Methods of construction of a Transcriptome

Transcriptomics is the quantitative science that encompasses the assignment of a list of strings ("reads") to the object ("transcripts" in the genome). To calculate the expression strength, the density of reads corresponding to each object is counted. Initially, transcriptomes were analysed and studied using expressed sequence tags libraries and serial and cap analysis of gene expression (SAGE).

Currently, the two main transcriptomics techniques include DNA microarrays and RNA-Seq. Both techniques require RNA isolation through RNA extraction techniques, followed by its separation from other cellular components and enrichment of mRNA.

There are two general methods of inferring transcriptome sequences. One approach maps sequence reads onto a reference genome, either of the organism itself (whose transcriptome is being studied) or of a closely related species. The other approach, *de novo* transcriptome assembly, uses software to infer transcripts directly from short sequence reads and is used in organisms with genomes that are not sequenced

DNA microarrays (this is just an intro: more on will be on lecture 7)
The first transcriptome studies were based on microarray techniques (also known as DNA chips or biochip). Microarrays consist of thin glass layers with spots on which oligonucleotides, known as "probes" are arrayed; each spot contains a known DNA sequence.

When performing microarray analyses, mRNA is collected from a control and an experimental sample, the latter usually representative of a disease. The RNA of interest is converted to cDNA to increase its stability and marked with fluorophores of two colours, usually green and red, for the two groups. The cDNA is spread onto the surface of the microarray where it hybridizes with oligonucleotides on the chip and a laser is used to scan. The fluorescence intensity on each spot of the microarray corresponds to the level of gene expression and based on the colour of the fluorophores selected, it can be determined which of the samples exhibits higher levels of the mRNA of interest.

One microarray usually contains enough oligonucleotides to represent all known genes; however, data obtained using microarrays does not provide information about unknown

genes. During the 2010s, microarrays were almost completely replaced by next-generation techniques that are based on DNA sequencing.

RNA-SEQUENCING (Seq) METHODS FOR SPECIFIC GOALS

Tag-Based Methods for Gene Expression Profiling

Digital gene expression, DGE-Seq

Digital gene expression (DGE)-Seq, or Tag-Seq, is a deep sequencing method derived from SAGE. As in SAGE, the method involves attachment of mRNA to beads via the poly(A) tail, first and second strand cDNA syntheses on the beads, and digestion of double-stranded cDNA with a frequent cutting restriction enzyme. The remaining 3′ fragment attached to the beads is then ligated to its 5′ end adapter with a recognition site for another restriction enzyme, called tagging enzyme. The tagging enzyme cleaves the cDNA and generates a short 21 bp tag, which is then ligated to a second adapter at its 3′ end. The cDNA is then amplified by PCR, followed by sequencing. Because only a short tag is sequenced from the whole transcript, DGE-Seq is more economical than traditional RNA-Seq for a given depth of sequencing and can provide a higher dynamic range of detection when the same number of reads is generated. By design, DGE-Seq preserves RNA strandedness. This method has been commercialized by several companies and is useful especially when simple gene expression profiling is the goal. It is also the method of choice when the complete genome or transcriptome is not available for full alignment of RNA-Seq reads.

3′ End Sequencing

A number of methods specifically sequence the 3′ end region of transcripts. Most of these methods were first developed to interrogate alterative cleavage and polyadenylation sites, a widespread phenomenon in all eukaryotes. As in DGE-Seq, the data from these methods can also be used to study gene expression.

In essence, DGE-Seq and 3′ end sequencing are tag-based approaches that use one fragment to represent a transcript. While efficient for gene expression analysis, they can have higher variability than ‘shotgun’ style RNA-Seq, where one transcript is represented by multiple fragments. Biases from fragmentation, adapter ligation and PCR can make tag-based data more prone to batch effects.

