Transcriptomes

Transcriptomics.

Transcriptomics is the study of the transcriptome—the complete set of RNA transcripts that are produced by the genome, under specific circumstances or in a specific cell—using high-throughput methods, such as microarray analysis. Comparison of transcriptomes allows the identification of genes that are differentially expressed in distinct cell populations, or in response to different treatments.

The **transcriptome** is the set of all RNA transcripts, including coding and non-coding, in an individual or a population of cells. The term can also sometimes be used to refer to all RNAs, or just mRNA, depending on the particular experiment. The term *transcriptome* is a portmanteau of the words *transcript* and *genome*; it is associated with the process of transcript production during the biological process of transcription.

The early stages of transcriptome annotations began with cDNA libraries published in the 1980s. Subsequently, the advent of high-throughput technology led to faster and more efficient ways of obtaining data about the transcriptome. Two biological techniques are used to study the transcriptome, namely DNA microarray, a hybridization-based technique and RNA-seq, a sequence-based approach. RNA-seq is the preferred method and has been the dominant transcriptomics technique since the 2010s. Single-cell transcriptomics allows tracking of transcript changes over time within individual cells. Data obtained from the transcriptome is used in research to gain insight into processes such as cellular differentiation, carcinogenesis, transcription regulation and biomarker discovery among others. Transcriptome-obtained data also finds applications in establishing phylogenetic relationships during the process of evolution and in *in vitro* fertilization. The transcriptome is closely related to other -ome based biological fields of study; it is complementary to the proteome and the metabolome and encompasses the translatome, exome, meiome and thanatotranscriptome which can be seen as ome fields studying specific types of RNA transcripts. There are numerous publicly available transcriptome databases.

Applications of Transcriptomics

- Sequencing and Next-Generation Sequencing
- Genotyping

- Copy number variation (CNV)
- Methylation Analysis
- Gene Expression
- miRNA Analysis
- ChIP-on-chip (Chromatin Immunoprecipitation)
- Metagenomics: Bacteria and Fungus Identification.

Etymology and history

The word *transcriptome* is a portmanteau of the words *transcript* and *genome*. It appeared along other neologisms formed using the suffixes *-ome* and *-omics* to denote all studies conducted on a genome-wide scale in the fields of life sciences and technology. As such, transcriptome and transcriptomics were one of the first words to emerge along with genome and proteome. The first study to present a case of a collection of a cDNA library for silk moth mRNA was published in 1979. With the rise of high-throughput technologies and bioinformatics and the subsequent increased computational power, it became increasingly efficient and easy to characterize and analyze enormous amount of data. Attempts to characterize the transcriptome became more prominent with the advent of automated DNA sequencing during the 1980s. During the 1990s, expressed sequence tag sequencing was used to identify genes and their fragments. This was followed by techniques such as serial analysis of gene expression (SAGE), cap analysis of gene expression (CAGE), and massively parallel signature sequencing (MPSS).

Transcription

The transcriptome encompasses all the ribonucleic acid (RNA) transcripts present in a given organism or experimental sample. RNA is the main carrier of genetic information that is responsible for the process of converting DNA into an organism's phenotype. A gene can give rise to a single-stranded messenger RNA (mRNA) through a molecular process known as transcription; this mRNA is complementary to the strand of DNA it originated from. The enzyme RNA polymerase II attaches to the template DNA strand and catalyses the addition of ribonucleotides to the 3' end of the growing sequence of the mRNA transcript.

In order to initiate its function, RNA polymerase II needs to recognize a promoter sequence, located upstream (5') of the gene. In eukaryotes, this process is mediated by transcription

factors, most notably Transcription factor II D (TFIID) which recognizes the TATA box and aids in the positioning of RNA polymerase at the appropriate start site. To finish the production of the RNA transcript, termination takes place usually several hundred nuclecotides away from the termination sequence and cleavage takes place. This process occurs in the nucleus of a cell along with RNA processing by which mRNA molecules are capped, spliced and polyadenylated to increase their stability before being subsequently taken to the cytoplasm. The mRNA gives rise to proteins through the process of translation that takes place in ribosomes.

