MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2013) **13**, 559–572 doi: 10.1111/1755-0998.12109

PRIMER

Principles of transcriptome analysis and gene expression quantification: an RNA-seq tutorial

JOCHEN B. W. WOLF*†
*Department of Evolutionary Biology, Uppsala University, Uppsala, Sweden, †Science of Life Laboratory, Uppsala, Sweden

Abstract

Genome-wide analyses and high-throughput screening was long reserved for biomedical applications and genetic model organisms. With the rapid development of massively parallel sequencing nanotechnology (or next-generation sequencing) and simultaneous maturation of bioinformatic tools, this situation has dramatically changed. Genome-wide thinking is forging its way into disciplines like evolutionary biology or molecular ecology that were historically confined to small-scale genetic approaches. Accessibility to genome-scale information is transforming these fields, as it allows us to answer long-standing questions like the genetic basis of local adaptation and speciation or the evolution of gene expression profiles that until recently were out of reach. Many in the eco-evolutionary sciences will be working with large-scale genomic data sets, and a basic understanding of the concepts and underlying methods is necessary to judge the work of others. Here, I briefly introduce next-generation sequencing and then focus on transcriptome shotgun sequencing (RNA-seq). This article gives a broad overview and provides practical guidance for the many steps involved in a typical RNA-seq work flow from sampling, to RNA extraction, library preparation and data analysis. I focus on principles, present useful tools where appropriate and point out where caution is needed or progress to be expected. This tutorial is mostly targeted at beginners, but also contains potentially useful reflections for the more experienced.

Keywords: bioinformatics, next-generation sequencing, shotgun sequencing, transcriptome sequencing

Received 24 January 2013; revision received 18 March 2013; accepted 21 March 2013

Introduction

Only a decade ago, the study of gene expression was reserved to the realm of human medical genetics or genetic model systems like the mouse, fruit fly and nematodes. For these systems, microarrays and serial analyses of gene expression were the only tools available for examining features of the transcriptome and global patterns of gene expression. For eco-evolutionary model species, this important layer of biological information between genotype and phenotype was simply not accessible. Gene expression studies were restricted to small-scale quantitative PCR analyses of candidate genes or relied on cross-species hybridization on microarrays (Naurin *et al*. 2008). With the rapid development of massively parallel sequencing (or next-generation sequencing) (Margulies *et al*. 2005) and the maturation of analytical tools during the last few years, the situation has changed dramatically. Whole-genome or whole-transcriptome analyses have become a realistic option for genetic nonmodel organisms, even for individual laboratories (Ellegren *et al*. 2012; Lamichhaney *et al*. 2012), and will soon be standard practice in molecular ecological studies.

This article is not meant to be an exhaustive review of the latest developments in sequencing technology, specific downstream analyses or available software packages, nor a comprehensive summary of past applications in the ecological sciences. Up-to-date reviews exist for most of these aspects (Quick links in Box 1). Rather, I intend to give a broad overview and provide practical guidance for the many steps involved during a typical RNA-seq work flow (Fig. 1).

RNA-seq: applications

Before going into technical detail, I will briefly highlight the potential of RNA-seq within the ecological and evolutionary sciences. One of the most basic and still common applications of the method is the mere characterization of a species' transcriptome. While this is descriptive and generates little biological insight, it is often an important first step and constitutes a valuable

Correspondence: Jochen B. W. Wolf, Fax: 0046184716310;
E-mail: jochen.wolf@ebc.uu.se

Box 1 Quick links to useful entry points to the field

Overview on high-throughput sequencing and RNA-seq

1 *Principles of high throughput sequencing technology:* (Metzker 2010)
2 *Principles of RNA-seq:* (Wang *et al.* 2009; Oshlack *et al.* 2010; Ozsolak & Milos 2011)
3 *Principles of transcriptome assembly* (Martin & Wang 2011) *with particular reference to plants* (Jain 2012)

Applications in ecology and evolutionary biology

1 *General next-generation sequencing applications including RNA-seq:* (Ekblom & Galindo 2011)
2 *Special issues on next-generation sequencing including RNA-seq:* (Stapley *et al.* 2010; Tautz *et al.* 2010; Orsini *et al.* 2013)

Practical guidance and examples for useful tools

1 *Review on computational methods and tools:* (Pepke *et al.* 2009; Magi *et al.* 2010; Bao *et al.* 2011; Garber *et al.* 2011; Lee *et al.* 2012)
2 *Guidance in the design and analysis of RNA-seq experiments:* (De Wit *et al.* 2012; Vijay *et al.* 2013)
3 *Statistical consideration for RNA-seq data:* (Bullard *et al.* 2010; Kvam *et al.* 2012)
4 *Preprocessing and quality control tools:* NGSQCtoolkit *(Patel & Jain 2012),* fastQCtoolkit (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/),
5 *Mapping tools:* (Trapnell & Salzberg 2009; Bao *et al.* 2011)
6 *Gene name assignment:* e.g. BLAST2GO, SATSUMA, SPINES (for details and references see Vijay *et al.* 2013)
7 *Data visualization tools:* e.g. MapView (Bao *et al.* 2009), IGV (Thorvaldsdóttir *et al.* 2013), Tablet (Milne *et al.* 2010)
8 *Utility suites saving own effort*: BEDtools (Quinlan & Hall 2010), SAMtools (Li *et al.* 2009)
9 *Variant calling and genotyping: review* (Nielsen *et al.* 2011), GATK (DePristo *et al.* 2011), freebayes (http://bioinformatics.bc.edu/marthlab/FreeBayes);
10 *Gene function:* Gene ontology *(Gene Ontology Consortium 2004)*, gene ontology tools http://neurolex.org/wiki/Category:Resource:Gene_Ontology_Tools
11 *Gene interaction pathways:* e.g. KeGG pathway (Ogata *et al.* 1999; http://www.genome.jp/kegg/pathway.html), STRING database (Szklarczyk *et al.* 2011; http://string-db.org/) (http://string-db.org/)
12 *Galaxy: a useful online platform to analyse RNA-seq data:* (Goecks *et al.* 2010)
13 *The Bioconductor package: (Gentleman et al. 2004)* www.bioconductor.org
14 *Differential expression software:* e.g. DESeq (Anders & Huber 2010), edgeR (Robinson *et al.* 2010), baySeq (Hardcastle & Kelly 2010), NOIseq (Tarazona *et al.* 2011)
15 *Alternative splicing software:* e.g. Cufflinks (Trapnell *et al.* 2012), DEXSeq (Anders *et al.* 2012), EBSeq (Leng N *et al.* 2013), MISO (Katz *et al.* 2010)

Where to find help?

When help is needed one can often build on the experience of an online community. Current examples of active fora are: www.seqanswers.org; http://www.molecularecologist.com/next-gen-fieldguide-2013/; http://www.rna-seqblog.com/; http://www.biostars.org/. Stanford's SimpleFool's Guide to RNAseq specifically targets an organismal biologist audience (http://sfg.stanford.edu/).

resource for further analyses. An advantage over other next-generation approaches that reduce the genome to a more manageable size like restriction-site-associated DNA tags (RAD: Baird *et al.* 2008), multiplexed-shotgun genotyping (MSG: Andolfatto *et al.* 2011) or genotyping-by-sequencing (GBS: Elshire *et al.* 2011) is that RNA-seq data are directly derived from functional genomic elements, mostly protein-coding genes. Transcriptome sequence thus constitutes a meaningful resource to develop a large number of popular molecular markers such as single-nucleotide polymorphisms and microsatellites. In situations where full sequencing cannot be afforded, but the application requires the use of many markers (e.g. genome scans), the transcriptome provides a useful functionally relevant subset of the genome. For example, Lamichhaney *et al.* (2012) recently suggested a cost-efficient method to infer population allele frequencies by mapping genome-wide sequencing data of pooled individuals onto a *de novo* assembled transcriptome backbone.

