**RNA-Seq: Current Methods and Potential Applications**

Philip Pauerstein
Computational Molecular Biology
Winter 2011

Gene expression patterns hold valuable information regarding the specific functions of particular cell and tissue types. Initially, the analysis of transcripts was limited to testing changes in expression of only a few genes at a time using low-throughput methods such as *in situ* hybridization and RT-PCR. In the past decade, with the ability to study genetic information at the genome-wide scale, microarrays have become the primary high-throughput method for gene expression analysis. Based on relative changes in the amount of hybridization of cDNA, relative expression values are computed for populations of genes. Microarray analysis has certain limitations, including the inability to identify novel transcripts, a limited dynamic range for detection, and difficulty in replicability and inter-experimental comparison. RNA sequencing (RNA-Seq) overcomes many of these problems. Making use of high-throughput next-generation sequencing methods, sequencing the entire transcriptome permits both transcript discovery and robust digital quantitative analysis of gene expression levels. This document reviews basic biological principles of RNA-Seq, basic computational methods for analysis and use of RNA-Seq data, and potential medical and clinical applications of RNA-Seq.

*Biological Principles and Basic Methods of Transcriptome Analysis by RNA Sequencing*

RNA-Seq was developed in 2008, and was first applied to analysis of yeast, mouse, and Arabidopsis transcriptomes (Lister et al., 2008; Mortazavi et al., 2008; Nagalakshmi et al., 2008; Wilhelm et al., 2008). To achieve its two major functions – transcript discovery and quantification of gene expression – this method relies on the generation of short reads of transcript sequence information

which are then assembled into full-length transcripts (contigs) and mapped to the genome. To generate this data, isolated RNA populations (e.g., total, polyadenylated, small RNAs) are converted to cDNA for sequencing. More recently developed methods involve direct sequencing of RNA to avoid artifacts generated by reverse transcription and subsequent modification steps (Ozsolak and Milos, 2011). For longer nucleic acid sequences (>500 bp), fragmentation of either RNA or cDNA is necessary to allow processing by next-generation sequencing, a step which introduces bias. Fragmentation of cDNA results in a bias toward 3' ends of transcripts, while fragmentation of RNA molecules results in more homogenous coverage of most of the transcript with loss of information at either end (Wang et al., 2009). Adaptors are attached to the ends of these fragments to enable efficient sequencing, and short sequences are generated. These sequences are aligned to known sequences and are used to generate gene expression data – including expression levels, sequences, and splice forms (Wang et al., 2009).