Sequencing to Reveal Alternative Splicing and Gene Fusion

Almost all multi-exon genes display alternative splicing (AS). AS plays an important role in regulation of cellular processes, and aberrations of the process are associated with many human diseases. A more direct approach to examine AS is to sequence the exon-exon junction region directly. Using oligo pairs targeted to specific exon-exon junction sequences, the Fu lab developed RASL-Seq, which provides analysis of specific splice junction regions. However, prior knowledge of the exon–exon junction sequences is required for the oligo design.

Similar to splicing, gene fusion events can place two noncontinuous genomic regions together in a single transcript. Created by chromosomal rearrangements, gene fusions are present in approximately 20% of cancer. Fusion events can be detected using RNA-Seq data along with specific bioinformatic methods. Detection of a fusion event is typically revealed by reads containing fusion junctions or by differences in expression between the 5′ and 3′ ends of genes that are fused. Regular RNA-Seq methods are typically not sufficiently sensitive to detect fusion junctions. Several methods have been developed, including

(1) enrichment of RNA-Seq reads for genes of interest,

(2) exon capture, and

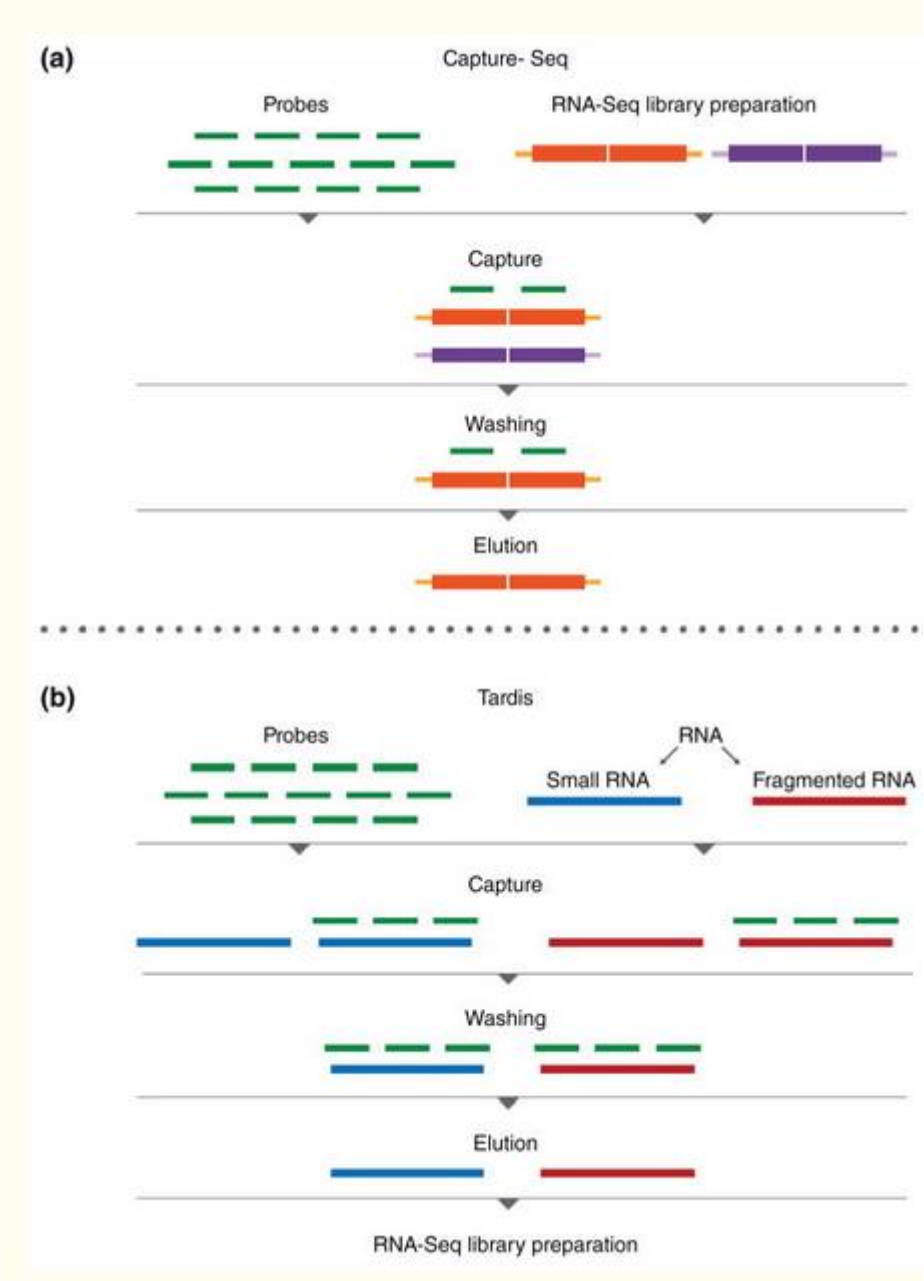
(3) amplicon sequencing.