Types of RNA transcripts

In accordance with the central dogma of molecular biology, the transcriptome initially encompassed only protein-coding mRNA transcripts. Nevertheless, several RNA subtypes with distinct functions exist. Many RNA transcripts do not code for protein or have different regulatory functions in the process of gene transcription and translation. RNA types which do not fall within the scope of the central dogma of molecular biology are non-coding RNAs which can be divided into two groups of long non-coding RNA and short non-coding RNA. Long non-coding RNA includes all non-coding RNA transcripts that are more than 200 nucleotides long. Members of this group comprise the largest fraction of the non-coding transcriptome.

Short non-coding RNA includes the following members:

- transfer RNA (tRNA)
- micro RNA (miRNA): 19-24 nucleotides (nt) long. Micro RNAs up- or downregulate expression levels of mRNAs by the process of RNA interference at the post-transcriptional level.
- small interfering RNA (siRNA): 20-24 nt
- small nucleolar RNA (snoRNA)
- Piwi-interacting RNA (piRNA): 24-31 nt. They interact with Piwi proteins of the Argonaute family and have a function in targeting and cleaving transposons.
- enhancer RNA (eRNA)

In the human genome, about 5% of all genes get transcribed into RNA. The transcriptome consists of coding mRNA which comprise around 1-4% of its entirety and non-coding RNAs

which comprise the rest of the genome and do not give rise to proteins. The number of non-protein-coding sequences increases in more complex organisms. Several factors render the content of the transcriptome difficult to establish. These include alternative splicing, RNA editing and alternative transcription among others. Additionally, transcriptome techniques are capable of capturing transcription occurring in a sample at a specific time point, although the content of the transcriptome can change during differentiation. The main aims of transcriptomics are the following:

- catalogue all species of transcript, including mRNAs, non-coding RNAs and small RNAs;
- to determine the transcriptional structure of genes, in terms of their start sites, 5′ and 3′ ends, splicing patterns and other post-transcriptional modifications; and
- to quantify the changing expression levels of each transcript during development and under different conditions.

The term can be applied to the total set of transcripts in a given organism, or to the specific subset of transcripts present in a particular cell type. Unlike the genome, which is roughly fixed for a given cell line (excluding mutations), the transcriptome can vary with external environmental conditions. Because it includes all mRNA transcripts in the cell, the transcriptome reflects the genes that are being actively expressed at any given time, with the exception of mRNA degradation phenomena such as transcriptional attenuation. The study of transcriptomics, (which includes expression profiling, splice variant analysis etc), examines the expression level of RNAs in a given cell population, often focusing on mRNA, but sometimes including others such as tRNAs and sRNAs.

Toxicogenomics.

Toxicogenomics is a field of science that deals with the collection, interpretation, and storage of information about gene and protein activity within particular cell or tissue of an organism in response to toxic substances. Toxicogenomics combines toxicology with genomics or other high throughput molecular profiling technologies such as transcriptomics, proteomics and metabolomics. Toxicogenomics endeavours to elucidate molecular mechanisms evolved in the expression of toxicity, and to derive molecular expression patterns (i.e., molecular biomarkers) that predict toxicity or the genetic susceptibility to it. In pharmaceutical research toxicogenomics is defined as the study of the structure and function of the genome as it responds to adverse xenobiotic exposure. It is the toxicological subdiscipline of

pharmacogenomics, which is broadly defined as the study of inter-individual variations in whole-genome or candidate gene single-nucleotide polymorphism maps, haplotype markers, and alterations in gene expression that might correlate with drug responses Though the term toxicogenomics first appeared in the literature in 1999. It was already in common use within the pharmaceutical industry as its origin was driven by marketing strategies from vendor companies. The term is still not universal accepted, and others have offered alternative terms such as chemogenomics to describe essentially the same area. The nature and complexity of the data (in volume and variability) demands highly developed processes of automated handling and storage. The analysis usually involves a wide array of bioinformatics and statistics, regularly involving classification approaches. In pharmaceutical drug discovery and development toxicogenomics is used to study adverse, i.e., toxic, effects, of pharmaceutical drugs in defined model systems in order to draw conclusions on the toxic risk to patients or the environment.