Sequence-based polymorphisms is not the only biologically relevant layer of segregating variation.

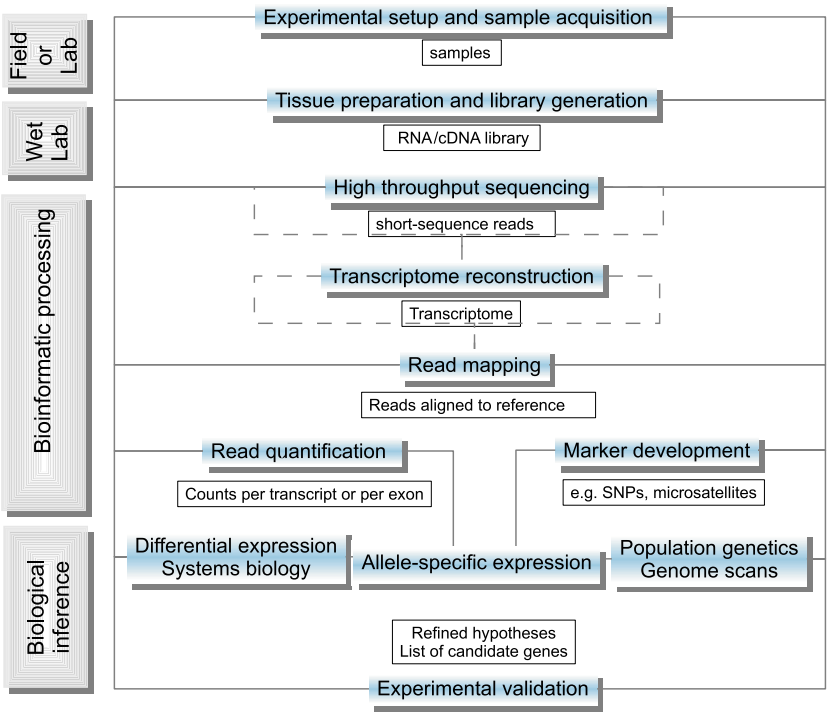

Fig. 1 Flow chart of a typical RNA-seq experiment.

The great advantage of RNA-seq data over other next-generation-sequencing applications is that it allows users to investigate differences in gene expression patterns between populations, for example in the context of speciation (Wolf *et al.* 2010) or eco-type-specific adaptation (Lenz *et al.* 2013). Gene regulatory variation need not be confined to gene expression levels. Pending further methodological development, we will see more studies quantifying segregation between transcript isoforms and quantification of their relative expression levels (Harr & Turner 2010). Simultaneous information on sequence variation at individuals' genomes and transcriptomes allows inferring patterns of allele-specific expression that can be relevant to environmental response and adaptation (Guo *et al.* 2004; Tirosh *et al.* 2009) and has yet to be examined in the wild. With the increasing ease of large-scale sequencing, the field of molecular ecology will expand its boundaries and merge with other disciplines such as phylogenetics, comparative genomics or systems biology, to their mutual benefit.

RNA-seq: the principle

RNA-seq, also called whole-transcriptome shotgun sequencing, refers to the use of high-throughput sequencing technologies (see below) for characterizing the RNA content and composition of a given sample. Due to technological limitations at present, sequence information from transcripts cannot be retrieved as a whole, but is randomly decomposed into short reads of up to several hundred base pairs (Fig. 2). In the absence of genome or transcriptome information, transcripts first need to be reconstructed from these reads (or read pairs), which is referred to as *de novo* assembly. In the case where transcript or genome information is readily available, reads can be directly aligned onto the reference. Further, counting the reads that fall onto a given transcript provides a digital measurement of transcript abundance, which serves as the starting point for biological inference (Fig. 1).

RNA-seq and microarrays

Until the advent of RNA-seq, microarrays were the standard tool for gene expression quantification. It is thus not surprising that the first RNA-seq studies in nonmodel organisms used transcriptome information obtained by sequencing from a single individual or a pool of individuals to construct microarrays for quantifying individual gene expression (Vera *et al.* 2008). Decreasing costs, increasing yields and improving bioinformatic data processing now make it possible to obtain both sequence information and a measure of gene expression for several individuals directly by sequencing (Wolf *et al.* 2010). Although both RNA-seq and microarrays are generally in good agreement when it comes to relative gene expression quantification (Nookaew *et al.* 2012), RNA-seq has clear advantages and will soon be the standard even for large

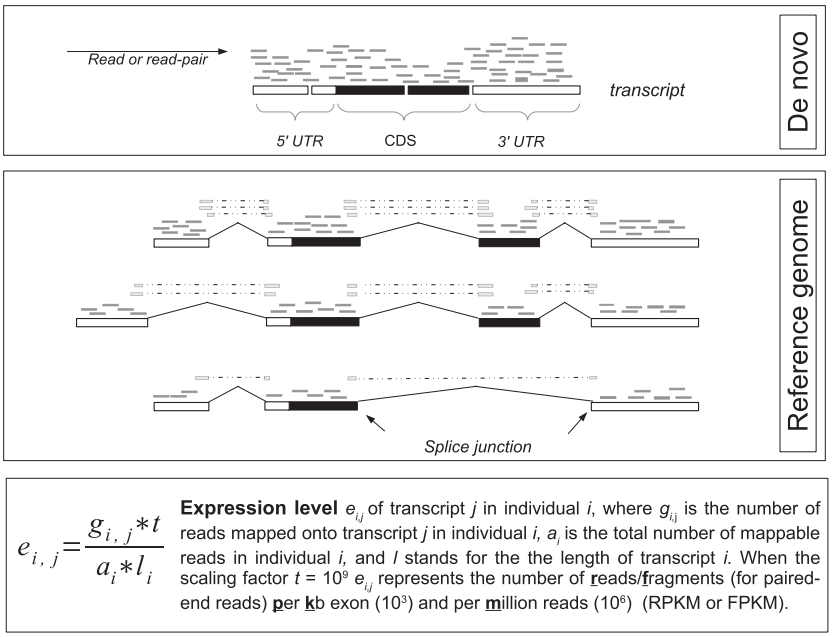

Fig. 2 Measuring gene expression using RNA-seq. (A) Processed mRNA transcripts can be inferred *de novo* without an existing genome by assembling short reads (grey bars) into contigs (long grey bars). In the best case, the longest contig represents one full-length transcript. Alternatively spliced isoforms (as shown in B) are difficult to infer and are either represented by the most common variant, merged or partially distributed into different contigs. (B) When a (distant) reference genome exists, reads can be directly mapped. Reads (or read pairs) that fall on splice junctions (dotted lines) are informative about alternatively splicing and allow the reconstruction of transcript isoforms. C) Basic expression measure. For both *de novo* and (spliced) mapping assemblies, gene expression levels can be inferred as read counts that are generally normalized by initial sequencing depth and transcript length. For further normalization options, see text. Black bars characterize coding sequence (CDS), and white bars indicate untranslated regions.

experiments. Compared with microarrays, RNA-seq at sufficient coverage captures a wider range of expression values. As a digital measure (count data), it scales linearly even at extreme values, whereas microarrays show saturation of analog-type fluorescent signals (Marioni *et al.* 2008). RNA-seq further provides information on RNA splice events; these are not readily detected by standard microarrays (Mortazavi *et al.* 2008). While still in its infancy in genetic nonmodel organisms, the segregation of alternative isoforms between populations or incipient species is an exciting area to explore with RNA-seq data. Another disadvantage of microarrays is its propensity for cross-hybridization to introduce biases in gene expression measurements. A comparable problem also exists for RNA-seq when reads align ambiguously, and I will discuss this in more detail in the next section.

