential of HT-SuperSAGE is positively superior to microarray techniques, since it is based on an unprecedented ultra-high (deep) sequencing of tags, the digital printout of quantitative tag counts, and a high-throughput capacity.

DeepSuperSAGE can also employ different anchoring enzymes, of which *Nla*III is the standard enzyme in all the many versions of SAGE. However, as described by Sharbel *et al.* [29], cDNA is frequently not efficiently digested by *Nla*III, but instead by *Dpn*II, at least in certain species. Theoretically, any 4-bp cutter restriction endonuclease can be part of the deepSuperSAGE protocol and the change in the sequence of adapter ends is often welcomed. Actually, the frequency of sites for 4-bp cutter enzymes in cDNA is generally not consistent. Experimental results in *A. thaliana* show that *Nla*III or *Dpn*II digestion could recover tags from 92 to 93% of expressed genes, while *Bfa*I produced tags from about 80% of the cDNAs. Similar biases of restriction sites in the predicted genes were also reported by *in silico* scans of *D. melanogaster* and *C. elegans* genomes [32]. Since the restriction endonuclease *Bfa*I recognizes the sequence 5′-CTAG-3′, which includes a stop codon (TAG), this site may be less represented in cDNA sequences. However, the results demonstrate that *Nla*III or *Dpn*II are appropriate endonucleases for deepSuperSAGE, and most (above 99%) of the expressed genes could be monitored by these two enzymes.

1.5 Perspectives

NGS technologies are great innovations, and have revolutionized genomics and transcriptomics. The NGS platforms are continuously being improved and expanded, and new sequencing technologies are already being released or will be released in the near future, such as single-molecule sequencing from Pacific Biosciences [33]. This, and other next-next-generation sequencing technologies will read long fragments (above 1000 bp) in one path without amplification of the template DNA. However, the number of sequencing reads per run will be reduced compared to current massively parallel sequencing. Single-molecule sequencing will assist whole-genome analysis, even in *de novo* sequencing of genomes owing to efficient sequence assembling and less errors by PCR amplification. In transcriptomics, the new sequencing methods will be an effective tool to sequence cDNA directly and may allow us to read millions of full-length cDNA sequences accurately at a time. We expect that deepSuperSAGE in combination with massively parallel sequencing will remain advantageous even after the emergence of the next-next generation of sequencers. One of its merits is quantitative expression analysis, for which the number of sequence reads (tag counts) determines its accuracy and potential as a gene discovery tool. Moreover, multiplexing will assist in the measurement of gene expression of many different samples synchronously. Also, sequencing costs are still an issue and the costs for an RNA-

seq experiment still exceed the costs of a deepSuperSAGE experiment by a factor of 10. Therefore, the current deepSuperSAGE is still superior to single-molecule sequencing of cDNA or tags/tag concatemers. Instead, the immense accumulation of whole-genome and long cDNA sequences in the databases will greatly support the application of deepSuperSAGE in many aspects of eukaryotic biology.

Acknowledgments

H.M. is supported by the Program for the Promotion of Basic and Applied Researches for Innovations in Bio-oriented Industry (BRAIN). This work is also supported by JSPS grant 22380009. G.K. acknowledges research support by DFG (grant DFG 332/22-1) and GTZ (grant 08.7860.3-001.00). All proprietary names and registered tradenames for all materials, equipment, software, and so on, are acknowledged throughout this chapter.

References

1 Bustin, S.A., Benes, V., Garson, J.A., Hellemans, J., Huggett, J., Kubista, M., Mueller, R., Nolan, T., Pfaffl, M.W., Shipley, G.L., Vandesompele, J., and Wittwer, C.T. (2009) The MIQE Guidelines: minimum information for publication of quantitative real-time PCR experiments. *Clin. Chem.*, **55**, 611–622.

2 Bustin, S.A. (2010) Why the need for qPCR publication guidelines? The case for MIQUE. *Methods*, **50**, 217–226.

3 Bustin, S.A., Beaulieu, F.A., Huggett, J., Jaggi, R., Kibenge, F.S.B., Olsvik, P.A., Penning, L.C., and Toegel, S. (2010) MIQE precis: practical implementation of minimum standard guidelines for fluorescence-based quantitative real-time PCR experiments. *BMC Mol. Biol.*, **11**, 74.

