Data Analysis

A number of organism-specific transcriptome databases have been constructed and annotated to aid in the identification of genes that are differentially expressed in distinct cell populations. RNA-seq is emerging (2013) as the method of choice for measuring transcriptomes of organisms, though the older technique of DNA microarrays is still used. RNA-seq measures the transcription of a specific gene by converting long RNAs into a library of cDNA fragments. The cDNA fragments are then sequenced using high-throughput sequencing technology and aligned to a reference genome or transcriptome which is then used to create an expression profile of the genes.

Applications

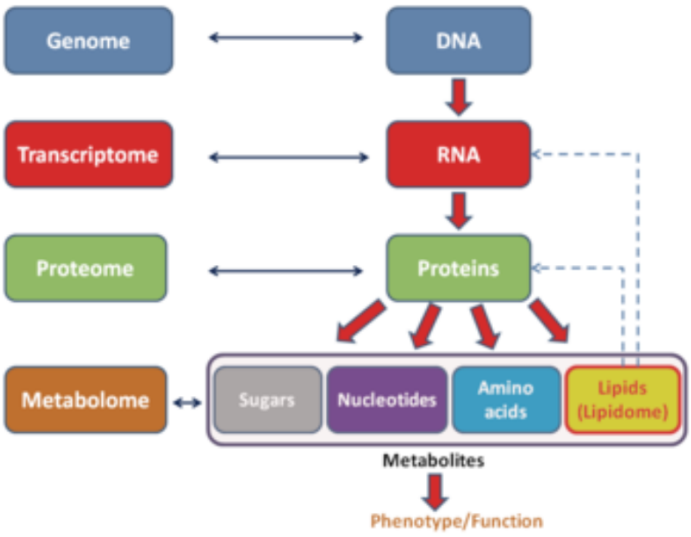
Mammals

The transcriptomes of stem cells and cancer cells are of particular interest to researchers who seek to understand the processes of cellular differentiation and carcinogenesis. A pipeline using RNA-seq or gene array data can be used to track genetic changes occurring in stem and precursor cells and requires at least three independent gene expression data from the former cell type and mature cells. Analysis of the transcriptomes of human oocytes and embryos is used to understand the molecular mechanisms and signaling pathways controlling early embryonic development, and could theoretically be a powerful tool in making proper embryo selection in in vitro fertilisation. Analyses of the transcriptome content of the placenta in the first-trimester of pregnancy in *in vitro* fertilization and embryo transfer (IVT-ET) revealed differences in genetic expression which are associated with higher frequency of adverse perinatal outcomes. Such insight can be used to optimize the practice. Transcriptome analyses can also be used to optimize cryopreservation of oocytes, by lowering injuries associated with the process. Transcriptomics is an emerging and continually growing field in biomarker discovery for use in assessing the safety of drugs or chemical risk assessment.
Transcriptomes may also be used to infer phylogenetic relationships among individuals or to detect evolutionary patterns of transcriptome conservation. Transcriptome analyses were used to discover the incidence of antisense transcription, their role in gene expression through interaction with surrounding genes and their abundance in different chromosomes. RNA-seq was also used to show how RNA isoforms, transcripts stemming from the same gene but with different structures, can produce complex phenotypes from limited genomes.

Plants

Transcriptome analysis have been used to study the evolution and diversification process of plant species. In 2014, the 1000 Plant Genomes Project was completed in which the transcriptomes of 1,124 plant species from the families viridiplantae, glaucophyta and rhodophyta were sequenced. The protein coding sequences were subsequently compared to infer phylogenetic relationships between plants and to characterize the time of their diversification in the process of evolution. Transcriptome studies have been used to characterize and quantify gene expression in mature pollen. Genes involved in cell wall metabolism and cytoskeleton were found to be overexpressed. Transcriptome approaches also allowed to track changes in gene expression through different developmental stages of pollen, ranging from microspore to mature pollen grains; additionally such stages could be compared across species of different plants including *Arabidopsis*, rice and tobacco

Relation to other ome fields



General schema showing the relationships of the genome, transcriptome, proteome, and metabolome (lipidome).