The ultimate solution to unravel the complexity of alternatively splicing and gene fusion isoforms is to sequence each transcript from the beginning to the end. Two strategies have been established to this end. Single-molecule real-time sequencing (SMRT) on the PacBio sequencing platform offers long reads up to 5 kb. However, this method is costly, and has a high error-rate and low multiplexing capacity. Hybrid sequencing methods have been introduced which combine SMRT data with short, standard RNA-Seq reads.

Targeted RNA-Seq

Selection a specific set of transcripts for sequencing is often desirable when a defined group of genes is of interest. Lowly expressed genes that cannot be readily analysed using whole transcriptome sequencing can also be detected using targeted RNA-Seq methods. Two

general approaches have been used, namely, target capture and amplicon sequencing. The target capture approach involves selection of specific genes using a set of biotinylated probes which bind cDNA, or RNA (Figure Below). By contrast, amplicon sequencing employs gene-specific primers for the amplification of cDNA targets (Figure below). Approaches differ in the amplicon design, including two specific primers after cDNA synthesis by template switch, nested PCR with one specific primer and one common primer for adapter, and specific targeting primers in combination with a primer for poly(A) tail priming.



RNA sequencing

RNA sequencing is a next-generation sequencing technology; as such it requires only a small amount of RNA and no previous knowledge of the genome. It allows for both qualitative and quantitative analysis of RNA transcripts, the former allowing discovery of new transcripts and the latter a measure of relative quantities for transcripts in a sample.

The three main steps of sequencing transcriptomes of any biological samples include RNA purification, the synthesis of an RNA or cDNA library and sequencing the library. The RNA purification process is different for short and long RNAs. This step is usually followed by an assessment of RNA quality, with the purpose of avoiding contaminants such as DNA or technical contaminants related to sample processing. RNA quality is measured using UV spectrometry with an absorbance peak of 260 nm. RNA integrity can also be analysed quantitatively comparing the ratio and intensity of 28S RNA to 18S RNA reported in the RNA Integrity Number (RIN) score. Since mRNA is the species of interest and it represents only 3% of its total content, the RNA sample should be treated to remove rRNA and tRNA and tissue-specific RNA transcripts.

The step of library preparation with the aim of producing short cDNA fragments, begins with RNA fragmentation to transcripts in length between 50 and 300 base pairs. Fragmentation can be enzymatic (RNA endonucleases), chemical (trismagnesium salt buffer, chemical hydrolysis) or mechanical (sonication, nebulisation). Reverse transcription is used to convert the RNA templates into cDNA and three priming methods can be used to achieve it, including oligo-DT, using random primers or ligating special adaptor oligos.

