Lecture 3. Functional genomics.

Learning outcomes:

1. Give the definition to the following terms: functional genomics, genome organization, informational capacity of the genome, informational density of the genome.
2. Characterize the different types of genes by their functions, give the specific examples.
3. Explain how the human genome encodes 100 thousands proteins if it contains only 25-30 thousands of genes?
4. Compare the genomes of several absolutely different organisms with the human genome by their structure and informartional properties, analyze the differences and similarities.
5. Describe the mechanisms of gene expression: transcription, post-transcriptional modyfications, translation and post-translational modyfications of proteins.

Functional genomics is a field of molecular biology that attempts to describe **gene (and protein**) **functions and interactions**. Functional genomics make use of the vast data generated by genomic and transcriptomic projects (such as genome sequencing projects and RNA sequencing). Functional genomics focuses on the dynamic aspects such as gene transcription, translation, regulation of gene expression and protein–protein interactions, as opposed to the static aspects of the genomic information such as DNA sequence or structures. A key characteristic of functional genomics studies is their genome-wide approach to these questions, generally involving high-throughput methods rather than a more traditional "gene-by-gene" approach.

Genome organization is the **total structure** of the genome (DNA- or RNA-containing, the number and conformation of DNA molecules (linear or circle molecules), the number of chromosomes, organization of genes (mono- and polycistronic) and etc.). For example, some viruses (retroviruses, adenoviruses, coronaviruses and etc.) have only the RNA molecules in their genomes. All other organisms have DNA-containing genomes. **Informational capacity of the genome** is the number of gene in the genome. **Informational density of the genome** is the ratio of the number of genes to the size (to the length) of the genome. For example, informational density of viral and prokaryotic genomes are very high in comparison with eukaryotic genomes (with the human genome in this number too).

There are several types of genes. **Protein-coding genes** synthesize the **mRNAs** that further are translated to **proteins**. **RNA-coding genes** synthesize the molecules of all other types of **RNA** (**tRNA, rRNAs, miRNAs** and many other different types of **small RNAs)** and don't encode the proteins. **Pseudogenes** are non-active regions of DNA that are very similar to protein-coding or RNA-coding genes and that could be active in the past but don't have the clear functions now.

Gene expression is the process by which information from a gene is used in the synthesis of a functional gene product that enables it to produce protein as the end product. These products are often proteins, but in non-protein-coding genes such as transfer RNA (tRNA) or small nuclear RNA (snRNA) genes, the product is a functional RNA. Gene expression is summarized in the central dogma of molecular biology first formulated by Francis Crick in 1958, further developed in his 1970 article, and expanded by the subsequent discoveries of reverse transcription and RNA replication.

The production of a RNA copy from a DNA strand is **called transcription**, and is performed by RNA polymerases, which add one ribonucleotide at a time to a growing RNA strand as per the complementarity law of the nucleotide bases. This RNA is complementary to the template 3′ → 5′ DNA strand, with the exception that thymines (T) are replaced with uracils (U) in the RNA. While transcription of prokaryotic protein-coding genes creates messenger RNA (mRNA) that is ready for translation into protein, transcription of eukaryotic genes leaves a **primary transcript** of RNA (**pre-RNA**), which first has to undergo a **series of modifications** to become **a mature RNA**. The **processing of pre-mRN**A includes **5′ capping**, which is set of enzymatic reactions that add **7-methylguanosine (m7G)** to the 5′ end of pre-mRNA and thus protect the RNA from

