Plant Signaling & Behavior 5:11, 1353-1358; November 2010; © 2010 Landes Bioscience

Exploring molecular signaling in plant-fungal symbioses using high throughput RNA sequencing

Murray P. Cox,[1-3,*] Carla J. Eaton[1,4] and D. Barry Scott[1,2]

[1]Institute of Molecular BioSciences; Massey University Palmerston North; [2]The Bio-Protection Research Centre; [3]The Allan Wilson Centre for Molecular Ecology and Evolution; New Zealand; [4]Department of Plant Pathology and Microbiology; University of California Riverside; USA

Key words: symbiosis, second generation sequencing, RNA-seq

Plant-fungal symbioses are a common feature in nature. They vary from pathogenic interactions, where fungi subvert plant resources for their own use, to mutualistic associations, where both fungus and host benefit from the interaction. Although the ecological importance of plant-fungal symbioses has long been recognized and the biology of several key associations are now well studied, new technologies have the potential to allow fresh insight into the molecular basis of plant-fungal interactions. One such technique—high throughput RNA sequencing—has recently been used to explore the molecular basis of cross-species communications. Here, we give a brief overview of this emerging technology, and present a general guide for employing the methodology to dissect plant-fungal symbiosis.

Plant-fungal associations play an important role in terrestrial ecosystems. These interactions range from detrimental (i.e., antagonistic) associations to mutualistic (i.e., beneficial) associations. Plant-fungal signaling plays an important role across this interaction continuum. In antagonistic associations, fungi attempt to circumvent host defense responses, whereas in mutualistic states, both host and fungus must constantly communicate to maintain the tightly balanced symbiosis. Although this inter-species signaling is a vital function in many ecosystems, little is known about how it is controlled at the molecular level.

Studying cross-species signaling has traditionally proven difficult. The communication process is both spatially and temporally dynamic, and involves interactions between two very different and very complex biological systems. A wide range of approaches has been applied to understanding this signaling. One such technique is suppression subtractive hybridization (SSH),[1] which facilitates PCR amplification of cDNA fragments that are differentially expressed between control and experimental transcriptomes. Complementary DNA (cDNA) copies of RNA transcripts that have similar abundance in both samples preferentially hybridize to form double stranded DNA. These are then removed, leaving only differentially expressed sequences. In an alternative approach, serial analysis of gene expression (SAGE),[2] a short sequence tag is extracted from a unique position within each transcript. These tags are concatenated, sequenced and counted to quantify rates of gene expression. Yet by far the most dominant method has been the use of cDNA microarrays,[3] in which gene-specific oligonucleotide probes are spotted onto a solid surface. Labeled gene transcripts anneal to these probes, with the resulting signal being proportional to the number of transcripts bound. Using this method, considerable insight has been gained into the molecular mechanisms underlying symbiosis.

Microarrays have been widely applied to fungal systems.[4] The first fungal microarrays were developed in 1997 for yeast systems,[5,6] but were adapted for filamentous fungi from 2002 onwards.[7,8] Microarrays have been particularly usefully applied to studies of mycorrhizal symbioses, including analyses of gene expression changes in tomato (*Solanum lycopersicum*) over a time course of infection with the arbuscular mycorrhizal fungus *Glomus intraradices*.[9] This particular study revealed that host gene expression changes were noticeable even before fungal colonization was fully established. Laser microdissection coupled with microarray analysis has been used to identify the plant determinants of arbuscule development following colonization of tomato plants by *Glomus mosseae*.[10] Similarly, microarrays have been employed to examine changes in rice gene expression when exposed to mutant strains of *Magnaporthe oryzae mgb1* and *mst12*, which are unable to invade the plant host and cause infection.[11] Expression of phytoalexins was found to be much lower in plants inoculated with these mutant strains relative to wild type inoculations, thus suggesting that fungal invasion is required to trigger the phytoalexin response. Microarrays containing entire fungal gene sets were developed from 2005 onwards (some of the earliest were for *Aspergillus fumigatus* and *A. niger*[12,13]), thereby providing a single platform for studying changes in fungal gene expression under a wide range of experimental conditions.