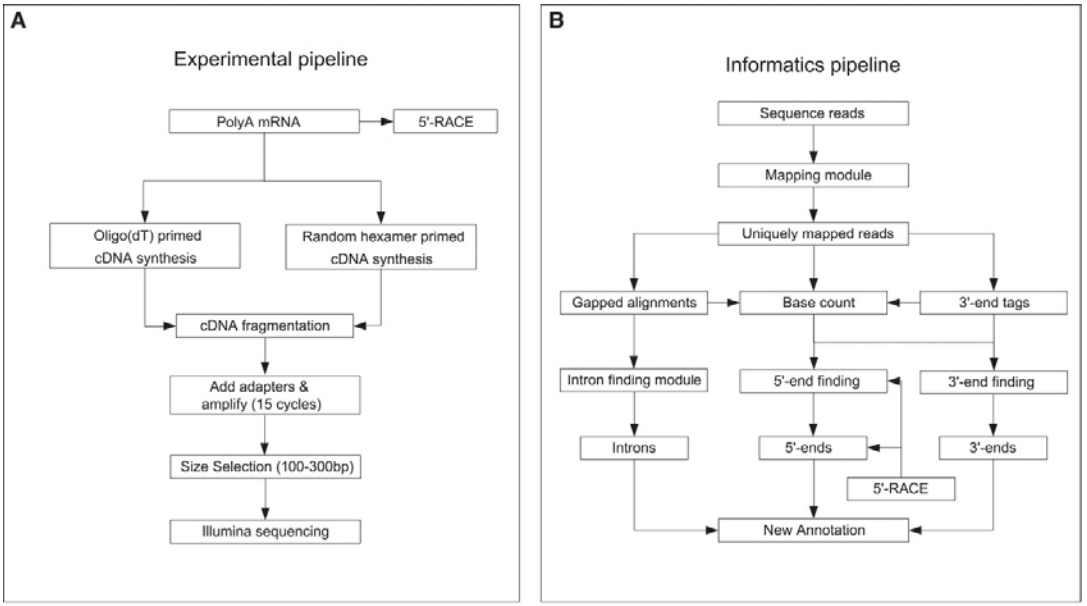


Figure 1. Outline of an example RNA-Seq procedure, from one of the four original reports using this method (adapted from Nagalakshmi et al., 2008).

Early methods for RNA-Seq involved cDNA conversion and some amplification, and sequence reads were short at less than 50 bp per read. Each of these steps resulted in opportunities for bias and

error – reverse transcriptase is error-prone, resulting in reduced sequence quality, amplification methods may introduce unknown biases toward certain transcripts over others, and short reads increase the risk of incorrect mapping of transcripts with regions of high homology or repeated sequences. Recent developments including direct RNA sequencing, paired-end sequencing, and the use of sequencing methods that generate longer reads have helped to overcome some of these problems (Ozsolak et al., 2009; Ozsolak et al., 2011; Wang et al., 2009). In particular, direct RNA sequencing using single-molecule sequencing platforms avoids cDNA conversion and amplification steps and requires femtomole quantities of RNA – this reduction in the amount of input RNA required could in the long term make direct RNA sequencing useful for identifying transcriptome characteristics of small samples or rare cell types. Direct RNA sequencing is currently limited, however, to polyadenylated transcripts, making it unsuitable for analysis of small or non-coding RNAs which may have major functions in cells and tissues.

Currently, hybridization-based microarrays are the primary method for global gene expression analysis. Because of its advantages in sensitivity,