While RNA-seq will most likely take the lead role in transcriptome analysis in the near future, one should not forget that RNA-seq data collection and statistical analysis are still under development. Before starting an RNA-seq experiment, one should thus bear in mind that RNA-seq data collection and analysis is more involved, and does not benefit from the decades of experience available for microarray analysis. Thus, microarrays should not be dismissed by default, and it is worth considering which application is best suited for addressing the question at hand before engaging in a large RNA-seq experiment.

RNA-seq: limitations

Before going through the individual steps of an RNA-seq experiment as outlined in Fig. 1, I would like to raise awareness of what can and cannot be accomplished in a typical RNA-seq experiment. A general consideration relates to the biological reality that will be captured. Implicitly, the quantity of interest is often not the intermediate mRNA, but the final protein products of the functional cell machinery. Methods are available for direct protein abundance estimation (Leskinen *et al.* 2012), and mRNA levels are mostly used as a proxy. However, when measuring steady-state mRNA levels, we are largely ignorant of mRNA stability or turnover rates; these rates eventually determine protein abundance. It is thus important to keep in mind that a gene's expression level alone can be a poor predictor of protein abundance (Vogel *et al.* 2010). Second, gene expression is highly tissue specific (Brawand *et al.* 2011), and even within tissues, we may only be interested in expression patterns of single-cell types generating a phenotype of

interest (e.g. melanocytes within hair follicles that are measured from expression in skin). Thus, caution is needed in the interpretation of gene expression patterns, as they often reflect expression of a heterogeneous mix of cell populations even in cases where tissues have been carefully selected. Third, mRNA is transcribed only from one template DNA strand. While protocols for strand-specific RNA-sequencing exist (Levin *et al.* 2010), standard RNA-seq experiments, at present, do not incorporate strand specificity. As a consequence, transcripts from overlapping genes encoded on different strands are indistinguishable and will be incorrectly lumped in the analysis. In particular, for the annotation of novel genome assemblies, strand-specific protocols should be considered, as they greatly improve gene annotation.

The experimental set-up

Purpose of the study

A natural starting point of every experiment is to define its specific goals and assess its feasibility with respect to the budget and available methodology. For an RNA-seq experiment, questions like the following should be considered as early as possible in the planning process, as they will have direct consequences on the experimental design and the analysis pipeline: Which aspects of the transcriptome am I most interested in, protein-coding mRNA or regulatory noncoding RNA? Are there other sequencing resources publicly available? Do I care about alternative splicing or is it enough to characterize broad scale expression patterns? Which sequence coverage will I need, and accordingly what sequencing technology suits my purpose best? Do I want to characterize the transcriptome (or compare expression) between treatment groups or populations? What is the statistical power for my sample to detect differential expression between genes with an x-fold expression difference at a given expression level?

Statistical design

Biological replication and statistical model selection is necessary to make inferences with some generality (Auer & Doerge 2010). While this is obvious and common practice in most fields of biological research, it is still worth mentioning here, as the statistical treatment of RNA-seq data has only recently shifted from single sample analyses to incorporate biological replication and allow for more complex statistical designs such as generalized linear regressions models (Hardcastle & Kelly 2010; Robinson *et al.* 2010). When designing the experimental set-up (e.g. blocked designs, interactions, Bayesian approaches), one should consider what the analysis tools can currently handle and find a compromise between experimental complexity and feasibility.

Common garden environment

Gene expression is notoriously plastic and highly sensitive to environmental conditions. While uncontrolled experiments from wild specimens can provide a valuable first step in hypothesis building (Wolf *et al.* 2010), firm conclusions on the underlying evolutionary dynamics of expression divergence can only be drawn under controlled experimental conditions.

The right choice of tissue, timing and study organism

Both transcript abundance and isoform identity are substantially different across tissues and change dramatically not only during embryological development, but throughout an individual's life (e.g. with reproductive status) and even the course of a day (circadian rhythms). It is therefore essential to consider in which tissue and at which physiological stage one is most likely to observe a difference relevant to the question at hand (e.g. ecological signal in juvenile, but not adult individuals: Nolte *et al.* 2009). For comparisons between groups, it is crucial to keep the variance within groups as low as possible while making sure that between-group differences do not arise from systematic differences in the sampling regime (e.g. different time of day, different physiological status) that will obscure the differences of interest (e.g. effect of population divergence on gene expression).

It strongly depends on one's organism what tissues can be selected. For small organisms like insects, single individuals may not yield sufficient RNA material for analysis and several individuals may have to be pooled. In these situations, building tissue-specific transcriptomes may be out of question. Likewise, when working with large animals, blood or skin are often the only accessible tissues. One should definitely premeditate which tissues, at which developmental time point and under which environmental condition need to be sampled before engaging in an RNA-seq experiment. Considering possible physiological pathways of the phenotype of interest can be a good starting point. To give an example, when targeting a coloration phenotype, one may first want to sample adult skin tissue to monitor genes involved in pigment production. Embryological stages where pigment cell precursors mature and migrate out of the neural crest may be likewise informative. Further, brain tissue from the hypothalamus associated with hormonal control of pigmentation may contribute to the understanding of trait evolution (Ducrest *et al.* 2008).

Work in the wet laboratory

Laboratory work will be increasingly outsourced to companies and genome centres; thus, most researchers do not need to worry about the specifics of laboratory protocols. Still, sample preparation, RNA extraction and library preparation are important steps for a successful RNA-seq experiment and predefine which information will be retrieved from the transcriptome. Everyone working with RNA-seq data should thus be familiar with the most common steps and their impact on the resulting data.

Sample collection

Immediate shock-freezing tissue in liquid nitrogen is arguably still the most reliable method to prevent fragmentation and eventual loss of RNA due to RNase activity. Where field situations preclude its use, commercially available buffers (e.g. RNAlater, Trizol) or more economic home-made solutions (De Wit *et al.* 2012) generally do a good job of protecting RNA at room temperature for some time.

Contamination

In contrast to microarrays, every RNA molecule stands a chance to appear in the final data. Special precaution should thus be taken to avoid any form of contamination (RNase-free pre-PCR area, separate ventilation system). To exclude contamination with abundant mitochondrial DNA (e.g. in muscle tissue) or DNA contamination from microorganisms, treatment with high-quality DNase is recommended.

RNA extraction and quality assessment

RNA extraction needs to be adjusted to the focal RNA-species: small RNA molecules (<200 bp) will get lost during standard mRNA extraction following typical LiCl precipitation or commercially available kits. Different extraction protocols will be needed for small transcripts such as micro-RNAs. The assessment of RNA integrity [e.g. by micro-capillary electrophoresis (Schroeder *et al.* 2006)] is a critical first step for obtaining meaningful gene expression measurements and should be reported in the final publication.

rRNA depletion or poly-A enrichment

Ribosomal RNA (rRNA) constitutes the predominant fraction of the transcriptome. To avoid wasting sequencing effort on a few superabundant molecules, rRNA needs to be removed prior to library preparation. Where the sequence is known, rRNA can be directly subtracted from the transcript pool. Alternatively, poly-adenylated mRNA molecules can be enriched by capture on oligo-dT-coated magnetic beads or membranes. As poly-adenylation occurs at the 3′ UTR, the latter can introduce a bias in sequencing coverage towards the 3′ end (see fig. 4 Künstner *et al.* 2010).

cDNA synthesis

Most sequencing platforms typically require RNA to be converted to cDNA prior to sequencing. The enzymatic reaction of the reverse transcriptase can be primed either by the hybridization of an oligo-dT primer onto the poly-A tail of the mRNA template or by random hexamer primers. The former can aggravate the 3′ UTR bias, while the latter may introduce biases by sequence context. The jury is still out on which is preferable, and it is sometimes recommended to use a combination.