Despite the importance of these studies, existing methods of surveying gene expression have significant limitations. Both SSH and SAGE are technically demanding; biased enrichment is a common outcome of the hybridization step in SSH, as well as the tag capture step in SAGE. Hence, microarrays have become

*Correspondence to: Murray P. Cox; Email: m.p.cox@massey.ac.nz
Submitted: 07/05/10; Accepted: 07/06/10
Previously published online:
www.landesbioscience.com/journals/psb/article/12950

the primary method for detecting differential gene expression, including complex symbiotic interactions. However, microarrays have their limitations too. The level of fluorescence when transcripts bind to microarray probes yields an essentially analogue signal. These signals are indistinguishable from noise at low rates of gene expression, but equally, signals saturate rapidly at high rates of gene expression. Therefore, microarrays are limited to parametric statistics that can capture the continuous nature (and associated noise) of these binding signals. More importantly for many plant-fungal symbioses, microarray construction is limited to organisms with a reasonable level of genome information, and even then, the choice of genes and probes that are placed on the microarray implicitly biases the study's outcome. Because microarray probes can only be designed and synthesized if gene sequences are known, microarray design is typically restricted to a small range of model organisms. Although hundreds of fungal genomes are now being sequenced and therefore are potentially being opened up for microarray studies, many other fungi—particularly from relatively understudied symbiotic systems—still lack published genome data. Even when fungal sequences are available, their host plant is likely to remain relatively uncharacterized. The complexity of plant genomes (e.g., their size and ploidy) means that plant genome sequencing is still in its infancy. In most cases, understanding the host half of a plant-fungal symbiosis with microarrays is not possible even if the fungus itself is relatively well known. Well-studied crop species (e.g., maize, rice and tomato) are key exceptions. Finally, due to the difficulty of normalizing plant versus fungal transcripts, studying both plant and fungal genes from the same sample is a demanding application when using microarrays. This precludes many study designs that might otherwise prove extremely useful, especially for symbiotic systems.

Second-generation Sequencing

Recently, a new tool has been added to the arsenal of techniques available to probe plant-fungus interactions. High throughput RNA sequencing[15] now provides a new approach to investigate plant-fungal signaling. Second-generation sequencing methods were first brought to commercial use when 454 pyrosequencing,[16] a technology sold by Roche, was launched in 2005. 454 sequencing currently generates around one million reads per run with read lengths of ~400 nucleotides. This is clearly a significant improvement on traditional first-generation Sanger sequencing, which produced orders of magnitude less data, even when automated via advanced robotic processing. Solexa sequencing technology, since acquired by Illumina, was released in 2006. Illumina adopted a sequencing-by-synthesis approach, which is designed to produce large numbers of relatively small reads. Illumina sequencing is currently capable of producing 125 nucleotide paired-end reads (i.e., two 125 nucleotide sequences linked by an unsequenced region of approximately known size), and generates around 300 million reads per run—two orders of magnitude more than 454 sequencing. SOLiD, released by ABI in 2007, uses a sequencing-by-ligation approach, and can currently generate around two billion 50 nucleotide paired-end reads per run—three orders of magnitude more than 454 sequencing. Extraordinarily, third-generation sequencing methods are expected to produce orders of magnitude more data even than this. The first third-generation sequencing machines are currently in the early stages of commercialization.

These new technologies have changed the playing field in molecular biology, and they offer real advantages for dissecting plant-fungal interactions. The power of these methods comes from the huge numbers of reads that they produce. This leads to several key advantages over existing techniques:

Deep read coverage. A wide range of genes can be detected, regardless of absolute expression level (e.g., potentially down to one transcript per cell).

Full gene coverage. Entire genes can be surveyed, not just small probe regions.

Discrete read counts. The number of reads that map to a gene is digital, which opens up an improved range of statistics.

Larger dynamic range. A gene expressing tens of reads can be detected equally well as a gene expressing hundreds of thousands of reads.

Genome sequences not required. EST generation and gene quantification can be performed on the same dataset, thus removing any need for a reference genome.

Joint plant-fungus analysis. Both plant and fungal transcriptomes can be analyzed simultaneously in the same sample.

Dissecting Symbiosis

By way of example, we recently applied high throughput RNA sequencing to a beneficial symbiosis between *Epichloë festucae* and perennial ryegrass (*Lolium perenne*) (**Fig. 1**). Epichloë/Neotyphodium species (epichloë endophytes) are constitutive symbionts of cool season grasses. Here, the fungus occupies a specialized niche, the aerial grass tissues, where it gains access to plant-derived nutrients and a means of dissemination through colonization of the host's seeds. In return, the fungus promotes host survival through improved nutrient acquisition and protection from mammalian and insect herbivory via the production of biologically active secondary metabolites. However, relatively little is known about how plant-fungal cross-species signaling coordinates this complex symbiotic state.

