Microsynth

RNA Sequencing

An introduction to efficient planning and execution of RNA sequencing (RNA-Seq) experiments.

Motivation for RNA Sequencing

RNA sequencing (RNA-Seq) refers to a method that is based on next generation sequencing (NGS) technologies to study transcriptomes. While microarray-based technologies depend on measuring hybridization intensities to predesigned probes, RNA-Seq relies on discrete counting of sequenced molecules. Thus, in contrast to microarrays, RNA-Seq requires no prior knowledge about the genome and can be considered hypothesis free. In addition, the counting of RNA molecules gives RNA-Seq experiments a much higher dynamic range compared to the hybridization intensities measured in microarray experiments. Also considering the falling costs of NGS technologies, it is not surprising that RNA-Seq experiments have become the gold standard in transcriptome studies. Novel isoforms, alternative splice sites, rare transcripts and gene-fusions, non-coding transcripts, and additional, even novel mechanism, can be detected all in a single experiment. Usually, the main goal of RNA-Seq studies is to obtain expression profiles, pathways and gene networks linked to the experimental condition studied. A generalized workflow of a typical RNA-Seq study is depicted in ***Figure 1***. Given the complex organization of genomes together with the huge amount of fragmented data, the interpretation of RNA-Seq experiments may appear a daunting task, especially for eukaryotic organisms [1]. For instance, the most recent human reference assembly – GRCh38 – has 244'550 unique exons with a mean length of 330 bases and a total amount of 80 million bases scattered across the 3 billion bases of the human genome. There are 2'879 annotated non-coding micro RNA (miRNA) [2], 52'000 transcripts from 26'475 genes, 22'302 associated gene ontology (GO) terms [3] and 330 KEGG pathways [4] (as of August 2018). Thus, a good experimental strategy is important to reliably identify the desired genes and gene networks, for example the transcription targets of a stressor, or a specific gene or pathway of interest after treatment, or gene-expression differences between different genotypes. This white paper will give an overview on how to handle RNA-Seq data by presenting a selection of workflows.

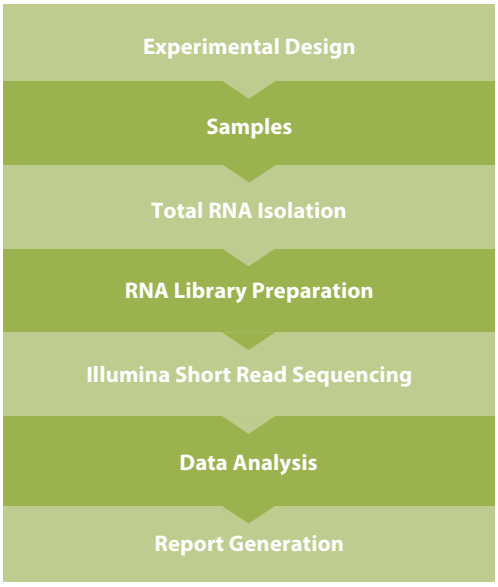

Figure 1. *This figure depicts a generalized RNA sequencing workflow that may be used for differential expression analysis.*

Microsynth AG, Switzerland
Schützenstrasse 15 · P.O. Box · CH-9436 Balgach · Phone +41 71 722 83 33 · Fax +41 71 722 87 58 · info@microsynth.ch · www.microsynth.ch

Sequencing and Differential Expression Analysis of Coding and Non-coding RNA

Selected Applications of RNA-Seq

Transcriptome studies are well suited to understand disease mechanisms, developmental mechanisms, or response to various stressors. Differential expression analysis of RNA-Seq data relies on the comparison of data sets obtained from experimental conditions (e.g. drug treatments) and controls to determine the difference in transcript abundance. The focus here is on messenger RNA (mRNA). In addition, non-coding miRNAs, which often have gene regulatory purposes, may be used to develop biomarkers specific to a medical condition. Such differentially expressed miRNA, can then be experimentally verified to develop diagnostic qPCR kits for instance.