Library preparation

Library preparation for sequencing is platform specific and cannot be discussed within the scope of this manuscript. Yet, some general issues deserve attention.

Single end vs. paired end—During library preparation, cDNA is fragmented into smaller pieces, which then serve as the template for sequencing. When a single-end strategy is chosen, the fragments are partially sequenced from one end, where paired-end sequencing of short sequences is read from both ends. Paired-end sequencing can be useful for initial transcriptome assembly and for isoform detection, but one should be aware that the insert size should not be too large (generally <300 bp), as otherwise the small size fraction of transcripts will be lost. On the other hand, too short insert sizes can result in adapter contamination, which requires trimming or read removal and complicates the analysis.

Polymerase chain reaction—Currently, most sequencing platforms require appreciable amounts of starting material, which is usually achieved by PCR-based transcriptome amplification during the cDNA synthesis step (but see Raz *et al.* 2011). The efficiency of the PCR naturally depends on template length and sequence content, which will result in a biased, nonlinear relationship between the initial concentration of a gene before and after the PCR. It is advisable to use as few amplification cycles as possible and check the final data for obvious PCR distortions. If present, PCR duplicates need to be dealt with. For genome sequencing at moderate sequencing depth, the likelihood of duplicated reads, *that is*, two reads being randomly drawn from exactly the same

genomic location, is very low. The general recommendation would be to simply purge duplicated reads. For RNA-seq, where the number of reads determines the actual gene expression estimate, duplications are more likely to occur by chance in highly expressed genes, and the removal of duplicated reads would downward bias the expression estimate. Mathematical prediction of the expected number of duplicated reads is necessary to adjust to PCR artefacts, but is complicated by the non-random distribution of reads (due to, for example, 3′ UTR bias, edge effects, GC bias). This makes it less straightforward to make a decision as to the proportion of identical reads that shall be purged for a given expression level. Usage of paired-end reads helps to some degree in this regard. Identical read pairs are more likely to indicate PCR artefacts, as they have a much reduced probability of random duplication due to insert size variation between read pairs. The best way forward, though, may be the use of standardized spike-in controls that naturally integrate many of the PCR-sensitive parameters (Jiang *et al.* 2011).

Library normalization—For initial transcriptome characterization, it has been suggested one can homogenize gene expression levels across genes by library normalization. Sequencing effort will then be distributed more evenly across transcripts and in theory should result in a more complete description of the transcriptome. However, library normalization is a costly and sensitive process that does not necessarily yield a broader representation of the transcriptome (Künstner *et al.* 2010; Vijay *et al.* 2013) and cannot be directly used for gene expression quantification. Given the high throughput of current sequencing platforms, I generally advise against normalizing libraries.

Sequencing strategy

Sequencing platform

At present, the most commonly used sequencing platforms are the pyrosequencing-based 454 system by Roche, the sequencing-by-synthesis-based GA/HiSeq/MiSeq machines from Illumina and the sequencing-by-ligation SOLiD system (Mardis 2008). Others are under development or emerging in the market, such as the semiconductor chip-based IonTorrent system, Helicos' solid-phase-based Genetic Analysis Platform and the single-molecule real-time sequencing-based approach from Pacific Biosciences or Oxford Nanopore (Eid *et al.* 2009; Raz *et al.* 2011; Merriman & Rothberg 2012). As any review will lag behind the fast development in this area, information can be best retrieved on the manufacturers' homepages and through online forums (Box 1). To choose an adequate technology, the important parameters to consider are price per base pair, error rate and error profiles, total output and read length. Where there is a trade-off between read length and total output, the latter seems more important for RNA-seq. While longer reads help in *de novo* transcriptome assembly, paired-end reads perform similarly well. What counts in the end is the number of correctly aligned reads per gene, which determines the accuracy of gene expression measurement and inferential power.

Error profiles

Each technology has its own errors. Attempts are being made to provide standardized error probabilities for each base in the unit of *phred* quality cores that were originally developed for Sanger sequencing (Ewing & Green 1998). Still, error profiles differ between technology and need to be considered when interpreting data. For example, incorrect homopolymer runs, which are common artefacts for the 454 and Ion Torrent technologies, will be more prone to mis-alignment and can thereby influence regional read coverage. Illumina sequencing is sensitive to the GC content of the template, which can affect inferences of gene expression patterns (Wolf & Bryk 2011).

Sequence coverage

What amount of sequence coverage should be targeted in an RNA-seq experiment? Naturally, this depends on the question that shall be addressed. If the entire transcriptome shall be characterized with all lowly expressed genes and most alternatively spliced isoforms, sequencing effort needs to be considerably higher than if a broad inventory of expressed genes is the primary goal. Fortunately, technology has already developed to a point where hundreds of millions of reads are generated per run at moderate costs and barcoding of multiple samples allows adjusting individual coverage. As a rule of thumb, 100 million reads (>100 bp) should provide a decent basis for transcriptome characterization and capture most of the genes present in an RNA sample (Wang *et al.* 2011; Vijay *et al.* 2013). A fraction of this (~10 million reads) will be sufficient to accurately quantify gene expression for individual samples across a broad range of expression levels (Vijay *et al.* 2013). Ultimately, the required coverage depends on the application and expected effect sizes, and the numbers given above can only been seen as rough guidelines. Simulation is a useful tool to assess the coverage necessary for transcriptome inventory or differential expression analysis (Vijay *et al.* 2013), and it may even be warranted to run a subset of the samples to get a feeling for the organism and

genes in question. As the output of the sequencing machines increases and prices drop, the problem of trading-off financial resources vs. analytical power will largely disappear.

Bioinformatic processing

A single run of any sequencing platform generates an appreciable amount of sequencing data, quickly reaching hundreds of gigabytes. Before engaging into RNA-seq, one should thus make sure that the necessary computing, data storage resources and basic bioinformatic expertise are in place. In this section, I will touch upon the routines for processing the raw read data. Most are relevant not only for RNA-seq, but for other high-throughput sequencing applications and have been reviewed in more detail elsewhere (Garber *et al.* 2011; De Wit *et al.* 2012; Lee *et al.* 2012). Finally, I will shortly introduce the statistical methods used to treat this type of gene expression data.

Computing resources

D*e novo* transcriptome assembly consumes more resources than genome-guided approaches. To facilitate the assembly in a reasonable time frame, a computer should contain at least 8 cores and 256 GB of RAM, with a fast storage system in the former case and an 8-core machine with 32 GB of RAM in the latter. Downstream analyses like inference of differential expression can be performed on a desktop computer. For a moderately sized RNA-seq experiment, at least one terabyte of storage space should be set aside. Where sufficient resources are not in place, commercially available cloud computing services may be an attractive alternative (Schatz *et al.* 2010).

Programming skills

Although easy-to-use online tools like the Galaxy platform (Goecks *et al.* 2010) are publicly available, basic knowledge in UNIX shell programming and Perl/Python scripting for data modification come as a great advantage. Moreover, some familiarity with the R programming environment is useful, as softwares for many of the downstream analyses are collected in the Bioconductor suite of R packages (www.bioconductor.org).