4 Derveaux, S., Vandesompele, J., and Hellemans, J. (2010) How to do successful gene expression analysis using real-time PCR. *Methods*, **50**, 227–230.

5 Schena, M., Shalon, D., Davis, R.W., and Brown, P.Q. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, **270**, 467–470.

6 Shendure, J. (2008) The beginning of the end for microarrays? *Nat. Methods*, **5**, 585–587.

7 Irizarry, R.A., Warren, D., Spencer, F., Kim, I.F., Biswal, S., Frank, B.C., Gabrielson, E., Garcia, J.G.N., Geoghegan, J., Germino, G., Griffin, C., Hilmer, S.C., Hoffman, E., Jedlicka, A.E., Kawasaki, E., Martínez-Murillo, F., Morsberger, L., Lee, H., Petersen, D., Quackenbush, J., Scott, A., Wilson, M., Yang, Y., Qing Ye, S., and Yu, W. (2005) Multiple-laboratory comparison of microarray platforms. *Nat. Methods*, **2** 345–350.

8 Larkin, J.E., Frank, B.C., Gavras, H., Sultana, R., and Quackenbush, J. (2005) Independence and reproducibility across microarray platforms. *Nat. Methods*, **2**, 337–344.

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

10 Saha, S., Sparks, A.B., Rago, C., Akmaev, V., Wang, C.J., Vogelstein, B., Kinzler, K.W., and Velculescu, V.E. (2002) Using the transcriptome to annotate the genome. *Nat. Biotechnol.*, **20**, 508–512.

11 Metzker, M.L. (2010) Sequencing technologies – the next generation. *Nat. Rev. Genet.*, **11**, 31–46.

12 Matsumura, H., Reich, S., Ito, A., Saitoh, H., Kamoun, S., Winter, P., Kahl, G., Reuter, M., Krüger, D.H., and Terauchi, R. (2003) Gene expression analysis of host–pathogen interactions by SuperSAGE. *Proc. Natl. Acad. Sci. USA*, **100**, 15718–15723.

13 Meisel, A., Bickle, T.A., Kruger, D.H., and Schroeder, C. (1992) Type III restriction enzymes need two inversely oriented recognition sites for DNA cleavage. *Nature*, **355**, 467–469.

14 Moncke-Buchner, E., Rothenberg, M., Reich, S., Wagenführ, K., Matsumura, H., Terauchi, R., Kruger, D.H., and Reuter, M. (2009) Functional characterization and modulation of the DNA cleavage efficiency of Type III restriction endonuclease *Eco*P15I in its interaction with two sites in the DNA target. *J. Mol. Biol.*, **387**, 1309–1319.

15 Wagenfuhr, K., Pieper, S., Mackeldanz, P., Linscheid, M., Kruger, D.H., and Reuter, M. (2007) Structural domains in the Type III restriction endonuclease *Eco*P15I: characterization by limited proteolysis, mass spectrometry and insertional mutagenesis. *J. Mol. Biol.*, **366**, 93–102.

16 Matsumura, H., Ito, A., Saitoh, H., Winter, P., Kahl, G., Reuter, M., Kruger, D.H., and Terauchi, R. (2004) SuperSAGE. *Cell. Microbiol.*, **7**, 11–18.

17 Raftery, M.J., Moncke-Buchner, E., Matsumura, H., Giese, T., Winkelmann, A., Reuter, M., Terauchi, H., Schonrich, G., and Kruger, D.H. (2009) Unravelling the interaction of human cytomegalovirus with dendritic cells by using SuperSAGE. *J. Gen. Virol.*, **90**, 2221–2233.

18 Nasir, K.B.H., Takahashi, Y., Ito, A., Saitoh, H., Matsumura, H., Kanzaki, H., Shimizu, T., Ito, M., Sharma, P.C., Ohme-Takagi, M., Kamoun, S., and Terauchi, R. (2005) High-throughput in plant expression screening identifies a class II ethylene-responsive element binding factor-like protein that regulates plant cell death and non-host resistance. *Plant J.*, **43**, 491–505.

19 Matsumura, H., Bin Nasir, K.H., Yoshida, K., Ito, A., Kahl, G., Krüger, D.H., and Terauchi, R. (2006) SuperSAGE array: the direct use of 26-base-pair transcript tags in oligonucleotide arrays. *Nat. Methods*, **3**, 469–474.