We explored the role of a fungal stress-activated MAP kinase (*sakA*) in maintaining this mutualistic association.[17] Deletion of *E. festucae sakA* induces a switch from a mutualistic to an antagonistic interaction with the host; infected plants exhibit dramatic changes in development, stunted growth and premature senescence. To determine the molecular basis of these changes, we undertook a high throughput RNA sequencing survey of gene expression in both the fungus and its plant host. From a molecular perspective, high throughput RNA sequencing revealed striking changes in fungal gene expression consistent with the transition from restricted to proliferative growth, including upregulation of hydrolytic enzymes and transporters, and downregulation of genes involved in the production of host protective metabolites. Corresponding analysis of the plant transcriptome revealed upregulation of host genes involved in pathogen defense, as well

as major changes to plant hormone-related gene expression and increased activation of transposases. Together, these results highlight the fine balance between mutualism and antagonism in this plant-fungal interaction.

The study also illustrates the power of deep RNA sequencing to dissect the molecular processes that underpin symbiosis. We believe that the same general approach can be applied to study plant-fungal communication in a range of symbiotic systems, although this approach is so new that clear experimental guidelines are currently lacking. Here, we present a relatively straightforward workflow that can be adapted to most plant-fungus interactions, covering the gamut from strict mutualism to rampant pathogenism.

High throughput RNA sequencing

Experimental design is the first, and most important, step. Consider a simple two-sample system: (i) a wild type fungus and (ii) another fungal strain (e.g., a natural isolate with some interesting property, or an experimental strain in which some gene of interest has been disrupted). In some instances, it may be possible to manipulate both the fungus and the plant genetically. In other cases, such as arbuscular mycorrhizal fungi—many of which cannot be cultured, let alone manipulated genetically—such studies may be limited to natural isolates or mutated host plants. Because we are interested in the plant-fungal interaction, both strains must necessarily be grown in planta. Much of the complexity in the following study design ultimately derives from this key requirement. First, the plant host is infected with a fungal strain, and a biological sample is collected when colonization has been established. The timing and anatomical location (e.g., meristem, leaf ligule, fruiting body) of this sample collection will vary depending on the question being asked. In most cases, biological replicates should be performed (i.e., multiple instances of the same strain grown in, and collected from, multiple plants). In practice, this strategy can often be difficult to implement. High throughput RNA sequencing requires a large starting quantity of total RNA, often on the order of 5–10 μg. If only a small portion of the plant is being screened (for instance, the leaf ligule), it will frequently be necessary to combine samples from multiple plants. There is no consensus yet about whether biological replicates should be used in this sort of pooled study design, or indeed, what such biological replicates would strictly mean. Nevertheless, in non-pooled study designs, we advise that projects should be planned with biological replicates wherever possible.

Looking towards the future, we anticipate that library construction methods will improve rapidly, especially by reducing the amount of RNA required for library preparation. These improvements will allow access to a wider range of sampling techniques, such as laser microdissection to target very small anatomical structures. Such techniques will reduce concerns around screening homogeneous versus heterogeneous tissue types; genes may be upregulated in one tissue type, but downregulated in adjacent cells. When taken to extremes, it may even be preferable to undertake high throughput RNA analysis on single cells, thus circumventing the need to average gene expression levels across