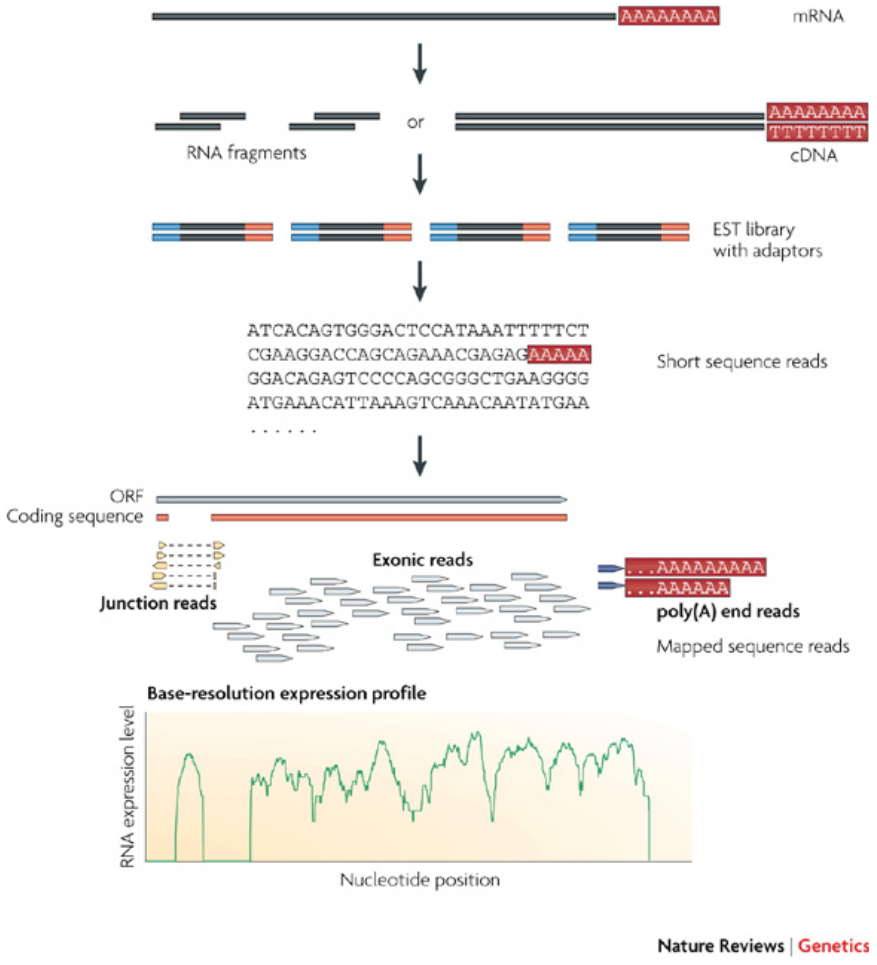


Figure 2. Summary of the RNA-Seq procedure. mRNA is converted to cDNA, and fragments from that library are used to generate short sequence reads. Those reads are assembled into contigs which may be mapped to reference sequences (Wang et al., 2009).

quantification, and replicability of experiments, RNA-Seq has the potential to replace microarrays in transcriptome analysis. The number of reads for a given transcript (calculated in reads per kilobase of exon model per million reads, or RPKM) corresponds to the absolute expression level of that particular gene in the cell type or tissue in question, providing an absolute quantification method with essentially no ceiling – dynamic ranges of detection of up to five orders of magnitude difference have been reported (Mortazavi et al., 2008; Ozsolak et al., 2011; Wang et al., 2009). This contrasts with the relatively limited dynamic range of microarrays that depends on relative, rather than absolute, quantification of hybridization intensities. Because this technique is based on sequencing, no prior knowledge of transcript sequence is required, bypassing another limitation of microarray analysis – namely, that complementary probes must be synthesized before hybridization, and these probes must generally correspond to known transcripts (Mortazavi et al., 2008; Wang et al., 2009). RNA-Seq thus permits the discovery of novel transcripts. These may include mRNA molecules that encode fusion proteins in tumor cell populations or, as can be detected using sequence analysis, alternate splice forms (Ozsolak et al., 2011; Mortazavi et al., 2008).

Questions of replicability of microarray data have been raised in recent years – comparisons of microarray data from separate experiments and accurate repetition of experiments is difficult (Ioannidis et al., 2009). RNA-Seq has proven to be more repeatable

| Technology | Tiling microarray | RNA-Seq |
|---|---|---|
| *Technology specifications* | | |
| Principle | Hybridization | High-throughput sequencing |
| Resolution | From several to 100 bp | Single base |
| Throughput | High | High |
| Reliance on genomic sequence | Yes | In some cases |
| Background noise | High | Low |
| *Application* | | |
| Simultaneously map transcribed regions and gene expression | Yes | Yes |
| Dynamic range to quantify gene expression level | Up to a few-hundredfold | >8,000-fold |
| Ability to distinguish different isoforms | Limited | Yes |
| Ability to distinguish allelic expression | Limited | Yes |
| *Practical issues* | | |
| Required amount of RNA | High | Low |
| Cost for mapping transcriptomes of large genomes | High | Relatively low |

Table 1. Comparison of microarray and RNA-Seq analysis of gene expression. (Adapted from Wang et al., 2009.)

and robust across experimental groups than microarrays, providing yet another advantage over the previous generation of methods for global gene expression analysis (Mortazavi et al., 2008; Ozsolak et al., 2011). The high resolution provided by RNA-Seq has also been used to demonstrate its strength over microarray analysis – in a recent case study, microarray data had supported the hypothesis that X-linked genes are expressed at levels equal to autosomal genes, suggesting a form of dosage compensation, but RNA-Seq analysis demonstrated that X-encoded transcripts are in fact present at half the levels of autosomal transcripts, indicating that dosage compensation does not occur in mammals (Xiong et al., 2010).

With its single-base pair resolution, sensitivity, and replicability, RNA sequencing has the potential to replace microarray analysis as the preferred method of whole-transcriptome-scale gene expression analysis. Doing so will require technical improvements in sample preparation and sequencing methods to permit the use of smaller sample sizes and the ability to efficiently process and utilize the large amounts of data generated in RNA-Seq experiments.