<u>Single-cell transcriptomics</u>

Transcription can also be studied at the level of individual cells by single-cell transcriptomics. Single-cell RNA sequencing (scRNA-seq) is a recently developed technique that allows the analysis of the transcriptome of single cells. With single-cell transcriptomics, subpopulations of cell types that constitute the tissue of interest are also taken into consideration. This approach allows to identify whether changes in experimental samples are due to phenotypic cellular changes as opposed to proliferation, with which a specific cell type might be overexpressed in the sample. Additionally, when assessing cellular progression through differentiation, average expression profiles are only able to order cells by time rather

than their stage of development and are consequently unable to show trends in gene expression levels specific to certain stages. Single-cell trarnscriptomic techniques have been used to characterize rare cell populations such as circulating tumor cells, cancer stem cells in solid tumors, and embryonic stem cells (ESCs) in mammalian blastocysts.

Although there are no standardized techniques for single-cell transcriptomics, several steps need to be undertaken. The first step includes cell isolation, which can be performed using low- and high-throughput techniques. This is followed by a qPCR step and then single-cell RNAseq where the RNA of interest is converted into cDNA. Newer developments in single-cell transcriptomics allow for tissue and sub-cellular localization preservation through cryo-sectioning thin slices of tissues and sequencing the transcriptome in each slice. Another technique allows the visualization of single transcripts under a microscope while preserving the spatial information of each individual cell where they are expressed.

SEQUENCING OF OTHER RNA SPECIES

Small RNAs

Small non-coding RNAs below 30 nt, such as miRNAs, piRNAs, and endosiRNAs, are processed from primary transcripts. miRNAs can be efficiently captured by direct ligation with adapters (see above). Thanks to the 5′ phosphate group and 3′ OH group of miRNAs, no additional processing of RNA is necessary before the ligations. As discussed above, this method naturally preserves the strandedness of RNA but introduces substantial biases due to influence of sequence on ligation. Using degenerate random nucleotides at the ligation ends of adapters can effectively mitigate bias. Alternatively, the 5′ adapter ligation step, which appears to be more prone to bias than the 3′ adapter ligation (personal observation), can be eliminated if the single-stranded cDNA is circularized by DNA ligase and amplified by PCR.

Because of the short insert size of the cDNA library for small RNAs, it is necessary to use a specific method to separate them from contaminant DNAs, such as PCR products without any inserts. These methods include electrophoresis separation or adding the RT primer to block the 3′ adapter from ligating with the 5′ adapter.

Circular RNA

One recent surprising finding has been the discovery of circular RNAs, which are generated by back-splicing. Circular RNAs can be sequenced by digesting away linear RNAs using exonuclease R, followed by regular RNA-Seq methods involving fragmentation, RT and PCR.

CITED WORKS

Morozova, O., Hirst, M., Marra, M.A., 2009. Applications of new sequencing technologies for transcriptome analysis. Annual Review of Genomics & Human Genetics 10, 135–151.

Pietu, G., Mariage-Samson, R., Fayein, N.A., et al., 1999. The genexpress IMAGE knowledge base of the human brain transcriptome: A prototype integrated resource for functional and computational genomics. Genome Research 9, 195–209.

Fundamentals of Advanced Omics Technologies: From Genes to Metabolites (Comprehensive Analytical Chemistry) 1st Edition, Kindle Edition by Carolina Simó, Alejandro Cifuentes, Virginia García-Cañas

Genomics, Proteomics and Metabolomics in Nutraceuticals and Functional Foods (Hui: Food Science and Technology) 2nd, Kindle Edition by Debasis Bagchi, Anand Swaroop, Manashi Bagchi

Metabolome Analyses: Strategies for Systems Biology 2005th Edition by Seetharaman Vaidyanathan, George G. Harrigan, Royston Goodacre

Peralta, Mihaela (2012). "The Human Transcriptome: An Unfinished Story". Genes. 3 (3): 344–360.