Experimental Design

For a successful experiment, many aspects, including experimental setup, sampling, and funding are to be considered. In addition, the number of biological replicates and the number of reads produced for each replicate are essential parameters to produce valid results [5], especially to detect the maximal number of differentially expressed genes which includes rare transcripts. As gene expression analysis builds on counting reads from the respective transcription unit, single-end reads of 75 bases length usually suffice for accurate mapping. However, paired-end sequencing (and in some cases longer reads, for instance as produced by Pacific Biosciences (PacBio) sequencing technologies) is required if highly accurate transcript quantification, determination of gene fusions or novel splice variant detection is envisaged. In contrast to single-end sequencing, paired-end sequencing enables reading both ends of a (c)DNA fragment. Generally, it is recommended to work at least in triplicates per experimental condition and sequence 30 million single-end reads per replicate for eukaryotic organisms and 10 million single-end reads for each replicate for prokaryotic organisms. For miRNA the read numbers may be halved. It is also worth mentioning that the External RNA Controls Consortium (ERCC) has developed a set of external RNA controls designed to mimic natural eukaryotic mRNA sequences [6]. These sequences may be spiked in after RNA isolation and can be used to estimate the uncertainty in the subsequent measurements.

RNA Isolation

Obtaining high quality RNA is critical. RNA degradation is detrimental to the experiments since it may introduce 3′ biases during polyadenylation (polyA) enrichment or may distort the transcript profile by differentially affecting different RNAs. Thus, great care is needed to preserve the integrity of the input RNA. The acronym GIGO "garbage in, garbage out" holds true in this case as well. A notable exception is RNA extracted from formaldehyde-fixed paraffin-embedded (FFPE) tissue obtained by laser-assisted dissection methods, where a certain amount of degradation is unavoidable. Transcript profiles may also be distorted upon amplification of low amounts of input RNA, using either transcription-based or PCR-based amplification methods. However, with careful controls and a sufficient number of biological replicates, the adverse effects can be minimized.

Sequencing Library Construction

Depending on the desired RNA type (coding or non-coding) and type of organism studied (eukaryote or prokaryote), different sequencing library types are constructed. For instance, to sequence mRNA, a polyA enrichment step is performed for eukaryotes, while a ribosomal RNA depletion step is carried out for prokaryotes. Kits for non-coding RNA library construction may use alternative techniques to enrich the relevant RNA fraction. The constructed libraries are stranded, meaning they retain the strand information of the sequenced molecule, which results in a more reliable quantification of gene expression [7]. One typical method to keep the RNA strandness makes use of uracil instead of thymine for incorporation during second strand cDNA synthesis. After adapter ligation and before PCR amplification, uracil-DNA glycosylase (UNG) is added to degrade the second strand. As a result, all reads start in the same orientation, allowing the identification of the transcribed strand. A schematic depiction of how total RNA is turned into a sequenceable Illumina cDNA library is shown in ***Figure 2***.

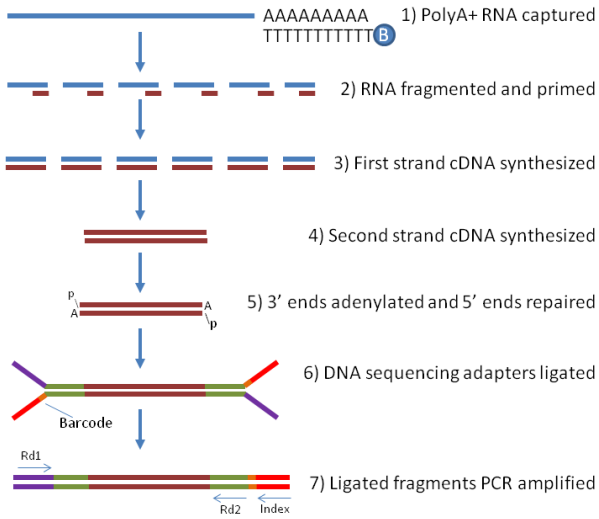

Figure 2. *Schematic description of a poly-A enriched RNA Illumina library ready for sequencing. Image: David Corney.*

Next Generation Sequencing

Illumina short-read sequencing by synthesis (SBS) technology, as depicted in ***Figure 3***, is especially well suited for RNA-Seq, as it is fast, accurate and cost effective [8]. Sequenced reads are produced in the standard fastq format [9] that incorporates both sequence information and quality scoring and can be further processed in downstream analyses.