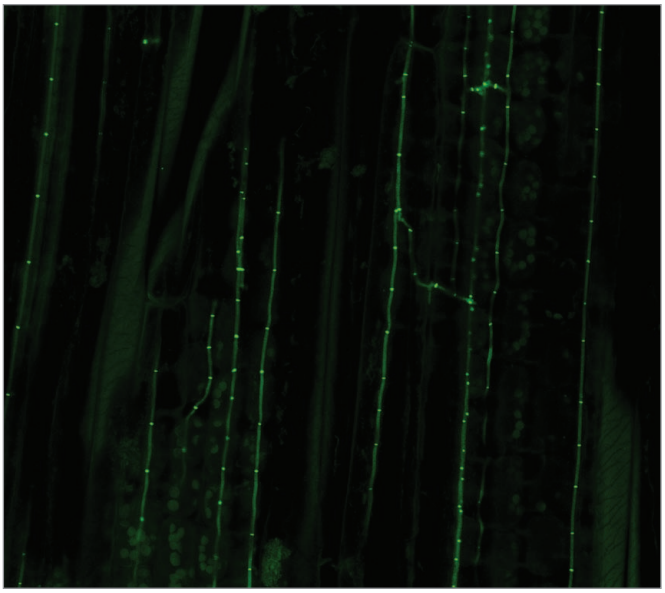

Figure 1. Confocal depth series image (4 μm) of a longitudinal section through a perennial ryegrass (*Lolium perenne*) leaf stained with Alexafluor (WGA-AF488) and aniline blue. The image shows hyphae (fluorescent green) of *Epichloë festucae* growing in close association with plant cells. Strongly illuminated points indicate hyphal septa.

thousands or even millions of cells. High throughput mRNA analysis of a single cell has already been accomplished,[18] although such extreme applications of the technology are not currently considered routine.

Once biological samples are collected, total RNA can be extracted using any standard method. A key proviso is that the final results of high throughput sequencing rely heavily on the quality of this starting material. It is worth spending considerable time optimizing the best method for extracting very high quality RNA (e.g., minimal fragmentation and contaminants). It is also advisable to screen for at least one fungal gene with RT-PCR; high throughput RNA sequencing of an uninfected plant host would be a costly mistake. For most researchers, the samples will then be sent to an external service provider. Given the intricate nature of second-generation sequencing and the relative infrequency with which most individual labs perform it, outsourcing is often the most cost effective approach. Second-generation sequencing services are increasingly being offered by local university core facilities, but a number of commercial providers are also available. All of the main sequencing companies (Roche, Illumina and ABI) have commercial sequencing arms as well.

Depending on the exact biological questions being asked, the total RNA is converted to cDNA and treated either to remove rRNA (i.e., ribo-depletion) or enrich for mRNA (i.e., poly-A selection). Poly-A selection is most common for studies that focus on gene expression, while ribo-depletion is preferred if the study aims to examine non-coding RNAs (i.e., not only poly-A tailed mRNAs). The important biological roles played by non-coding RNAs are only just starting to be recognized.[19] This is a rapidly growing field and many of the techniques described below can

be generalized to it. However, for brevity, we focus here on quantifying expression of protein coding genes. To build sequencing libraries, the cDNA is randomly sheared, platform-specific adapters are added, and the resulting sample undergoes clonal amplification. The exact details of these procedures vary widely between platforms, but are discussed in greater detail elsewhere.[20]

The choice of sequencing platform is an important consideration. For high throughput RNA sequencing, the emphasis is firmly on the total number of reads. Each read provides a gene count, and obtaining large numbers of reads (i.e., millions to billions) generates the power of this approach. Therefore, Illumina and SOLiD sequencing are preferred over the (relatively) low yields obtained with 454. We further note that SOLiD sequencing gains much of its power by leveraging off a known genome sequence. SOLiD is therefore most useful for transcriptome studies of model organisms, while Illumina sequencing is typically adopted for transcriptome studies of non-model organisms. This category includes most species involved in plant-fungal symbioses.

Bioinformatics

A key trend in second-generation sequencing is that much of the technical work is transferred from the laboratory bench to the computer screen. Indeed, for most researchers, bench work will comprise a relatively small part of the overall project (often less than ten percent). Conversely, computational analyses are becoming increasingly complex. Because the Illumina and SOLiD technologies generate so many sequence reads, their output files are usually many gigabytes in size. Analyzing such large datasets requires good computer skills, usually including knowledge of the UNIX operating system, and at least in the case of non-model organisms, basic programming skills as well. At present, users must be familiar with the command line interface. Although several commercial software suppliers have point-and-click products on the market, these currently do not compare well with open-source command lines alternatives. For most biologists, these computational requirements will necessitate collaboration with colleagues who can provide substantial bioinformatic support.

As with traditional Sanger sequencing, data quality is a vital consideration. Although these new sequencing technologies generate vast numbers of reads, not all reads are usable. Further, the error profiles of second-generation sequencing technologies differ dramatically from Sanger sequences. 454 is prone to large numbers of erroneous insertion/deletion events, while Illumina and SOLiD sequences show point mutation error rates that increase rapidly along the read length. In general, poor quality reads should be removed from further analysis, and the ends of reads with poor quality (e.g., Phred scores <30) should be trimmed. Dynamic trimming (i.e., on a read-by-read basis) is preferable over more common static trimming, where all reads are trimmed to the same length. This quality control step can easily lead to millions of reads being discarded, but this is of little concern when starting from tens of millions of reads. However, it should still be borne in mind that read quantity dictates the depth of the transcriptome analysis (i.e., genes with very low expression can only be quantified accurately with a sufficiently large number of total reads).