File formats

Apart from programming, it is advisable to get familiar with a number of cross-platform file formats including .fasta, .fastq, .sam, .bam, .vcf, .gtf or .gff files.

Quality control

Raw data come with errors and should be preprocessed before being fed into downstream analyses like mapping or assembly. Basic tasks such as adapter removal, duplicate quantification and summary statistics on quality score can be performed by standard tools like the fastQC toolkit (Box 1). The large amount of data precludes comprehensive visual inspection, but spot tests are still important to get a feeling about important aspects of the data like assembly quality, coverage distribution, GC biases and coverage edge effects. Several visualization tools are available for this purpose and some are listed in Box 1. Otherwise, quality control is not yet formally established for RNA-seq data, and it is largely unclear how raw data trimming and quality filtering affect the end results (see also PCR). Importantly, one should always make sure to use a blocked statistical design in setting up the sequencing reaction, as error profiles can differ considerably across sequencing runs and across subsections (e.g. lanes for Illumina) within one sequencing run. If all libraries from one treatment were run on one lane, and all libraries from the other treatment group on another lane, there is no way of knowing whether differences in expression were due to treatment effect or simply reflect differences in sequencing quality.

Transcriptome characterization

RNA-seq data can be used for the identification of transcripts either by mapping reads to an existing genome or by assembling them *de novo*. As the production of a high-quality genome is still an expensive and laborious endeavour, the choice is typically between *de novo* assembly and mapping reads to an existing genome assembly of a distantly related species. Both have their merit. In a simulation study, Vijay *et al.* (2013) showed that 'mapping assemblies' to genomes as distantly related at 15% sequence divergence compared favourably with *de novo* assemblies. Alternatively spliced isoforms, in particular, seem to be better inferred by mapping than by *de novo* tools. Mapping assemblies can, however, only retrieve what has been annotated in the distant reference genome, and all artefacts of the reference will be carried along.

Variant calling

One main application of RNA-seq is the development of molecular markers within the putatively functional genomic elements of transcribed DNA. Many tools like the highly flexible GATK pipeline (DePristo *et al.* 2011) are available for variant calling (reviewed in (Nielsen *et al.* 2011). Different tools, however, will call only partially

overlapping sets of variants, as they take different statistical approaches and differ in which aspects of the data are used (e.g. base pair quality, population allele frequency, substitution rates). As a general strategy, it is advisable to use the intersect from several programs. Moreover, note that variant sites discovered on the basis of small population samples will show clear ascertainment bias to high frequency variants, which significantly biases downstream population genetic analyses (Albrechtsen *et al.* 2010). A specific issue for transcriptome data is allele-specific expression, which makes it difficult (or even impossible) to confidently judge diploid genotypes.

Gene expression quantification

The same consideration – *de novo* vs. using a distant reference – applies to the quantification of transcripts. Read counts can likewise be based on the number of reads mapping to a transcript reference that has been assembled *de novo* (Fig. 2A) or to a distant genomic reference (Fig. 2B). Simulations again suggest that the mapping strategy better represents the initial transcript concentration and outperforms *de novo*-based inference in differential expression analyses (Vijay *et al.* 2013).

Mapping strategy

Aligning millions of short query sequences with sequencing errors onto a genome or transcriptome reference (Fig. 2) is a complicated problem (Trapnell & Salzberg 2009), and accurate read quantification crucially depends on choosing the right strategy. Several decisions need to be taken. First, an appropriate mapping tool must be chosen. Working with nonmodel species will often require that reads be aligned to distant references for which not all existing mapping tools are suitable. For example, the hybrid mapping strategy of *stampy* (Lunter & Goodson 2011) seems to well equipped for handling distances of sequence divergences up above 15% (Vijay *et al.* 2013). Naturally, when mapping to divergent references, the default options for the number of accepted mismatches per read need to be adjusted accordingly.

A second, general challenge with short read alignment is how to deal with ambiguity in read mapping (Treangen & Salzberg 2012). As the similarity between regions of the reference increases (e.g. by copy number variation, multigene families, repetitive domains), the confidence in placing a read at a given location will decrease. There are four basic options for dealing with this issue. First, ambiguously mapped reads are discarded and only uniquely mapped reads are kept. Second, all matches – maybe within a general quality cut-off – are retained, potentially increasing the amount of mapped reads beyond the number of raw reads. Third, the scoring function of the alignment algorithm evaluates the best possible alignment and, in the case of ties, distributes reads randomly across equally good loci. Fourth, mapping algorithms implemented in software packages like RSEM (Li & Dewey 2011) or TopHat (Trapnell *et al.* 2009) divide ambiguous reads in relative proportion according to probabilistic inference. Simulations suggest that the latter strategy best reflects underlying transcript abundance and produces the least bias in inferring (differential) gene expression (Vijay *et al.* 2013).

A last consideration relates to the special situation when mapping transcriptome data to a genomic reference, which requires the use of mapping algorithms that can handle spliced-read alignment (Fig. 2). Several packages such as ERANGE (Mortazavi *et al.* 2008) are available for this purpose and are instrumental for inference of alternative splicing (see below, Box 1). The TopHat spliced-read mapper is particularly attractive, as it does not rely on a fully annotated genome and merely requires a raw genome sequence as a backbone (Trapnell *et al.* 2009).

Gene name assignment

Gene name assignment is a vital step for drawing biologically meaningful conclusions from RNA-seq experiments and for comparing results among different studies (see *gene function and interaction below*). In mapping approaches to annotated reference genomes, gene names come for free, but in the case of *de novo* assemblies, contigs provide no information about the sequenced gene, and their assignment to orthologous genes from (distantly) related genomes is not always straightforward. Suffix-tree-based methods such as *NUCmer* and *PROmer* seem to work well for closely related species, whereas BLAST-based orthology detection seems to be a better alternative for more distantly related species (Vijay *et al.* 2013). In the latter case, conservative filtering and reciprocal-assignment should be applied to guard against false positives such as paralogues with high sequence similarity.

Statistical treatment of RNA-seq data

After mapping, per transcript read counts can be used as a relative measure of transcript abundance. In a perfect world, transcript abundance of steady-state mRNA should be directly proportional to the number of reads: a transcript from gene A with twice the cellular concentration of transcript B should have twice as many reads. This relationship should hold across a large range of expression levels spanning several orders of magnitude.

Normalization

Normalization is a crucial component for RNA-seq data. When comparing transcripts of different length, it is intuitive to control for length, as a longer transcript will be covered with more reads than an equally expressed shorter transcript (Fig. 2). In genetic nonmodel organisms, however, we are often ignorant about target transcript size, particularly for lowly expressed transcripts of which only fragments are assembled. In such a situation, transcript length of a related species can be used as a reasonable proxy (Wolf *et al.* 2010). Transcript length is less of a problem for comparisons of the same transcript across treatments (differential expression) and is thus often not incorporated in statistical methods used for inferring differential expression.

When comparing gene expression profiles between two samples, another important aspect of normalization is to control for differences in sequencing effort and quality between two samples. Consider an RNA sample of individual A that has been sequenced to twice the depth of coverage as individual B. All genes from individual A will appear to be expressed at a higher level, even if they have the same relative concentration in the cell. A simple way to handle this is to divide read counts by the total number of mappable reads (or quantiles of mappable reads) (Bullard *et al.* 2010). Such basic normalization controlling for transcript length and sequencing effort is captured by the commonly used RPKM or FPKM measures (reads or fragments per kilobase exon per million reads, Fig. 2), but more refined normalization may be necessary for firm inference. Standardized spike-in RNA controls of known concentration, defined length and GC content may be best suited to assure comparability across transcripts, samples, protocols and platforms (Jiang *et al.* 2011).