*Basic Bioinformatic Methods for Analysis of RNA-Seq Datasets*

Analysis of RNA-Seq datasets requires the ability to map short sequence reads to reference genomes or transcriptomes. Subsequently, two major goals must be achieved during data analysis: counting of transcripts to calculate gene expression levels and identification of novel transcripts or gene variants, such as gene fusions, or transcript characteristics, such as splice regions or transcription start sites. Known genes are relatively simple to identify and quantify in a normalized manner in reads per kilobase per million (RPKM), values which reflect actual RNA levels (Mortazavi et al., 2008; Pepke et al., 2009). Exon sequences and known splice sites in reference sequences can be applied to RNA-Seq reads.

ERANGE (Enhanced Read Analysis of Gene Expression) was a software package described in one of the initial reports of RNA-Seq to align sequences to the reference genome, assign splice sites, and count the expression levels of transcripts (Mortazavi et al., 2008).

The exon models used in a particular experiment can influence RPKM values, depending on the presence of novel transcripts or transcripts that do not match reference sequences (Pepke et al., 2009). While most analyses of RNA-Seq data rely on alignment to reference genome sequences, certain recently developed software packages such as Rnnotator are able to assemble RNA-Seq data into transcriptomes without consulting a reference sequence by assembling short contiguous reads from RNA sequencing datasets (Martin et al., 2010). This software allows the identification of novel transcripts and the unbiased detection of transcripts from multiple sources, allowing more efficient application of RNA-Seq to transcript discovery and characterization of transcriptomes in undefined populations, such as non-model organisms or mixed populations consisting of multiple organisms (such as microbial populations) (Martin et al., 2010).

Splice site identification presents another major analytic question in RNA-Seq data processing. Initially, ungapped contig sequences containing multiple exons from a transcript sequence were mapped across known splice sites from reference transcript sequences to identify splice forms (Pepke et al., 2009). Recent attempts to bypass the requirement for known splice sites have made use of methods for assembling transcript sequence fragments to the reference genome and mapping possible splice sites by noting gap locations. TopHat performs novel splice site identification using this method with a seed-and-extend algorithm (Trapnell et al., 2009). Thus, based on sequence data, RNA-Seq in combination with specific software algorithms can identify known and novel splice forms of transcripts. Longer read lengths help to improve accurate splice site detection, as do paired-end reads (Ozsolak et al., 2011).

Multireads, or sequences that match multiple independent genes, present a difficulty in quantifying transcript levels. Highly conserved domains of certain protein families and repetitive

sequences can map to multiple locations in the reference sequence, preventing accurate quantification of transcript levels for those short reads (Mortazavi et al., 2008). Extension of read length by use of sequencing technologies capable of longer reads or by application of paired-end reads to provide short sequences from both ends of cDNA fragments can improve proper mapping of these sequences to single locations in the genome (Wang et al., 2009; Ozsolak et al., 2011).

Data output for RNA-Seq must address both quantification of expression levels and transcript discovery in accessible formats. Software and infrastructure for microarray-based gene expression analysis should be applicable to RNA-Seq datasets – RPKM values effectively replace hybridization intensities, and datasets can be analyzed in biologically meaningful ways with Gene Ontology-based methods. Transcript discovery requires the identification of individual events that do not match known transcript sequences or characteristics, including splice site variations, promoter variations, and allele sequence variations (Wang et al., 2009).

| | Primary category | Discovery | Need genomic assembly | Associated read mapper | Splice junctions | Quantitation |
|---|---|---|---|---|---|---|
| ABySS v1.0.11 | Short-read assembler | Yes | No | NA | Assembled | Read coverage |
| BASIS V1 | Existing transcript quantitation | No | Yes | External | From existing models | Read coverage |
| ERANGE v3.1 | Existing and novel gene quantitation | Yes | Yes | BlatBowtieEland | From existing modelsNovel with blat | RPKM from gene annotations and novel transfrags |
| G-Mo.R-Se v1.0 | Novel gene model annotation | Yes | Yes | SOAP | Predicted from transfrags | No |
| QPALMA v0.9.9.2 | Spliced read mapper | Yes | Yes | Integrated | Predicted from transfrags | No |
| RNA-mate v1.1 | Existing and novel gene quantitation | Yes | Yes | Map reads | From existing models | Deprecated in v1.1 |
| RSAT v0.0.3 | Existing transcript quantitation | No | No; requires transcript sequences | Eland, SeqMap (bundled) | From supplied transcript sequences | RPKM from transcript sequences |
| TopHatv 1.0.10 | Existing and novel gene quantitation | Yes | Yes | Bowtie | Predicted from transfrags From existing models | RPKM from supplied annotations |
| Velvet v0.7.47 | Short read assembler | Yes | No | NA | Assembled | Fold coverage |

Table 2. Software for RNA-Seq analysis as of 2009 (Pepke et al., 2009).