Once cleaned reads are available, these are mapped to a set of gene sequences. For now, we assume that a genome sequence is already available for your organism. (This was true in our Epichloë study, although we discuss alternative approaches below). The reads must be mapped to gene sequences from this genome, a process traditionally performed using blast-like strategies. However, blast comparisons, which take at least a few seconds per read, are too slow for mapping millions of reads. Instead, a new technique based on the Burrows-Wheeler transform has been developed.[21] This method can map reads to a reference gene set orders of magnitude faster than blast-like strategies. A range of free software, which tends to grow and change frequently, is available to perform this task. At present, two popular programs are BWA[22] and Bowtie.[23] Subsequently, the number of reads that map to each gene must be counted. To our knowledge, no software that performs this task has been publicly released. This is just one instance where collaboration with a computational biologist will be necessary. The counting process is repeated for both the wild type and experimental transcriptome, thus providing a list of reads that map to each gene for both the experimental case and the wild type control.

The next step is to apply appropriate statistics to determine which genes are differentially expressed between the wild type and experimental transcriptome. This task is extremely important, but in our experience, is often performed poorly by many biology-oriented research groups. At this stage of the project, we strongly advise collaborating with researchers who have strong statistical skills. The general goal is to show whether expression levels (i.e., the number of reads mapping to a given gene) differ between two samples. A range of statistical tests are commonly employed for this task, including Fisher's exact test[24] and the likelihood ratio test.[25] New methods designed specifically for high throughput RNA sequencing data have also been developed.[26] It is important to correct for multiple tests—statistics are typically run for each gene, which for most organisms can total in the tens of thousands. Large numbers of false positives will result if no correction is applied to individual significance values. As with microarray data, the currently favored approach is to apply a false discovery rate (FDR) control.[27] Another concern is that the total number of reads will usually differ between the two samples. Thus, read counts must be normalized for each sample when calculating fold changes in gene expression. This is typically performed via simple division with the total number of reads. We do not advise normalizing on housekeeping gene expression as, in our experience, these genes actually change expression under many experimental treatments. Importantly, none of these statistics need be developed from scratch. A range of packages in the R statistical language are already available to perform these manipulations, including DEGseq[26] and edgeR.[28] New statistical approaches are also continually being developed.[29]

We note that the large number of reads obtained in high throughput RNA sequencing experiments can lead to an issue that is typically unusual in biology—'too much' statistical

power. As the number of reads increases into the millions, a gene with 10,000 versus 10,100 mapped reads (i.e., a 1% change) in two samples, respectively, can become statistically significant. However, *statistical* significance is not necessarily the same thing as *biological* significance. One approach commonly employed at present is to only further analyze those genes that change by a predefined proportion (say, doubled or halved). However, as sequencing technologies improve and read counts inevitably increase, this issue of biological versus statistical significance will need to be addressed in a less *ad hoc* way by the research community.

We also recognize that a genome sequence (and hence, a set of gene sequences) may not be available for many plant-fungus symbiotic partners. In this case, researchers will need to develop their own library of gene sequences (or rather expressed sequence tags, ESTs). Conveniently, such libraries can be generated from the same read data that is used for gene mapping. Again, the sheer number of sequences means that traditional assembly methods, such as those used to assemble the human genome, are no longer computationally feasible. However, next generation de novo read assemblers based on de Bruijn graph theory[30,31] have recently become available. Two popular assemblers are ABySS[32,33] and Velvet.[34] De novo assembly is a complex process, and is described in more detail elsewhere.[35] De novo assembly produces a list of (partial) gene sequences that can subsequently be employed in the map-and-count strategy described above. Note that EST libraries produced in this manner seldom represent a complete set of gene sequences; they are usually fragmentary and will often be biased towards genes with moderate to high expression. However, where full genome sequences are unavailable (and this is still a common occurrence today), de novo assembly of EST libraries provides a feasible and powerful alternative strategy.