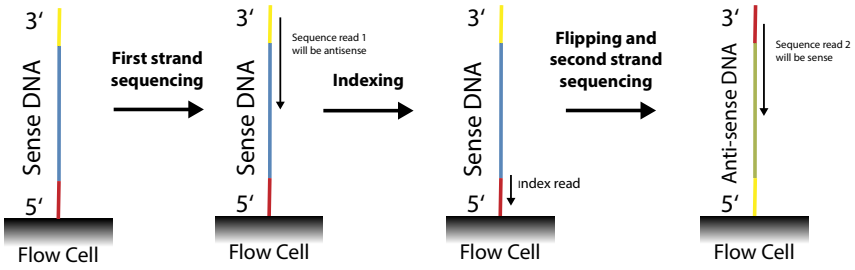

Figure 3: *Schematic of Illuminas paired-end sequencing workflow.*

Bioinformatics Analysis

As an example, let us consider an RNA-Seq experiment aimed at detecting changes in the gene expression profile of a human cell line after exposure to a drug. With control samples and additional samples taken at two different time points after drug application, each in three replicates, the total sample number would amount to nine. It is strictly recommended to not pool different replicates or conditions into a single sample for sequencing as this would eliminate all statistical power and the experiment would become useless. Assuming each sample in the outlined RNA-Seq experiment generates 30 million single-end, 75 bases long reads, the whole experiment produces 270 million reads or 20 billion bases. This large amount of data requires a dedicated analysis pipeline to extract meaningful information. Bioinformatics analysis of RNA-Seq data generally consists in: 1) quality control, optional size selection (e.g. to specifically separate non-coding RNA fractions) and filtering of the sequenced reads, 2) splice-aware mapping of the reads to the reference genome, 3) counting of uniquely mapped reads for each gene, 4) normalization of read counts across the experiment and 5) statistical evaluation of the normalized values comparing the different conditions (such as treatment and control) to each other to identify significant fold changes and up- or down-regulation of the genes [10]. ***Table 1*** presents an excerpt of a mRNA differential gene expression analysis. ***Figure 4*** shows how expression values of replicates group differ from the values of the respective controls.

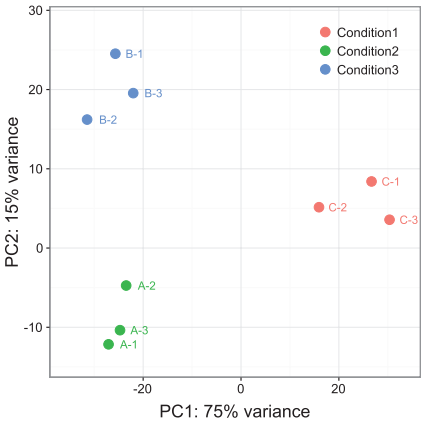

Figure 4. *Principal Component Analysis (PCA) plot to visualize grouping of samples in an RNA-Seq experiment. The three conditions depicted are clearly separated, indicating significant, differential gene expression patterns of the three analyzed conditions.*

Condition1 vs Condition2

5 records per page Search all columns:

ID	Image	From to logFC	From to p-Value	From to Adjusted p-Value
Palm3		-2.510	3.57e-13	7.39e-12
Masp1		-2.540	3.58e-13	7.39e-12
Dynap		-2.510	5.37e-13	1.10e-11
Gbp9		-5.990	6.35e-13	1.30e-11
Bdkrb2		-2.860	7.00e-13	1.42e-11

Table 1. *This excerpt of a table shows the main output of a differential gene expression analysis. In this experiment two conditions with three replicates are compared to each other. The table lists from left to right the gene identifier, boxplots representing expression level distributions of the replicates, the log_2 fold change of gene expression between condition 1 and condition 2, the propability value (p-value) of the log_2 fold change and the p-value adjusted for multiple testing.*