*Potential Medical and Clinical Applications of RNA-Seq Analysis*

RNA-Seq methods have already proven useful in addressing basic science questions, including the testing of dosage compensation from X chromosomes described above and the identification of novel transcripts, splice variants, and miRNA precursors (Mortazavi et al., 2008; Wilhelm et al., 2008; Xiong et al., 2010). In addition to these basic science applications, RNA-Seq has the potential, particularly as its cost drops in terms of both money and time required for sample preparation and analysis, to become a medically useful and clinically applicable technology. In disease classification and diagnosis, RNA-Seq could provide a powerful tool for high-resolution genomic analysis of tissues and cell populations to identify novel mutations and transcripts in cancers, to classify tumors based on gene expression patterns, or to identify microbial pathogens based on sequence identification. In realizing clinical possibilities, several barriers must be overcome, many of which are currently being addressed (see previous discussion of biological and computational aspects of RNA-Seq) (Ozsolak et al., 2011; Wang et al., 2009). Sample sizes required for RNA-Seq analysis must be small to accommodate rare and valuable patient tissue samples, efficient data analysis methods must exist for the effective interpretation of RNA-Seq results, and clinical personnel must be educated in the use of RNA-Seq datasets. The robustness of RNA-Seq datasets across experiments and the sensitivity of this method support the possibility for clinical use of RNA-Seq, particularly over hybridization-based gene expression analysis, but artifacts and errors in the sequencing procedure may lead to false identification of pathogenic transcripts, such as those encoding fusion proteins.

Several recent examples exist of nucleic acid detection for investigation of human phenotypic variations or for diagnosis of disease. In one case, RNA-Seq has allowed high-resolution inter-individual assessment of variation in gene expression levels at expression quantitative trait loci, which are believed

to underly some phenotypic diversity (Majewski and Pastinen, 2011). In a second example, investigators used maternal serum mRNA analysis to predict fetal aneuploidy at chromosome 21 by calculating expression levels of a single representative transcript from that chromosome (Lo et al., 2007). In that particular case, high-throughput analysis would permit both a global assessment of RNA levels and identification of aneuploidy of other chromosomes and the discovery of novel markers associated with such disease conditions. Gene expression thus is a useful marker for human phenotypes, suggesting that RNA-Seq could fill a role in medical and clinical fields.

Several recent examples of identification of novel mutations in tumor cell populations reveal the utility of RNA-Seq in disease classification. David Huntsman's group has applied RNA-Seq methodologies to the identification of mutations in gynecological tumors, identifying novel mutations in *FOXL2* in the previously ill-defined and treatment-resistant granulosa cell tumor of the ovary and in *ARID1A* in endometriosis-associated ovarian carcinomas (Shah et al., 2009; Wiegand et al., 2010). Diagnosis of granulosa cell tumors is difficult given the lack of knowledge of their pathogenesis and their relatively ambiguous histology, making a genetic variant associated with these tumors particularly valuable. Paired-end RNA-Seq analysis of several primary tumor samples identified mutations in the transcription factor *FOXL2* specifically in tumor cell populations (Shah et al., 2009). Another study used mRNA-seq of cell lines derived from tumors to discover somatic mutations associated with human activated B-cell-like diffuse large B-cell lymphoma, identifying *MYD88* as a candidate oncogenic mutation (Ngo et al., 2011). That result was used to focus studies of primary patient samples to rapidly identify mutations specific to this lymphoma subtype. These examples indicate the utility of RNA-Seq-based analysis of tumors to identify novel mutations. Because RNA-Seq provides full transcript sequences and is capable of assembling transcripts without relying on preexisting reference sequences (see computational methods above, Pepke et al., 2009; Martin et al., 2010), it can also analyze samples for fusion transcripts both known to be associated with cancer and previously unidentified (Wang et al., 2009; Ozsolak et al.,