Finally, the list of differentially expressed genes identified using high throughput RNA sequencing must be annotated and interpreted. This is arguably the most important part of the entire project, but the exact approaches used will vary as widely as the research questions being asked. A key tactic for gene characterization is to blast genes (or ESTs) to public reference databases, such as GenBank, which is hosted by the US National Center for Biotechnology Information. Because DNA sequences diverge rapidly, gene sequences are usually transformed to six-frame conceptual translation products and compared against non-redundant protein sequence databases (i.e., blastx). Possible functional classes can be determined manually, or by assigning genes to gene ontology (GO) categories. This process can even be automated using web-aware software, such as Blast2GO.[36] Ultimately, however, downstream analyses still rely heavily on time-intensive, manual interpretation by biologically skilled staff. In our experience, there is still no substitute for advanced biological knowledge at this stage of the project.

One last advantage of high throughput RNA sequencing is that the data can be used for many purposes besides determining whether a gene is expressed, calculating its expression rate or determining expression differences under different experimental conditions. The same read data are equally informative about promoter start sites, can distinguish splice variants and single nucleotide polymorphisms, and can be used to verify computationally annotated genes. As long as partial genome sequences are available, co-regulated gene clusters can also be predicted.[17] Statistically significant strings of contiguous up- and downregulated genes may indicate co-regulated regions of the genome. For instance, in our Epichloë-Lolium study, genes in the fungal pathways that produce bio-protective secondary metabolites are subject to a relatively small number of expression controls. As should now be clear, high throughput sequence data can frequently be applied to a wide range of subsidiary questions besides quantifying levels of gene expression.

Conclusions and Perspective

High throughput RNA sequencing is a very general approach that can be adapted to a wide range of study designs and research questions. Because its experimental components are commonly outsourced and the bioinformatic tools are largely already available, even small research groups can readily adopt this new approach to explore the molecular basis of plant-fungal symbiosis. Indeed, this field is growing rapidly. In future, key advances are likely to derive from combining traditional genetic approaches (such as gene 'knockout' mutants) with the power of high throughput sequencing techniques.[37] With some imagination and good bioinformatic support, we expect that high throughput RNA sequencing will drive the next generation of breakthroughs in studies of gene expression, particularly with regard to understanding the molecular signals underlying plant-fungal symbioses.

Acknowledgements

This work was partly supported by the Bio-Protection Research Centre, the Allan Wilson Centre for Molecular Ecology and Evolution, and the Institute of Molecular BioSciences (IMBS), Massey University Palmerston North, New Zealand. We thank Emma Brasell (IMBS) and Dmitry Sokolov (Manawatu Microscopy and Imaging Centre, IMBS) for preparation of the illustration.