Wang, Zhong; Gerstein, Mark; Snyder, Michael (January 2009). "RNA-Seq: a revolutionary tool for transcriptomics

Jiménez-Chillarón, Josep C.; Díaz, Rubén; Ramón-Krauel, Marta (2014). "Chapter 4 - Omics Tools for the Genome-Wide Analysis of Methylation and Histone Modifications". Comprehensive Analytical Chemistry

GK, Sim; FC, Kafatos; CW, Jones; MD, Koehler; A, Efstratiadis; T., Maniatis (December 1979). "Use of a cDNA library for studies on evolution and developmental expression of the chorion multigene families

E Velculescu, Victor; Zhang, Lin; Zhou, Wei; Vogelstein, Jacob; A Basrai, Munira; E Bassett Jr., Douglas; Hieter, Phil; Vogelstein, Bert; W Kinzler, Kenneth (1997). "Characterization of the Yeast Transcriptome"

Rhoades RA, Pflanzer RG (2002). Human Physiology (5th ed.). Thomson Learning. p. 584. ISBN 978-0-534-42174-8.

Janeway C (2001). Immunobiology (5th ed.). Garland Publishing.

Borghesi L, Milcarek C (2006). "From B cell to plasma cell: regulation of V(D)J recombination and antibody secretion". Immunologic Research. 36 (1–3): 27–32. doi:10.1385/IR:36:1:27. PMID 17337763. S2CID 27041937.

Pier GB, Lyczak JB, Wetzler LM (2004). Immunology, Infection, and Immunity. ASM Press.

Asmann YW, Klee EW, Thompson EA, Perez EA, Middha S, Oberg AL, Therneau TM, Smith DI, Poland GA, Wieben ED, et al. 3′ tag digital gene expression profiling of human brain and universal reference RNA using Illumina Genome Analyzer. BMC Genomics. 2009;10:531. doi: 10.1186/1471-2164-10-531. [PMC free article] [PubMed]

Morrissy AS, Morin RD, Delaney A, Zeng T, McDonald H, Jones S, Zhao Y, Hirst M, Marra MA. Next-generation tag sequencing for cancer gene expression profiling. Genome Res. 2009;19:1825–1835. doi: 10.1101/gr.094482.109. [PMC free article] [PubMed]

Chen EQ, Bai L, Gong DY, Tang H. Employment of digital gene expression profiling to identify potential pathogenic and therapeutic targets of fulminant hepatic failure. J Transl Med. 2015;13:22. doi: 10.1186/s12967-015-0380-9. [PMC free article] [PubMed]

Tian B, Manley JL. Alternative cleavage and polyadenylation: the long and short of it. Trends Biochem Sci. 2013;38:312–320. doi: 10.1016/j.tibs.2013.03.005. [PMC free article] [PubMed]

Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB. Alternative isoform regulation in human tissue transcriptomes. Nature. 2008;456:470–476. doi: 10.1038/nature07509. [PMC free article] [PubMed]

Wang GS, Cooper TA. Splicing in disease: disruption of the splicing code and the decoding machinery. Nat Rev Genet. 2007;8:749–761. doi: 10.1038/nrg2164. [PubMed]

Hafner M, Landgraf P, Ludwig J, Rice A, Ojo T, Lin C, Holoch D, Lim C, Tuschl T. Identification of microRNAs and other small regulatory RNAs using cDNA library sequencing. Methods. 2008;44:3–12. doi: 10.1016/j.ymeth.2007.09.009. [PMC free article] [PubMed]

Jayaprakash AD, Jabado O, Brown BD, Sachidanandam R. Identification and remediation of biases in the activity of RNA ligases in small-RNA deep sequencing. Nucleic Acids Res. 2011;39:e141. doi: 10.1093/nar/gkr693. [PMC free article] [PubMed]

Sun G, Wu X, Wang J, Li H, Li X, Gao H, Rossi J, Yen Y. A bias-reducing strategy in profiling small RNAs using Solexa. RNA. 2011;17:2256–2262. doi: 10.1261/rna.028621.111. [PMC free article] [PubMed]