2011). Additionally, its sensitivity and low background levels give RNA-Seq the ability to detect transcripts from single cells within populations, as might occur in analysis of patient blood samples for rare circulating tumor cells. Such circulating cells from solid tumors have already proven amenable to nucleic acid-based analysis, and extending RNA-Seq to identification and characterization of these cells would provide a highly sensitive and effective method for potentially diagnosis and monitoring of disease status (Helzer et al., 2009).

RNA-Seq-based profiling of patient samples could also be extended to microbiological diagnosis in cases of infection. Transcriptomic analyses by RNA-Seq have already been applied to several microbes, including *Salmonella*, *H. pylori,* and *M. tuberculosis* to reveal novel transcript and regulatory characteristics (Azhikina et al., 2010; Sharma et al., 2010; Güell et al., 2009; Perkins et al., 2009; Sittka et al., 2008). As RNA-Seq methods increase in speed and decrease in cost, sequence-based microbial diagnosis could become common and accurate. The capability to detect and properly classify transcripts from multiple organisms will be necessary for such applications, as will a prior knowledge of transcripts that uniquely identify certain microorganisms. From the perspective of translational science, RNA-Seq could be used to identify global changes in microbial populations within humans, such as the gut microbiome, or could be used to identify novel pathogens. Such methods would rely on metagenomic analysis and novel transcript discovery (Sorek and Cossart, 2010).

Despite the medical and clinical potential for RNA-Seq technologies, limitations to their application exist. Transcriptome-scale approaches to medical and clinical problems will still require linking of genetic information to phenotypes observed and validation of the clinical relevance of results from high-throughput sequencing datasets. The sensitivity of RNA-Seq could in some cases prove disadvantageous, as false positives could be called because of errors in sample preparation and sequencing or based on low-prevalence mutations that, although associated with disease states, may

not always cause disease (Ozsolak et al., 2011; Wang et al., 2009). Correlation of RNA-Seq outputs with phenotypic and clinical observations will be necessary for application of this technology.

*Conclusion*

RNA-based approaches may provide a simpler approach to global and unbiased gene analysis than whole-genome sequencing – given the reduced quantity of sequence data involved – while still remaining highly informative, providing both sequence and expression level data. As the speed improves, the ability to use smaller input quantities develops, and the cost drops, RNA-Seq is likely to become the preferred method for transcriptomic analysis. Its ability to generate absolute quantitative output about gene expression levels, its single-base resolution, and its potential for novel transcript identification and characterization demonstrate its superiority to microarray analysis. Additionally, methods for analysis of transcripts that do not require alignment to reference genomes and that permit the discovery of novel transcripts or the analysis of mixed cell populations will improve the ability of this technology to be applied to clinical problems. Appropriate tools for data analysis and efficient use of large volumes of output data in clinically relevant manners will be necessary for successful medical applications. The major conceptual advances have occurred to place RNA-Seq in line to become a major transcriptome analysis tool in both science and medicine, and primarily technical limitations must now be overcome to permit its widespread use.

## References


Azhikina, T., Skvortsov, T., Radaeva, T., Mardanov, A., Ravin, N., Apt, A., and Sverdlov, E. (2010). A new technique for obtaining whole pathogen transcriptomes from infected host tissues. BioTechniques *48*, 139-144.

Güell, M., van Noort, V., Yus, E., Chen, W., Leigh-Bell, J., Michalodimitrakis, K., Yamada, T., Arumugam, M., Doerks, T., Kühner, S., et al. (2009). Transcriptome complexity in a genome-reduced bacterium. Science *326*, 1268-1271.

Helzer, K. T., Barnes, H. E., Day, L., Harvey, J., Billings, P. R., and Forsyth, A. (2009). Circulating tumor cells are transcriptionally similar to the primary tumor in a murine prostate model. Cancer Res *69*, 7860-7866.