References

1. Diatchenko L, Lau YF, Campbell AP, Chenchik A, Moqadam F, Huang B, et al. Suppression subtractive hybridization: A method for generating differentially regulated or tissue-specific cDNA probes and libraries. Proc Natl Acad Sci USA 1996; 93:6025-30.
2. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. Serial analysis of gene expression. Science 1995; 270:484-7.
3. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 1995; 270:467-70.
4. Breakspear A, Momany M. The first fifty microarray studies in filamentous fungi. Microbiology 2007; 153:7-15.
5. DeRisi JL, Iyer VR, Brown PO. Exploring the metabolic and genetic control of gene expression on a genomic scale. Science 1997; 278:680-6.
6. Lashkari DA, DeRisi JL, McCusker JH, Namath AF, Gentile C, Hwang SY, et al. Yeast microarrays for genome wide parallel genetic and gene expression analysis. Proc Natl Acad Sci USA 1997; 94:13057-62.
7. Chambergo FS, Bonaccorsi ED, Ferreira AJ, Ramos AS, Ferreira Júnior JR, Abrahão-Neto J, et al. Elucidation of the metabolic fate of glucose in the filamentous fungus *Trichoderma reesei* using expressed sequence tag (EST) analysis and cDNA microarrays. J Biol Chem 2002; 277:13983-8.
8. Lewis ZA, Correa A, Schwerdtfeger C, Link KL, Xie X, Gomer RH, et al. Overexpression of White Collar-1 (WC-1) activates circadian clock-associated genes, but is not sufficient to induce most light-regulated gene expression in *Neurospora crassa*. Mol Microbiol 2002; 45:917-31.
9. Dermatsev V, Weingarten-Baror C, Resnick N, Gadkar V, Wininger S, Kolotilin I, et al. Microarray analysis and functional tests suggest the involvement of expansins in the early stages of symbiosis of the arbuscular mycorrhizal fungus *Glomus intraradices* on tomato (*Solanum lycopersicum*). Mol Plant Pathol 2010; 11:121-35.
10. Fiorilli V, Catoni M, Miozzi L, Novero M, Accotto GP, Lanfranco L. Global and cell-type gene expression profiles in tomato plants colonized by an arbuscular mycorrhizal fungus. New Phytol 2009; 184:975-87.
11. Kato T, Tanabe S, Nishimura M, Ohtake Y, Nishizawa Y, Shimizu T, et al. Differential responses of rice to inoculation with wild-type and non-pathogenic mutants of *Magnaporthe oryzae*. Plant Mol Biol 2009; 70:617-25.
12. MacKenzie DA, Guillemette T, Al-Sheikh H, Watson AJ, Jeenes DJ, Wongwathanarat P, et al. UPR-independent dithiothreitol stress-induced genes in *Aspergillus niger*. Mol Genet Genomics 2005; 274:410-8.
13. Nierman WC, Pain A, Anderson MJ, Wortman JR, Kim HS, Arroyo J, et al. Genomic sequence of the pathogenic and allergenic filamentous fungus *Aspergillus fumigatus*. Nature 2005; 438:1151-6.
14. Guenther JC, Hallen-Adams HE, Bücking H, Shachar-Hill Y, Trail F. Triacylglyceride metabolism by *Fusarium graminearum* during colonization and sexual development on wheat. Mol Plant Microbe Interact 2009; 22:1492-503.
15. Wang Z, Gerstein M, Snyder M. RNA-Seq: A revolutionary tool for transcriptomics. Nat Rev Genet 2009; 10:57-63.
16. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature 2005; 437:376-80.
17. Eaton CJ, Cox MP, Ambrose B, Becker M, Hesse U, Schardl CL, et al. Disruption of signaling in a fungal-grass symbiosis leads to pathogenesis. Plant Physiol 2010; 153:1780-94.
18. Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, et al. mRNA-seq whole-transcriptome analysis of a single cell. Nat Meth 2009; 6:377-82.
19. Mattick JS. The genetic signatures of noncoding RNAs. PLoS Genet 2009; 5:1000459.
20. Metzker ML. Sequencing technologies—the next generation. Nat Rev Genet 2010; 11:31-46.
21. Burrows M, Wheeler D. A block sorting lossless data compression algorithm. Technical Report 124, Digital Equipment Corporation 1994.
22. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics 2009; 25:1754-60.
23. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 2009; 10:25.
24. Fisher RA. On the interpretation of χ^2 from contingency tables, and the calculation of P. J R Statist Soc 1922; 85:87-94.
25. Neyman J, Pearson E. On the use and interpretation of certain test criteria for purposes of statistical inference, Part I. Biometrika 1928; 20:175-240.
26. Wang L, Feng Z, Wang X, Wang X, Zhang X. DEGseq: An R package for identifying differentially expressed genes from RNA-seq data. Bioinformatics 2010; 26:136-8.
27. Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. J R Statist Soc B 1995; 57:289-300.
28. Robinson MD, McCarthy DJ, Smyth GK. EdgeR: A Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 2010; 26:139-40.
29. Robinson M, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol 2010; 11:25.
30. de Bruijn NG. A combinatorial problem. Koninklijke Nederlandse Akademie van Wetenschappen 1946; 49:758-64.
31. Good IJ. Normal recurring decimals. J London Math Soc 1946; 21:167-9.
32. Birol I, Jackman SD, Nielsen CB, Qian JQ, Varhol R, Stazyk G, et al. De novo transcriptome assembly with ABySS. Bioinformatics 2009; 25:2872-7.
33. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I. ABySS: A parallel assembler for short read sequence data. Genome Res 2009; 19:1117-23.
34. Zerbino DR, Birney E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 2008; 18:821-9.
35. Miller JR, Koren S, Sutton G. Assembly algorithms for next-generation sequencing data. Genomics 2010; 95:315-27.
36. Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M. Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 2005; 21:3674-6.
37. Hawkins RD, Hon GC, Ren B. Next-generation genomics: An integrative approach. Nat Rev Genet 2010; 11:476-86.