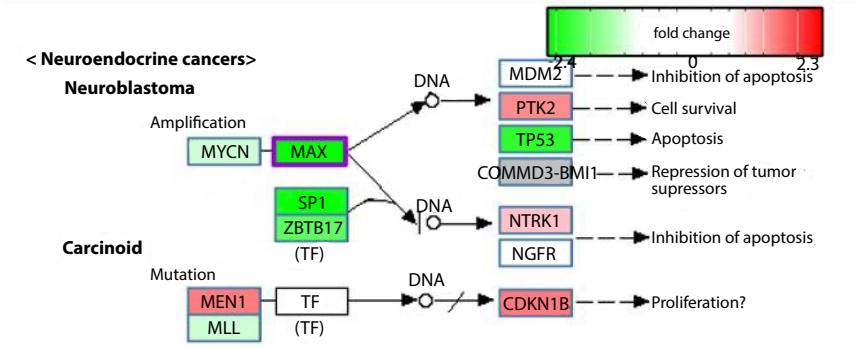

Figure 5. *Excerpt from the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway graph "TRANSCRIPTIONAL MISREGULATION IN CANCER", where colored nodes represent significantly up- or downregulated genes in the selected pathway.*

Based on the differential gene expression results and depending on the content of gene information published in databases, gene set and pathway analysis may be carried out to illuminate the larger context of the involved metabolic processes as exemplified in ***Figure 5***. A useful additional analysis in the case of miRNAs comprises a motif search to identify potential miRNA targets and to uncover additional, novel miRNAs [11]. The results of such analyses may be submitted to public databases such as miRNet [12] for further network-based visual analysis. ***Figure 6*** depicts such a motif identified by a miRNA analysis.

Figure 6. *A depiction of a significant de novo miRNA motif discovered in a miRNA Seq analysis. A miRNA motif is a region that is well conserved in many of the analyzed sequences.*

Summary

Obviously, RNA-Seq is not limited to dealing with questions of differential gene expression or identification of miRNA, which have been discussed in the previous sections. ***Table 2*** lists common RNA-Seq applications. The table can serve as a guide for selecting an appropriate approach to a research question. Another application of RNA Seq technology is, for example, *de novo* transcriptome assembly and annotation, which is useful when no annotated reference genome is available.
In short, RNA is collected from as many different stages and tissues as possible. The entire RNA is then enriched for polyadenylated mRNA. The pool of mRNA, which ideally represents all transcribed genes, is then normalized to reduce abundant mRNAs and enrich rare mRNAs. The normalized transcripts are sequenced, then assembled in a second step and annotated with various databases in a third step, resulting in a ready-to-use *de novo* transcriptome [13].
RNA-Seq provides a snapshot of the transcriptome in cells and cell populations, making it a very attractive and powerful method. However, the results of the RNA-Seq experiments are complex because they produce a large amount of fragmented data. However, with the right approach, the challenge of extracting knowledge is reduced to a manageable task.

Table 2. This table provides an overview of common scientifc questions in the field of RNA-Seq and gives a brief overview of the most important points that need to be considered in a RNA-Seq project. The table is intended as a quick reference guide.

Common Scientific Question	Explore the influence of a treatment on eukaryotic gene expression	Explore the influence of a treatment on prokaryotic gene expression	Develop biomarkers for specific medical conditions	Study cancer specific mechanisms	Study the transcriptome of a yet uncharted species
Analysis Method	Differential gene expression in **eukaryotes**	Differential gene expression in **prokaryotes**	**Non-coding RNA** differential expression analysis in eukaryotes	**Alternative splice-sites** and **gene-fusion** detection (novel isoforms)	***De novo*** transcriptome assembly
Experimental Setup	At least two conditions in replicates, **no pooling** of different conditions	At least two conditions in replicates, **no pooling** of different conditions	At least two conditions in replicates, **no pooling** of different conditions	At least two conditions in replicates, **no pooling** of different conditions	**Pooling** of different tissues, growth stadiums, etc. to capture the transcriptome in its entirety
Material and Resources	**mRNA** and availability of annotated reference genome	**mRNA** and availability of annotated reference genome	**e.g. miRNA** and availability of annotated non-coding RNA and reference genome	**mRNA** and availability of annotated reference Genome	**mRNA**, **missing** annotated reference Genome
Sample Preparation	Total RNA isolation; stranded **polyA enriched** sequencing library	Total RNA isolation; stranded **ribo-depleted** sequencing library	Total RNA isolation; **non-coding RNA enriched** sequencing library	Total RNA isolation; stranded **polyA enriched** sequencing library	Total RNA isolation; **normalized** mRNA sequencing library
Sequencing	**30 Mio single-end** reads, 75 bp length	**10 Mio single-end** reads, 75 bp length	**15 Mio single-end** reads, 75 bp length	**50 Mio paired-end** reads, 2 x 150 bp length	**20 Mio paired-end** reads, 2 x 300 bp length
Data Analysis	**Differential gene expression analysis**; pathway analysis if pathway database for the organism in question is available	**Differential gene expression analysis**; pathway analysis if pathway database for the organism in question is available	**Differential expression analysis**; motif search	**Statistical appraisal** of detected alternative splice sites and gene fusions	***De novo*** transcriptome assembly and annotation