An important aspect, which has only recently received attention, is carry-over effects of gene expression from a few genes to others. Imagine samples from two individuals, each containing 10 mRNA transcripts of equal length (no length normalization necessary). The total sequencing effort is the same in both samples, say 1000 reads (no normalization for sequencing effort needed). In sample 1, all genes are expressed at equal rates, so that 100 reads are expected per transcript. In sample 2, genes 1 through 9 are also expressed at equal rates and have exactly the same concentration in the cell as in sample 1. The concentration of gene 10, however, is 9 times higher than in sample 1. As sequencing effort (total number of reads) is distributed across all transcripts in a sample, genes 1 through 9 will only receive 55.5 reads each, while gene 10 receives 500 reads. Now, when comparing sample 1 to sample 2, all genes will appear to be differentially expressed, although genes 1 through 9 had exactly the same concentration in the cell in both samples. The expression difference in gene 10 influenced all other genes. Normalization methods, such as trimmed mean normalization (e.g. edgeR, Robinson & Oshlack 2010), have been suggested to address this carry-over effect. Another way that has received surprisingly little attention is the use of invariant internal control (housekeeping) genes for normalization (Brawand *et al.* 2011), as has long been the standard for quantitative PCR analyses.

Differential gene expression

Apart from normalization, it is important to find a statistical distribution approximating the nature of the data. The statistical properties of count data are generally well described by a Poisson process. However, many aspects of RNA-seq such as library preparation and mapping errors inflate the variance of read counts beyond that expected for a Poisson distribution (overdispersion). It has thus become common practice to model RNA-seq read count data as an overdispersed Poisson process or by a negative binomial distribution, which is routinely used to accommodate an overdispersed Poisson process (Kvam *et al.* 2012). With a statistical distribution at hand that captures the essential information of the data, parametric statistical inference is possible. This opens the door to differential expression analyses, *that is,* to formally compare expression levels between transcripts and samples across treatment groups. Several software packages performing this task at different levels of sophistication are currently available (e.g. DESeq, edgeR, baySeq, NOIseq, see Box 1). Statistical method development (e.g. how to best estimate overdispersion) is an active area of research and novel or refined methods are to be expected.

Alternative splicing

So far, I have conceptually simplified read quantification to one transcript per gene irrespective of transcript isoform. While this is reminiscent of traditional microarray profiling, it disregards the biological reality of alternative splicing (Fig. 2). One of the great powers of RNA-seq is uncovering this reality. Two basic approaches can be taken. First, one can try to reconstruct the most likely set of transcript isoforms *de novo* (transcriptome inventory). Without any prior annotation information of the gene, this is a difficult problem, and only few software developers have taken on the challenge of *de novo* isoform characterization (Grabherr *et al.* 2011). *De novo* approaches are error prone, and whenever an annotated, even distantly related, genome is available, one should make use of it (Vijay *et al.* 2013). When having a genome reference, it is easier (but still difficult) to infer full-length

transcript isoforms from spliced-reads and coverage differences between unique and shared parts of the isoforms. This is exploited by approaches such as MMSEQ (Turro *et al.* 2011), Solas (Richard *et al.* 2010) and Cufflinks (Trapnell *et al.* 2012), which at the same time can conduct isoform-specific differential expression analyses. Where a satisfactory transcript inventory has already been made, alternative approaches are possible that avoid isoform reconstruction altogether and directly look at expression differences between exons (Anders *et al.* 2012).

Gene function and interaction

A successful RNA-seq experiment will yield a set of candidate genes that differ between treatments or populations. The list itself is of limited interest and external information in necessary to infer their biological function. Are these genes over-represented in a metabolic pathway of interest? What potential role do they play in the organisms that could be relevant to adaptation? Several external sources of information and analytical procedures exist to address these and related questions (Box 1). The gene ontology database is arguably the most prominent initiative for comparing gene functions across species using a controlled vocabulary (Ashburner *et al.* 2000; Gene Ontology Consortium 2004). However, it is important to keep in mind that detailed gene function is mostly based on inbred-strains of model organisms and may have little to do with the function of the orthologue in the study organism. One should also be aware that gene ontology analyses invite to *a posteriori* story telling. In many cases, it may be more rewarding to map the genes of interest directly to candidate metabolic pathways (e.g. KeGG Ogata *et al.* 1999) or protein interaction networks (Szklarczyk *et al.* 2011; Leskinen *et al.* 2012).

Future applications and outlook

Next-generation sequencing has democratized the field of transcriptome analysis and brought it into the wild. We will soon see a plenitude of studies from (formerly) genetic nonmodel organisms attacking ecological and evolutionary questions with the tools outlined above. Whole-genome assembly and functional genome annotation with RNA have become a realistic goal (Ellegren *et al.* 2012), and a number of scientific questions such as dosage compensation that have been reserved to genetic models can now benefit from the contribution of a much extended sample of species (Wolf & Bryk 2011). Heralded by the human ENCODE project (The ENCODE Project Consortium 2012), it is conceivable that we will soon be able to harvest epigenetic information, characterize regulatory elements and address areas like isoform- or allele-specific expression and post-transcriptional processes, such as RNA editing in an evolutionary and population genetics context (Skelly *et al.* 2011). To achieve this goal, methodological integration across subdisciplines will be crucial. A list of differentially expressed genes, for example, makes much more sense when placed in a physiologically relevant context. Emerging tools from systems biology going beyond simple GO-terms like causal network modelling linking gene expression analysis to gene interaction information are sorely needed (Chindelevitch *et al.* 2012). Also, indirect experimental validation methods like cell-specific mRNA quantification (Larsson *et al.* 2010) or *in situ* hybridization will be instrumental in getting a grip on the functional aspects of evolution in wild organisms where transgenic constructs are generally not possible to test for the effect of single genes. Another challenge lies in methodological developments, both at the stage of sequence generation (single-molecule sequencing, direct RNA-sequencing, no PCR amplification) and downstream analyses. To date there are a vast number of different tools at different steps of the RNA-seq work flow that will influence the outcome of the experiment. As the field matures, we can hope to see methodological standardization, which will be highly welcomed by practitioners in the ecological and evolutionary sciences. What we definitely will see is exciting biology.

Acknowledgements

I would like to thank all key members of the Next-Generation Sequencing Club at the Evolutionary Biology Centre in Uppsala that over the last years have contributed to the fruitful discussion on often hard-to-digest methodological approaches to high-throughput sequencing. Special thanks go to Christen Bossu, Robert Ekblom, Axel Künstner, Aaron Shafer, Severin Uebbing and three anonymous reviewers who provided useful comments on the manuscript.

Glossary of the main terms

alignment: Similarity-based arrangement of DNA, RNA or protein sequences. In this context, subject and query sequence should be orthologous and should reflect evolutionary, not functional or structural relationships.

assembly: Computational reconstruction of a longer sequence (e.g. a transcript) from smaller sequence reads. ***De novo* assembly** refers to the reconstruction without making use of any reference sequence.

barcode: Short sequence identifier for individual labelling (**barcoding**) of sequencing libraries.

cDNA: Complementary DNA synthesized from a mRNA template.

contig: Contiguous RNA or DNA consensus sequence from a set of overlapping shorter segments (here reads).

coverage: S*equence coverage* refers to the average number of reads per locus and differs from *physical coverage*, a term often used in genome assembly referring to the cumulative length of reads or read pairs expressed as a multiple of genome size.