degradation by exonucleases. The m7G cap is then bound by cap binding complex heterodimer (CBC20/CBC80), which aids in mRNA export to cytoplasm and also protect the RNA from decapping. Another modification is **3′ cleavage and polyadenylation**. They occur if polyadenylation signal sequence (5′- AAUAAA-3′) is present in pre-mRNA, which is usually between protein-coding sequence and terminator. The pre-mRNA is first cleaved and then a series of ~200 adenines (A) are added to form poly(A) tail, which protects the RNA from degradation. The poly(A) tail is bound by multiple poly(A)-binding proteins (PABPs) necessary for mRNA export and translation re-initiation. In the inverse process of deadenylation, poly(A) tails are shortened by the CCR4-Not 3′-5′ exonuclease, which often leads to full transcript decay.
Pre-mRNA is spliced to form of mature mRNA. A very important modification of eukaryotic pre-mRNA is **RNA splicing**. The majority of eukaryotic pre-mRNAs consist of alternating segments called **exons** and **introns**. During the process of splicing, an RNA-protein catalytical complex known as spliceosome catalyzes two transesterification reactions, which remove an intron and release it in form of lariat structure, and then splice neighbouring exons together. In certain cases, some introns or exons can be either removed or retained in mature mRNA. This so-called **alternative splicing** creates series of different transcripts originating from a single gene. Because these transcripts can be potentially translated into different proteins, splicing extends the complexity of eukaryotic gene expression and the size of a species proteome. Extensive RNA processing may be an evolutionary advantage made possible by the nucleus of eukaryotes. In prokaryotes, transcription and translation happen together, whilst in eukaryotes, the nuclear membrane separates the two processes, giving time for RNA processing to occur.
In molecular biology and genetics, **translation** is the process in which ribosomes in the cytoplasm or endoplasmic reticulum synthesize proteins after the process of transcription of DNA to RNA in the cell's nucleus. In translation, messenger RNA (mRNA) is decoded in a ribosome, outside the nucleus, to produce a specific amino acid chain, or polypeptide.
Post-translational modifications (PTMs) are covalent modifications to proteins. Like RNA splicing, they help to significantly diversify the proteome. These modifications are usually catalyzed by enzymes. Additionally, processes like covalent additions to amino acid side chain residues can often be reversed by other enzymes. However, some, like the proteolytic cleavage of the protein backbone, are irreversible.

The questions for self - control:

1. What are the functional genomics, genome organization, informational capacity of the genome, informational density of the genome?
2. What are the differences between the protein-coding, RNA-coding genes and pseudogenes?
3. What is the alternative splicing and how it increases the informational opportunities of the genome?
4. What are the differences and similarities of viral, prokaryotic and eucariotic genomes? Give some specific examples and compare them with the human genome.
5. What are the gene expression, transcription, post-transcriptional modyfications, translation and post-translational modyfications of proteins?

Recommended readings:

1. Graur D, Zheng Y, Price N, Azevedo RB, Zufall RA, Elhaik E (20 February 2013). "On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE". *Genome Biology and Evolution*. **5** (3): 578–90. doi:10.1093/gbe/evt028. PMC 3622293. PMID 23431001
2. Crick FH (1958). "On protein synthesis". Symposia of the Society for Experimental Biology. 12: 138–63. PMID 13580867.
3. Crick F (August 1970). "Central dogma of molecular biology". Nature. 227 (5258): 561–3. Bibcode:1970Natur.227..561C. doi:10.1038/227561a0. PMID 4913914.

4. "Central dogma reversed". Nature. 226 (5252): 1198–9. June 1970. Bibcode:1970Natur.226.1198.. doi:10.1038/2261198a0. PMID 5422595.
5. Temin HM, Mizutani S (June 1970). "RNA-dependent DNA polymerase in virions of Rous sarcoma virus". Nature. 226 (5252): 1211–3. doi:10.1038/2261211a0. PMID 4316301.
6. Baltimore D (June 1970). "RNA-dependent DNA polymerase in virions of RNA tumour viruses". Nature. 226 (5252): 1209–11. doi:10.1038/2261209a0. PMID 4316300.
7. Iyer LM, Koonin EV, Aravind L (January 2003). "Evolutionary connection between the catalytic subunits of DNA-dependent RNA polymerases and eukaryotic RNA-dependent RNA polymerases and the origin of RNA polymerases". BMC Structural Biology. 3: 1. doi:10.1186/1472-6807-3-1. PMC 151600. PMID 12553882.
8. Brueckner F, Armache KJ, Cheung A, Damsma GE, Kettenberger H, Lehmann E, Sydow J, Cramer P (February 2009). "Structure-function studies of the RNA polymerase II elongation complex". Acta Crystallographica D. 65 (Pt 2): 112–20. doi:10.1107/S0907444908039875. PMC 2631633. PMID 19171965.
9. Krebs, Jocelyn E. (2017-03-02). Lewin's genes XII. Goldstein, Elliott S.,, Kilpatrick, Stephen T. Burlington, MA. ISBN 978-1-284-10449-3. OCLC 965781334
10. Walsh CT, Garneau-Tsodikova S, Gatto GJ (December 2005). "Protein posttranslational modifications: the chemistry of proteome diversifications". Angewandte Chemie. 44 (45): 7342–72. doi:10.1002/anie.200501023. PMID 16267872. S2CID 32157563.