Human Metabolome Project

The Human Metabolome Project was a Genome Canada funded project launched in January 2005. We are discussing about this project as it was a big milestone to the metabolome research. The purpose of the project was and still is to facilitate metabolomics research through several objectives to improve disease identification, prognosis and monitoring; provide insight into drug metabolism and toxicology; provide a linkage between the human metabolome and the human genome; and to develop software tools for metabolomics.
The project mandate is to identify, quantify, catalogue and store all metabolites that can potentially be found in human tissues and biofluids at concentrations greater than one micromolar. This data will be freely accessible in an electronic format to all researchers through the Human Metabolome Database (www.hmdb.ca). In addition, all compounds are publicly available through our Human Metabolome Library (www.metabolibrary.ca).
Already more than 800 compounds had been identified and by end of 2006, it is expected that more than 1400 metabolites will have been identified, quantified and archived into web accessible databases (www.hmdb.ca) and stored in -80°C freezers. However, the Human Metabolome Project is only mandated to provide chemical data and chemical compounds to the scientific community. It does not have the funding or the resources to use these "raw materials" for disease identification and characterization. Indeed, the intent of the Human Metabolome Project is to be an enabler of future metabolomic research, just as the Human

Genome Project has been an enabler of current genomic research. This research has been widely used in metabolomics, clinical chemistry, biomarker discovery and general biochemistry education.

Interactomics

In molecular biology, an interactome is the whole set of molecular interactions in a particular cell. The term specifically refers to physical interactions among molecules (such as those among proteins, also known as protein-protein interactions) but can also describe sets of indirect interactions among genes (genetic interactions). Mathematically, interactomes are generally displayed as graphs. The word "interactome" was originally coined in 1999 by a group of French scientists headed by Bernard Jacq.

Though interactomes may be described as biological networks, they should not be confused with other networks such as neural networks or food webs Interactomics is a discipline at the intersection of bioinformatics and biology that deals with studying both the interactions and the consequences of those interactions between and among proteins, and other molecules within a cell. Interactomics thus aims to compare such networks of interactions (i.e., interactomes) between and within species in order to find how the traits of such networks are either preserved or varied. Interactomics is an example of "top-down" systems biology, which takes an overhead, as well as overall, view of a biosystem or organism. Large sets of genome-wide and proteomic data are collected, and correlations between different molecules are inferred. From the data new hypotheses are formulated about feedbacks between these molecules. These hypotheses can then be tested by new experiments

Metabolic Network

A metabolic network is the complete set of metabolic and physical processes that determine the physiological and biochemical properties of a cell. As such, these networks comprise the chemical reactions of metabolism, the metabolic pathways, as well as the regulatory interactions that guide these reactions. With the sequencing of complete genomes, it is now possible to reconstruct the network of biochemical reactions in many organisms, from bacteria to human. Several of these networks are available online: Kyoto Encyclopedia of Genes and Genomes (KEGG), EcoCyc, BioCyc and metaTIGER .

Metabolic networks are powerful tools for studying and modelling metabolism.
Metabolic network reconstruction and simulation allows for an in-depth insight into the

molecular mechanisms of a particular organism. In particular, these models correlate the genome with molecular physiology. A reconstruction breaks down metabolic pathways (such as glycolysis and the Citric acid cycle) into their respective reactions and enzymes, and analyses them within the perspective of the entire network. In simplified terms, a reconstruction collects all of the relevant metabolic information of an organism and compiles it in a mathematical model. Validation and analysis of reconstructions can allow identification of key features of metabolism such as growth yield, resource distribution, network robustness, and gene essentiality. This knowledge can then be applied to create novel biotechnology.

In general, the process to build a reconstruction is as follows:

1. Draft a reconstruction
2. Refine the model
3. Convert model into a mathematical/computational representation
4. Evaluate and debug model through experimentation.

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]