DNase: Type of enzyme that catalyses the hydrolytic cleavage of phosphodiester linkages in the DNA backbone. A means of removing DNA from an RNA sample.

edge effects: As transcripts are of finite length, read coverage towards the end of a transcript will be lower than in the middle. rRNA depletion, cDNA preparation and GC sensitivity during sequencing can introduce additional nonrandom coverage variation along the transcript.

exon: Building block of pre-mRNA that is retained during splicing. An exon can include protein-coding sequence and untranslated regions.

GC content: The proportion of guanine and cytosine bases in a DNA/RNA sequence.

gene expression: The process by which information from a gene is used in the synthesis of a functional gene product.

gene expression level: Abundance of a gene product in a given reference set of cells.

gene expression profile: Composition of transcripts and their relative expression levels in a given reference set of cells.

gene ontology: Structured, controlled vocabularies and classifications of gene function across species and research areas.

GO-term: Gene ontology term.

insert size: Length of randomly sheared fragments (from the genome or transcriptome) sequenced from both ends.

library: Collection of RNA or DNA fragments modified in a way that is appropriate for downstream analyses such as high-throughput sequencing in this case.

mapping: A term routinely used to describe alignment of short sequence reads.

mRNA: Messenger RNA mediating information from the DNA molecule to the ribosome where it serves as a template for the amino acid sequence of a protein. A subset of the transcriptome.

noncoding RNA: Functional RNA molecule that is transcribed, but not translated into a protein sequence, *for example* micro-RNA, small-interfering RNA.

normalization: Mathematical procedure to ensure comparability of a measure across different conditions.

microarray: A muliplex array of oligonucleotides used for high-throughput screening of transcript abundance.

next-generation (or massively parallel) sequencing: Nano-technological application used to determine the base pair sequence of a DNA/RNA molecule at much larger quantities than previous end-termination (e.g. Sanger sequencing)-based sequencing techniques.

oligo-dT: A short sequence of consisting of deoxythymine nucleotides.

read: Short base pair sequence inferred from the DNA/RNA template by sequencing.

RNA-seq: High-throughput shotgun transcriptome sequencing. Here, not used synonymous to RNA-sequencing which implies direct sequencing of RNA molecules skipping the cDNA generation step.

RNase: Type of nuclease enzyme catalysing the degradation of RNA into smaller components.

splicing. Modification of pre-mRNA in which introns are removed and exons are retained. **Alternative splicing** refers to the retention of different combinations of exons.

transcript: An RNA molecule copied (transcribed) from a DNA template.

transcript isoform: Transcript with a unique combination of exons.

transcriptome: Set of all RNA molecules transcribed from a DNA template.

variant calling: Computational identification of locus-specific sequence polymorphism.

References

Albrechtsen A, Nielsen FC, Nielsen R (2010) Ascertainment biases in SNP chips affect measures of population divergence. *Molecular Biology and Evolution*, **27**, 2534–2547.

Anders S, Huber W (2010) Differential expression analysis for sequence count data. *Genome Biology*, **11**, R106.

Anders S, Reyes A, Huber W (2012) Detecting differential usage of exons from RNA-seq data. *Genome Research*, **22**, 2008–2017.

Andolfatto P, Davison D, Erezyilmaz D *et al.* (2011) Multiplexed shotgun genotyping for rapid and efficient genetic mapping. *Genome Research*, **21**, 610–617.

Ashburner M, Ball CA, Blake JA *et al.* (2000) Gene ontology: tool for the unification of biology. *Nature Genetics*, **25**, 25–29.

Auer PL, Doerge RW (2010) Statistical design and analysis of RNA sequencing data. *Genetics*, **185**, 405–416.

Baird NA, Etter PD, Atwood TS, *et al.* (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE*, **3**, e3376.

Bao H, Guo H, Wang J *et al.* (2009) MapView: visualization of short reads alignment on a desktop computer. *Bioinformatics*, **25**, 1554–1555.

Bao S, Jiang R, Kwan W *et al.* (2011) Evaluation of next-generation sequencing software in mapping and assembly. *Journal of Human Genetics*, **56**, 406–414.

Brawand D, Soumillon M, Necsulea A *et al.* (2011) The evolution of gene expression levels in mammalian organs. *Nature*, **478**, 343–348.

Bullard J, Purdom E, Hansen K, Dudoit S (2010) Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*, **11**, 94.

Chindelevitch L, Ziemek D, Enayetallah A *et al.* (2012) Causal reasoning on biological networks: interpreting transcriptional changes. *Bioinformatics*, **28**, 1114–1121.

De Wit P, Pespeni MH, Ladner JT *et al.* (2012) The simple fool's guide to population genomics via RNA-Seq: an introduction to high-throughput sequencing data analysis. *Molecular Ecology Resources*, **12**, 1058–1067.

DePristo MA, Banks E, Poplin R *et al.* (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, **43**, 491–498.

Ducrest AL, Keller L, Roulin A (2008) Pleiotropy in the melanocortin system, coloration and behavioural syndromes. *Trends in Ecology & Evolution*, **23**, 502–510.

Eid J, Fehr A, Gray J *et al.* (2009) Real-time DNA sequencing from single polymerase molecules. *Science*, **323**, 133–138.

Ekblom R, Galindo J (2011) Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity*, **107**, 1–15.

Ellegren H, Smeds L, Burri R *et al.* (2012) The genomic landscape of species divergence in Ficedula flycatchers. *Nature*, **491**, 756–760.

Elshire RJ, Glaubitz JC, Sun Q *et al.* (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE*, **6**, e19379.

Ewing B, Green P (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Research*, **8**, 186–194.

Garber M, Grabherr MG, Guttman M, Trapnell C (2011) Computational methods for transcriptome annotation and quantification using RNA-seq. *Nature Methods*, **8**, 469–477.

Gene Ontology Consortium (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Research*, **32**, 258D–261.

Gentleman R, Carey V, Bates D *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, **5**, R80.

Goecks J, Nekrutenko A, Taylor J, The Galaxy Team (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology*, **11**, R86.

Grabherr MG, Haas BJ, Yassour M *et al.* (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, **29**, 644–652.

Guo M, Rupe MA, Zinselmeier C *et al.* (2004) Allelic variation of gene expression in maize hybrids. *The Plant Cell Online*, **16**, 1707–1716.

Hardcastle TJ, Kelly KA (2010) baySeq: Empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics*, **11**, 422.

Harr B, Turner LM (2010) Genome-wide analysis of alternative splicing evolution among Mus subspecies. *Molecular Ecology*, **19**, 228–239.

Jain M (2012) Next-generation sequencing technologies for gene expression profiling in plants. *Briefings in Functional Genomics*, **11**, 63–70.

Jiang L, Schlesinger F, Davis CA *et al.* (2011) Synthetic spike-in standards for RNA-seq experiments. *Genome Research*, **21**, 1543–1551.

Katz Y, Wang ET, Airoldi EM, Burge CB (2010) Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nature Methods*, **7**, 1009–1015.

Künstner A, Wolf JBW, Backstrom N *et al.* (2010) Comparative genomics based on massive parallel transcriptome sequencing reveals patterns of substitution and selection across 10 bird species. *Molecular Ecology*, **19**, 266–276.

Kvam VM, Liu P, Si Y (2012) A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data. *American Journal of Botany*, **99**, 248–256.

Lamichhaney S, Barrio AM, Rafati N *et al.* (2012) Population-scale sequencing reveals genetic differentiation due to local adaptation in Atlantic herring. *Proceedings of the National Academy of Sciences*, **109**, 19345–19350.