Ioannidis, J. P. A., Allison, D. B., Ball, C. A., Coulibaly, I., Cui, X., Culhane, A. C., Falchi, M., Furlanello, C., Game, L., Jurman, G., et al. (2009). Repeatability of published microarray gene expression analyses. Nat. Genet *41*, 149-155.

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

Lo, Y. M. D., Tsui, N. B. Y., Chiu, R. W. K., Lau, T. K., Leung, T. N., Heung, M. M. S., Gerovassili, A., Jin, Y., Nicolaides, K. H., Cantor, C. R., et al. (2007). Plasma placental RNA allelic ratio permits noninvasive prenatal chromosomal aneuploidy detection. Nat. Med *13*, 218-223.

Majewski, J., and Pastinen, T. (2011). The study of eQTL variations by RNA-seq: from SNPs to phenotypes. Trends Genet *27*, 72-79.

Martin, J., Bruno, V. M., Fang, Z., Meng, X., Blow, M., Zhang, T., Sherlock, G., Snyder, M., and Wang, Z. (2010). Rnnotator: an automated de novo transcriptome assembly pipeline from stranded RNA-Seq reads. BMC Genomics *11*, 663.

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

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

Ngo, V. N., Young, R. M., Schmitz, R., Jhavar, S., Xiao, W., Lim, K., Kohlhammer, H., Xu, W., Yang, Y., Zhao, H., et al. (2011). Oncogenically active MYD88 mutations in human lymphoma. Nature *470*, 115-119.

Ozsolak, F., and Milos, P. M. (2011). RNA sequencing: advances, challenges and opportunities. Nat. Rev. Genet *12*, 87-98.

Ozsolak, F., Platt, A. R., Jones, D. R., Reifenberger, J. G., Sass, L. E., McInerney, P., Thompson, J. F., Bowers, J., Jarosz, M., and Milos, P. M. (2009). Direct RNA sequencing. Nature *461*, 814-818.

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

Perkins, T. T., Kingsley, R. A., Fookes, M. C., Gardner, P. P., James, K. D., Yu, L., Assefa, S. A., He, M., Croucher, N. J., Pickard, D. J., et al. (2009). A strand-specific RNA-Seq analysis of the transcriptome of the typhoid bacillus Salmonella typhi. PLoS Genet *5*, e1000569.

Shah, S. P., Köbel, M., Senz, J., Morin, R. D., Clarke, B. A., Wiegand, K. C., Leung, G., Zayed, A., Mehl, E., Kalloger, S.

E., et al. (2009). Mutation of FOXL2 in Granulosa-Cell Tumors of the Ovary. New England Journal of Medicine *360*, 2719-2729.

Sharma, C. M., Hoffmann, S., Darfeuille, F., Reignier, J., Findeisz, S., Sittka, A., Chabas, S., Reiche, K., Hackermuller, J., Reinhardt, R., et al. (2010). The primary transcriptome of the major human pathogen Helicobacter pylori. Nature *464*, 250-255.

Sittka, A., Lucchini, S., Papenfort, K., Sharma, C. M., Rolle, K., Binnewies, T. T., Hinton, J. C. D., and Vogel, J. (2008). Deep sequencing analysis of small noncoding RNA and mRNA targets of the global post-transcriptional regulator, Hfq. PLoS Genet *4*, e1000163.

Sorek, R., and Cossart, P. (2010). Prokaryotic transcriptomics: a new view on regulation, physiology and pathogenicity. Nat Rev Genet *11*, 9-16.

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

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet *10*, 57-63.

Wiegand, K. C., Shah, S. P., Al-Agha, O. M., Zhao, Y., Tse, K., Zeng, T., Senz, J., McConechy, M. K., Anglesio, M. S., Kalloger, S. E., et al. (2010). ARID1A Mutations in Endometriosis-Associated Ovarian Carcinomas. New England Journal of Medicine *363*, 1532-1543.

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

Xiong, Y., Chen, X., Chen, Z., Wang, X., Shi, S., Wang, X., Zhang, J., and He, X. (2010). RNA sequencing shows no dosage compensation of the active X-chromosome. Nat. Genet *42*, 1043-1047.