20 Coemans, B., Matsumura, H., Terauchi, R., Remy, S., Swennen, R., and Sagi, L. (2005) SuperSAGE combined with PCR walking allows global gene expression profiling of banana (*Musa acuminata*), a non-model organism. *Theor. Appl. Genet.*, **111**, 1118–1126.

21 Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.J., Chen, Z., Dewell, S.B., Du, L., Fierro, J.M., Gomes, X.V., Godwin, B.C., He, W., Helgesen, S., Ho, C.H., Irzyk, G.P., Jandom, S.C., Alenquer, M.L., Jarvie, T.P., Jirage, K.B., Kim, J.B., Knight, R., Lanza, J.R., Leamon, J.H., Lefkowitz, S.M., Lei, M., Li, J., Lohman, K.L., Lu, H., Makhijani, V.B., McDade, K.E., McKenna, M.P., Myers, E.W., Nickerson, E., Nobile, J.R., Plant, R., Puc, B.P., Ronan, M.T., Roth, G.T., Sarkis, G.J., Simons, J.F., Simpson, J.W., Srinivasan, M., Tartaro, K.R., Tomasz, A.,

Vogt, K.A., Volkmer, G.A., Wang, S.H., Wang, Y., Weiner, M.P., Yu, P., Begley, R.F., and Rothberg, J.M. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437** 376–380.

22 Terauchi, R., Matsumura, H., Krüger, D.H., and Kahl, G. (2008) SuperSAGE: the most advanced transcriptome technology for functional genomics, in *Handbook of Plant Functional Genomics* (eds G. Kahl and K. Meksem), Wiley-VCH, Weinheim, pp. 37–54.

23 Matsumura F H., Yoshida, K., Luo, S., Kimura, E., Fujibe, T., Albertyn, Z., Barrero, RA., Krüger, D.H., Kahl, G., Schroth, G.P., and Terauchi, R. (2010) High-throughput SuperSAGE for digital gene expression analysis of multiple samples using next generation sequencing. *PLoS ONE*, **5**, e1201.

24 Nielsen, K.L., Høgh, A.L., and Emmersen, J. (2006) DeepSAGE – digital transcriptomics with high sensitivity, simple experimental protocol and multiplexing of samples. *Nucleic Acids Res.*, **34**, e133.

25 Marioni, J.C., Mason, C.E., Mane, S.M., Stephens, M., and Gilad, Y. (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.*, **18**, 1509–1517.

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

27 Yamaguchi, H., Fukuoka, H., Arao, T., Ohyama, A., Nunome, T., Miyatake, K., and Negoro, S. (2010) Gene expression analysis in cadmium-stressed roots of a low cadmium-accumulating solanaceous plant, *Solanum torvum*. *J. Exp. Bot.*, **61**, 423–437.

28 Sharbel, T.F., Voigt, M.L., Corral, J.M., Thiel, T., Varshney, A., Kumlehn, J., Vogel, H., and Rotter, B. (2009) Molecular signatures of apomictic and sexual ovules in the *Boechera holboellii* complex. *Plant J.*, **58**, 870–882.

29 Sharbel, T.F., Voigt, M.L., Corral, J.M., Galla, G., Kumlehn, J., Klukas, C., Schreiber, F., Vogel, H., and Rotter, B. (2010) Apomictic and sexual ovules of *Boechera* display heterochronic global gene expression patterns. *Plant Cell*, **22**, 655–671.

30 Gilardoni, P.A., Schuck, S., Jüngling, R., Rotter, B., Baldwin, I.T., and Bonaventure, G. (2010) SuperSAGE analysis of the *Nicotiana attenuata* transcriptome after fatty acid-amino acid elicitation (FAC): identification of early mediators of insect responses. *BMC Plant Biol.*, **10**, 66.

31 Pinto, P.I., Matsumura, H., Thorne, M.A., Power, D.M., Terauchi, R., Reinhardt, R., and Canário, A.V. (2010) Gill transcriptome response to changes in environmental calcium in the green spotted puffer fish. *BMC Genomics*, **11**, 476.

32 Pleasance, E.D., Marra, M.A., and Jones, S.J. (2003) Assessment of SAGE in transcript identification. *Genome Res.*, **6**, 1203–1215.

33 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., Kuse, R., Lacroix, Y., Lin, S., Lundquist, P., Ma, 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., and Turner, S. (2009) Real-time DNA sequencing from single polymerase molecules. *Science*, **323**, 133–138.