Similar to other -ome based technologies, analysis of the transcriptome allows for an unbiased approach when validating hypotheses experimentally. This approach also allows for the discovery of novel mediators in signaling pathways. As with other -omics based technologies, the transcriptome can be analyzed within the scope of a multiomics approach. It is complementary to metabolomics but contrary to proteomics, a direct association between a transcript and metabolite cannot be established. There are several -ome fields that can be seen

as subcategories of the transcriptome. The exome differs from the transcriptome in that it includes only those RNA molecules found in a specified cell population, and usually includes the amount or concentration of each RNA molecule in addition to the molecular identities. Additionally, the transcritpome also differs from the translatome, which is the set of RNAs undergoing translation. The term meiome is used in functional genomics to describe the meiotic transcriptome or the set of RNA transcripts produced during the process of meiosis. Meiosis is a key feature of sexually reproducing eukaryotes, and involves the pairing of homologous chromosome, synapse and recombination. Since meiosis in most organisms occurs in a short time period, meiotic transcript profiling is difficult due to the challenge of isolation (or enrichment) of meiotic cells (meiocytes). As with transcriptome analyses, the meiome can be studied at a whole-genome level using large-scale transcriptomic techniques. The meiome has been well-characterized in mammal and yeast systems and somewhat less extensively characterized in plants.

The thanatotranscriptome consists of all RNA transcripts that continue to be expressed or that start getting re-expressed in internal organs of a dead body 24–48 hours following death. Some genes include those that are inhibited after fetal development. If the thanatotranscriptome is related to the process of programmed cell death (apoptosis), it can be referred to as the apoptotic thanatotranscriptome. Analyses of the thanatotranscriptome are used in forensic medicine. eQTL mapping can be used to complement genomics with transcriptomics; genetic variants at DNA level and gene expression measures at RNA level.

Relation to proteome

The transcriptome can be seen as a subset of the proteome, that is, the entire set of proteins expressed by a genome. Let us briefly discuss the ***proteome***

The proteome is the entire set of proteins that is, or can be, expressed by a genome, cell, tissue, or organism at a certain time. It is the set of expressed proteins in a given type of cell or organism, at a given time, under defined conditions. Proteomics is the study of the proteome.

Types of proteomes

While proteome generally refers to the proteome of an organism, multicellular organisms may have very different proteomes in different cells, hence it is important to distinguish proteomes in cells and organisms. A cellular proteome is the collection of proteins found in a particular cell type under a particular set of environmental conditions such as exposure to hormone stimulation. It can also be useful to consider an organism's complete proteome, which can be conceptualized as the complete set of proteins from all of the various cellular proteomes. This is very roughly the protein equivalent of the genome. The term *proteome* has also been used to refer to the collection of proteins in certain sub-cellular systems, such as organelles. For instance, the mitochondrial proteome may consist of more than 3000 distinct proteins. The proteins in a virus can be called a *viral proteome*. Usually viral proteomes are predicted from the viral genome but some attempts have been made to determine all the proteins expressed from a virus genome, i.e. the viral proteome. More often, however, virus proteomics analyzes the changes of host proteins upon virus infection, so that in effect *two* proteomes (of virus and its host) are studied.

Importance in cancer

The proteome can be used in order to comparatively analyze different cancer cell lines. Proteomic studies have been used in order to identify the likelihood of metastasis in bladder cancer cell lines KK47 and YTS1 and were found to have 36 unregulated and 74 down regulated proteins. The differences in protein expression can help identify novel cancer signaling mechanisms. Biomarkers of cancer have been found by mass spectrometry based proteomic analyses. The use of proteomics or the study of the proteome is a step forward in personalized medicine to tailor drug cocktails to the patient's specific proteomic and genomic profile. The analysis of ovarian cancer cell lines showed that putative biomarkers for ovarian cancer include "α-enolase (ENOA), elongation factor Tu, mitochondrial (EFTU), glyceraldehyde-3-phosphate dehydrogenase (G3P), stress-70 protein, mitochondrial (GRP75), apolipoprotein A-1 (APOA1), peroxiredoxin (PRDX2) and annexin A (ANXA)". Comparative proteomic analyses of 11 cell lines demonstrated the similarity between the metabolic processes of each cell line; 11,731 proteins were completely identified from this study. Housekeeping proteins tend to show greater variability between cell lines. Resistance to certain cancer drugs is still not well understood. Proteomic analysis has been used in order to identify proteins that may have anti-cancer drug properties, specifically for the colon cancer drug irinotecan. Studies of adenocarcinoma cell line LoVo demonstrated that 8

proteins were unregulated and 7 proteins were down-regulated. Proteins that showed a differential expression were involved in processes such as transcription, apoptosis and cell proliferation/differentiation among others.