References

[1] Steven L. Salzberg. Open questions: How many genes do we have? BMC Biology. 2018;16(94). doi:10.1186/s12915-018-0564-x

[2] Sam Griffiths-Jones, Russell J. Grocock, Stijn van Dongen, Alex Bateman, Anton J. Enright; miRBase: microRNA sequences, targets and gene nomenclature, Nucleic Acids Research, Volume 34, Issue suppl_1, 1 January 2006, Pages D140–D144, https://doi.org/10.1093/nar/gkj112

[3] The Gene Ontology Consortium; Expansion of the Gene Ontology knowledgebase and resources, Nucleic Acids Research, Volume 45, Issue D1, 4 January 2017, Pages D331–D338, https://doi.org/10.1093/nar/gkw1108

[4] Minoru Kanehisa, Miho Furumichi, Mao Tanabe, Yoko Sato, Kanae Morishima; KEGG: new perspectives on genomes, pathways, diseases and drugs, Nucleic Acids Research, Volume 45, Issue D1, 4 January 2017, Pages D353–D361, https://doi.org/10.1093/nar/gkw1092

[5] Schurch NJ, Schofield P, Gierliński M, et al. How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? RNA. 2016;22(6):839-851. doi:10.1261/rna.053959.115

[6] Lemire A, Lea K, Batten D, et al. Development of ERCC RNA Spike-In Control Mixes. Journal of Biomolecular Techniques : JBT. 2011;22(Suppl):S46

[7] Zhao S, Zhang Y, Gordon W, et al. Comparison of stranded and non-stranded RNA-seq transcriptome profiling and investigation of gene overlap. BMC Genomics. 2015;16(1):675. doi:10.1186/s12864-015-1876-7

[8] Online at: https://emea.illumina.com/systems/sequencing-platforms/nextseq/applications.html?langsel=/ch/, accessed 14.09.2018

[9] Online at: http://maq.sourceforge.net/fastq.shtml, accessed 14.09.2018

[10] Sandrine Borgeaud, Lisa C. Metzger, Tiziana Scrignari, Melanie Blokesch, The type VI secretion system of Vibrio cholerae fosters horizontal gene transfer, Science 02 Jan 2015: Vol. 347, Issue 6217, pp. 63-67. DOI: 10.1126/science.1260064

[11] Bhupesh K. Prusty, Nitish Gulve, Suvagata Roy Chowdhury, Michael Schuster, Sebastian Strempel, Vincent Descamps, Thomas Rudel. HHV-6 encoded small non-coding RNAs define an intermediate and early stage in viral reactivation. npj Genomic Medicine. 2018;3(25). 10.1038/s41525-018-0064-5

[12] Yannan Fan, Keith Siklenka, Simran K. Arora, Paula Ribeiro, Sarah Kimmins, Jianguo Xia; miRNet - dissecting miRNA-target interactions and functional associations through network-based visual analysis, Nucleic Acids Research, Volume 44, Issue W1, 8 July 2016, Pages W135–W141, https://doi.org/10.1093/nar/gkw288

[13] Neves, R.C., Guimaraes, J.C., Strempel, S. et al., Transcriptome profiling of Symbion pandora (phylum Cycliophora): insights from a differential gene expression analysis, Org Divers Evol (2017) 17: 111. https://doi.org/10.1007/s13127-016-0315-1