Larsson C, Grundberg I, Söderberg O, Nilsson M (2010) In situ detection and genotyping of individual mRNA molecules. *Nature Methods*, **7**, 395–397.

Lee HC, Lai K, Lorenc MT *et al.* (2012) Bioinformatics tools and databases for analysis of next-generation sequence data. *Briefings in Functional Genomics*, **11**, 12–24.

Leng N, Dawson JA, Thomson JA *et al.* (2013) EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. *Bioinformatics*, **29**, 1035–1043.

Lenz TL, Eizaguirre C, Rotter B, Kalbe M, Milinski M (2013) Exploring local immunological adaptation of two stickleback ecotypes by experimental infection and transcriptome-wide digital gene expression analysis. *Molecular Ecology*, **22**, 774–786.

Leskinen PK, Laaksonen T, Ruuskanen S, Primmer CR, Leder EH (2012) The proteomics of feather development in pied flycatchers (Ficedula hypoleuca) with different plumage coloration. *Molecular Ecology*, **21**, 5762–5777.

Levin JZ, Yassour M, Adiconis X *et al.* (2010) Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nature Methods*, **7**, 709–715.

Li B, Dewey CN (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, **12**, 323.

Li H, Handsaker B, Wysoker A *et al.* (2009) The Sequence Alignment/Map (SAM) format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Lunter G, Goodson M (2011) Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Research*, **21**, 936–939.

Magi A, Benelli M, Gozzini A *et al.* (2010) Bioinformatics for next generation sequencing data. *Genes*, **1**, 294–307.

Mardis ER (2008) The impact of next-generation sequencing technology on genetics. *Trends in Genetics*, **24**, 133–141.

Margulies M, Egholm M, Altman WE *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.

Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research*, **18**, 1509–1517.

Martin JA, Wang Z (2011) Next-generation transcriptome assembly. *Nature Reviews Genetics*, **12**, 671–682.

Merriman B, Rothberg JM, R&D Team IT (2012) Progress in Ion Torrent semiconductor chip based sequencing. *Electrophoresis*, **33**, 3397–3417.

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

Milne I, Bayer M, Cardle L *et al.* (2010) Tablet—next generation sequence assembly visualization. *Bioinformatics*, **26**, 401–402.

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

Naurin S, Bensch S, Hansson B *et al.* (2008) A microarray for large-scale genomic and transcriptional analyses of the zebra finch (Taeniopygia guttata) and other passerines. *Molecular Ecology Resources*, **8**, 275–281.

Nielsen R, Paul JS, Albrechtsen A, Song YS (2011) Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics*, **12**, 443–451.

Nolte AW, Renaut S, Bernatchez L (2009) Divergence in gene regulation at young life history stages of whitefish (Coregonus sp.) and the emergence of genomic isolation. *BMC Evolutionary Biology*, **9**, 59.

Nookaew I, Papini M, Pornputtpong N *et al.* (2012) A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in Saccharomyces cerevisiae. *Nucleic Acids Research*, **40**, 10084–10097.

Ogata H, Goto S, Sato K *et al.* (1999) KEGG: Kyoto Encyclopedia of genes and genomes. *Nucleic Acids Research*, **27**, 29–34.

Orsini L, Andrew R, Eizaguirre C (2013) Evolutionary ecological genomics. *Molecular Ecology*, **22**, 527–531.

Oshlack A, Robinson MD, Young MD (2010) From RNA-seq reads to differential expression results. *Genome Biology*, **11**, 220.

Ozsolak F, Milos PM (2011) RNA sequencing: advances, challenges and opportunities. *Nature Reviews Genetics*, **12**, 87–98.

Patel RK, Jain M (2012) NGS QC toolkit: a toolkit for quality control of next generation sequencing data. *PLoS ONE*, **7**, e30619.

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

Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.

Raz T, Causey M, Jones DR *et al*. (2011) RNA sequencing and quantitation using the Helicos Genetic Analysis System. In: *High-Throughput Next Generation Sequencing: Methods and Applications* (eds Young Min Kwon, Steven C. Ricke), pp. 37–49. Springer Science+Business Media, Heidelberg.

Richard H, Schulz MH, Sultan M *et al*. (2010) Prediction of alternative isoforms from exon expression levels in RNA-Seq experiments. *Nucleic Acids Research*, **38**, e112:1–e112:15.

Robinson MD, Oshlack A (2010) A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, **11**, R25.

Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.

Schatz MC, Langmead B, Salzberg SL (2010) Cloud computing and the DNA data race. *Nature Biotechnology*, **28**, 691–693.

Schroeder A, Mueller O, Stocker S *et al*. (2006) The RIN: an RNA integrity number for assigning integrity values to RNA measurements. *BMC Molecular Biology*, **7**, 3.

Skelly DA, Johansson M, Madeoy J, Wakefield J, Akey JM (2011) A powerful and flexible statistical framework for testing hypotheses of allele-specific gene expression from RNA-seq data. *Genome Research*, **21**, 1728–1737.

Stapley J, Reger J, Feulner PGD *et al*. (2010) Adaptation genomics: the next generation. *Trends in Ecology & Evolution*, **25**, 705–712.

Szklarczyk D, Franceschini A, Kuhn M *et al*. (2011) The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Research*, **39**, D561–D568.

Tarazona S, García-Alcalde F, Dopazo J, Ferrer A, Conesa A (2011) Differential expression in RNA-seq: a matter of depth. *Genome Research*, **21**, 2213–2223.

Tautz D, Ellegren H, Weigel D (2010) Next generation molecular ecology. *Molecular Ecology*, **19**, 1–3.

The ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.

Thorvaldsdóttir H, Robinson JT, Mesirov JP (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics*, **14**, 178–192.

Tirosh I, Reikhav S, Levy AA, Barkai N (2009) A yeast hybrid provides insight into the evolution of gene expression regulation. *Science*, **324**, 659–662.

Trapnell C, Salzberg SL (2009) How to map billions of short reads onto genomes. *Nature Biotechnology*, **27**, 455–457.

Trapnell C, Pachter L, Salzberg SL (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, **25**, 1105–1111.

Trapnell C, Roberts A, Goff L *et al*. (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols*, **7**, 562–578.

Treangen TJ, Salzberg SL (2012) Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews Genetics*, **13**, 36–46.

Turro E, Su S-Y, Goncalves A *et al*. (2011) Haplotype and isoform specific expression estimation using multi-mapping RNA-seq reads. *Genome Biology*, **12**, R13.

Vera JC, Wheat CW, Fescemyer HW *et al*. (2008) Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. *Molecular Ecology*, **17**, 1636–1647.

Vijay N, Poelstra JW, Künstner A, Wolf JBW (2013) Challenges and strategies in transcriptome assembly and differential gene expression quantification. A comprehensive in silico assessment of RNA-seq experiments. *Molecular Ecology*, **22**, 620–634.

Vogel C, de Sousa Abreu R, Ko D *et al*. (2010) Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line. *Molecular Systems Biology*, **6**, 400.

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

Wang Y, Ghaffari N, Johnson CD *et al*. (2011) Evaluation of the coverage and depth of transcriptome by RNA-Seq in chickens. *BMC Bioinformatics*, **12**, S5.

Wolf JBW, Bryk J (2011) General lack of global dosage compensation in ZZ/ZW systems? Broadening the perspective with RNA-seq. *BMC Genomics*, **12**, 91.

Wolf JBW, Bayer T, Haubold B *et al*. (2010) Nucleotide divergence versus gene expression differentiation: comparative transcriptome sequencing in natural isolates from the carrion crow and its hybrid zone with the hooded crow. *Molecular Ecology*, **19**, 162–175.