The proteome in bacterial systems

Proteomic analyses have been performed in different kinds of bacteria to assess their metabolic reactions to different conditions. For example, in bacteria such as *Clostridium* and *Bacillus*, proteomic analyses were used in order to investigate how different proteins help each of these bacteria spores germinate after a prolonged period of dormancy. In order to better understand how to properly eliminate spores, proteomic analysis must be performed.

Size and contents

The genomes of viruses and prokaryotes encode a relatively well-defined proteome as each protein can be predicted with high confidence, based on its open reading frame (in viruses ranging from ~3 to ~1000, in bacteria ranging from about 500 proteins to about 10,000). However, most protein prediction algorithms use certain cut-offs, such as 50 or 100 amino acids, so small proteins are often missed by such predictions. In **eukaryotes** this becomes much more complicated as more than one protein can be produced from most genes due to alternative splicing. Proteoforms. There are different factors that can add variability to proteins. SAPs (single amino acid polymorphisms) and non-synonymous single nucleotide polymorphisms (nsSNPs) can lead to different "proteoforms" or "proteomorphs".

Dark proteome. The term dark proteome defines regions of proteins that have no detectable sequence homology to other proteins of known three-dimensional structure and therefore cannot be modeled by homology.

Human proteome. Currently, several projects aim to map the human proteome, including the Human Proteome Map, ProteomicsDB and The Human Proteome Project (HPP). Much like the human genome project, these projects seek to find and collect evidence for all predicted protein coding genes in the human genome. The Human Proteome Map currently (October 2020) claims 17,294 proteins and ProteomicsDB 15,479, using different criteria. On October 16, 2020, the HPP published a high-stringency blueprint covering more than 90% of the

predicted protein coding genes. Proteins are identified from a wide range of foetal and adult tissues and cell types, including hematopoietic cells.

Methods to study the proteome

Analysing proteins proves to be more difficult than analysing nucleic acid sequences. While there are only 4 nucleotides that make up DNA, there are at least 20 different amino acids that can make up a protein. Additionally, there is currently no known high throughput technology to make copies of a single protein. Numerous methods are available to study proteins, sets of proteins, or the whole proteome. In fact, proteins are often studied indirectly, e.g., using computational methods and analyses of genomes. Only a few examples are given below.

Separation techniques and electrophoresis

Proteomics, the study of the proteome, has largely been practiced through the separation of proteins by two-dimensional gel electrophoresis. In the first dimension, the proteins are separated by isoelectric focusing, which resolves proteins on the basis of charge. In the second dimension, proteins are separated by molecular weight using SDS-PAGE. The gel is stained with Coomassie Brilliant Blue or silver to visualize the proteins. Spots on the gel are proteins that have migrated to specific locations

Mass spectrometry

Mass spectrometry is one of the key methods to study the proteome. Some important mass spectrometry methods include Orbitrap Mass Spectrometry, MALDI (Matrix Assisted Laser Desorption/Ionization), and ESI (Electrospray Ionization). Peptide mass fingerprinting identifies a protein by cleaving it into short peptides and then deduces the protein's identity by matching the observed peptide masses against a sequence database. Tandem mass spectrometry, on the other hand, can get sequence information from individual peptides by isolating them, colliding them with a non-reactive gas, and then cataloguing the fragment ions produced.

Chromatography

Liquid chromatography is an important tool in the study of the proteome. It allows for very sensitive separation of different kinds of proteins based on their affinity for a matrix. Some newer methods for the separation and identification of proteins include the use of monolithic capillary columns, high temperature chromatography and capillary electrochromatography.

Blotting

Western blotting can be used in order to quantify the abundance of certain proteins. By using antibodies specific to the protein of interest, it is possible to probe for the presence of specific proteins from a mixture of proteins.

Protein complementation assays and interaction screens

Protein-fragment complementation assays are often used to detect protein–protein interactions. The yeast two-hybrid assay is the most popular of them but there are numerous variations, both used *in vitro* and *in vivo*. Pull-down assays are a method to determine what kinds of proteins a protein interacts with. The Plasma Proteome database contains information on 10,500 blood plasma proteins. Because the range in protein contents in plasma is very large, it is difficult to detect proteins that tend to be scarce when compared to abundant proteins. There is an analytical limit that may possibly be a barrier for the detections of proteins with ultra low concentrations. Databases such as neXtprot and UniProt are central resources for human proteomic data.

The analysis of relative mRNA expression levels can be complicated by the fact that relatively small changes in mRNA expression can produce large changes in the total amount of the corresponding protein present in the cell. One analysis method, known as gene set enrichment analysis, identifies coregulated gene networks rather than individual genes that are up- or down-regulated in different cell populations. Although microarray studies can reveal the relative amounts of different mRNAs in the cell, levels of mRNA are not directly proportional to the expression level of the proteins they code for. The number of protein molecules synthesized using a given mRNA molecule as a template is highly dependent on translation-initiation features of the mRNA sequence; in particular, the ability of the

translation initiation sequence is a key determinant in the recruiting of ribosomes for protein translation.

Transcriptome databases

- Ensembl:
- OmicTools:
- Transcriptome Browser:
- ArrayExpress:

References

Proteomic analysis of cell lines to identify the irinotecan resistance proteins by Xingchen Peng, F. Gong, Meng Wei, X. Chen, Y. Chen, Ke Cheng, F. Gao, F. Xu, F. Bi, J. Liu

Chen Y, Barat B, Ray WK, Helm RF, Melville SB, Popham DL. Membrane Proteomes and Ion Transporters in Bacillus anthracis and Bacillus subtilis Dormant and Germinating Spores. J Bacteriol. 2019 Feb 25

Kozlowski LP. Proteome-pI: proteome isoelectric point database. Nucleic Acids Res. 2017 Jan 4

Leslie M. Outsize impact. Science. 2019 Oct 18;366(6463):296-299

Aebersold R, Agar JN, Amster IJ, Baker MS, Bertozzi CR, Boja ES, Costello CE, Cravatt BF, Fenselau C, Garcia BA, Ge Y, Gunawardena J, Hendrickson RC, Hergenrother PJ, Huber CG, Ivanov AR, Jensen ON, Jewett MC, Kelleher NL, Kiessling LL, Krogan NJ, Larsen MR, Loo JA, Ogorzalek Loo RR, Lundberg E, MacCoss MJ, Mallick P, Mootha VK, Mrksich M, Muir TW, Patrie SM, Pesavento JJ, Pitteri SJ, Rodriguez H, Saghatelian A, Sandoval W, Schlüter H, Sechi S, Slavoff SA, Smith LM, Snyder MP, Thomas PM, Uhlén M, Van Eyk JE, Vidal M, Walt DR, White FM, Williams ER, Wohlschlager T, Wysocki VH, Yates NA, Young NL, Zhang B. How many human proteoforms are there? Nat Chem Biol. 2018 Feb 14.

Shi Y, Xiang R, Horváth C, Wilkins JA. The role of liquid chromatography in proteomics. J Chromatogr A. 2004 Oct 22.

Gómez-Serrano M, Camafeita E, Loureiro M, Peral B. Mitoproteomics: Tackling Mitochondrial Dysfunction in Human Disease. Oxid Med Cell Longev. 2018 Nov 8;2018.

Viral proteomics: global evaluation of viruses and their interaction with the host by K. Viswanathan, K. Früh (2007)

Cruz IN, Coley HM, Kramer HB, Madhuri TK, Safuwan NA, Angelino AR, Yang M. Proteomics Analysis of Ovarian Cancer Cell Lines and Tissues Reveals Drug Resistance-associated Proteins. Cancer Genomics Proteomics. 2017