Genomics – Part 1: The Basics -an Obligation for Public Health

Arguing that 'genomics' is a 'laboratory' science and not of interest to public health is shortsighted. Even though working with people and communities, National Public Health organizations should also be aware of molecular biology and what is new in genetics, epigenetics, and gerontology.

A meaningful contribution of Public Health to control non-communicable diseases (NCDs), especially in low- and middle-income countries, requires a general insight into human metabolism and genetics (1). What applies to NCDs also is valid for infectious diseases, as the example of the controversial widespread use of genetically engineered vaccines during the Covid-19 period shows. Most public health personnel who received mRNA vaccines probably still don't know what mRNA means and that the development of the vaccine depends on genetic engineering. To make good for that, a previous entry into this blog familiarized the reader with the newest methodology for genetic engineering (2). Without an introduction to the newest developments in genetics, it is hard to understand what far-reaching possibilities scientists nowadays have in fiddling around with our genetic system regarding vaccination, medication, and epigenetics.

Epigenetics, for instance, has a direct link to one of the significant issues of concern for community medicine. Epigenetics expands our understanding of how the physical and social environment shapes the profile in our genome to increase the risk of acquiring diseases (3).

<u>The human cell</u>

What was known as 'genetic' is now termed 'genomics'. The 'Omic' technologies include genomics, proteomics, and metabolomics. The terminology change was chosen to illustrate the need to look at eukaryotes from the point of 'system biology,' which was explained previously in this blog (4). In genetics, it subscribes to the dynamic interchange of all the components which each other (5, 6).

To better understand recent developments in genomics, it is necessary first to recall basic genetics as being introduced in high school and maybe forgotten by now. Our physical existence is based on trillions of cells. Genes are within the cell and encode proteins responsible for the function of the cell. Human cells are eukaryotes containing a membrane surrounding the organelles, such as the nucleus, the cytoplasm, and the cytoskeleton. The cell nucleus contains most of the genetic code in form of deoxyribonucleic acid (DNA). The endoplasmic reticulum helps to process molecules and assists in transporting them. The Golgi apparatus further processes proteins, such as lysosomes and peroxisomes, which digest bacteria and eject toxic substances. Mitochondria transform energy from food so it can be used by the cell. Mitochondria have their own DNA, which is derived from the mother. Sperms have no mitochondria. The ribosomes enable the formation of proteins following the information provided by the genetic code (7, 8).

<u>DNA, nucleotides, and the Watson-Crick double helix</u>

DNA contains the genetic code; however, it is chemically inert and, therefore, relatively stable. For instance, it was possible to investigate 38,000-year-old DNA obtained from a Neanderthal skull. The implication of this has been described in entries of this blog (9, 10). Within the name abbreviated as DNA, the sugar deoxyribose is the main content, which connects to an inorganic phosphate group at the 5' end of the carbon atom from the sugar and a nitrogenous base at the carbon atom 3' position of the sugar. Through polyester bonds, they form some polymer. It was realized that the DNA molecule appears alongside fibers in an orderly fashion. Through the ingenious and complex method, the X-ray diffraction patterns*, it was known that DNA is helical, as Rosalind Franklin and Maurice Wilkins discovered in 1952.

To find out how DNA strands are arranged was a topic Francis Crick and James Watson were working on from 1951 to 1953. In February 1953, Watson paired the two bigger nucleotides of a wire-frame model with the two smaller ones and got the pairs adenine (A) to thymine(T) and guanine (G) to cytosine (C). With this pairing of nucleotides, they got the Watson-Crick double-helix and the Nobel Prize in 1962 with Maurice Wilkins. Wilkins' coworker, Rosalind Franklin, died of ovarian cancer three years before the Nobel Prize was awarded (11). (Nobel prizes are only dedicated to living scientists.) It is now generally accepted that 'evolution is a process that begins at the molecular level, inside the double helix of DNA' (12)(page 691).

From 5' to 3', alleles, histones, and microsomes

Our existence, health, and disease depend on the variation of the nucleotide in our DNA, which is 'read' by the ribonucleic acid (RNA) to finally result in proteins regulating our metabolism in health and disease. The construction of DNA and RNA is similar. Through hydrogen bonds between the bases, adenine (A) binds with thymine (T) and guanine (G) with cytosine (C). The base pairs form the rung of a DNA ladder. The two strains of nucleotides with sugar-phosphate backbones shape the double strand of the Watson-Crick 'double helix.' The DNA strain is 'read' in a specific direction from 5' (five primer) up to 3' (three primer). Within the double helix, 5' is constantly opposed by the other strain at 3'. One human individual comprises three billion bases, and almost all of them are the same for all of us, except for one percent. The order of the sequence of the bases contains the information on the genes which determine our heredity and organ functions. The bases are often compared to the alphabet, and the sequences' variations are the words. The DNA chains could be very long, such as 250 million nucleotides containing 4 the power of 250.000.000 sequences (13)(pages 170 – 178).

It is estimated that there are about twenty- to twenty-five thousand genes. There are two copies of each gene originating from our parents. Each of us differs from another, except we are identical twins. The differences are made because of the small variances of the codes in the gene alleles. An allele is a particular variant at a specific locus, whether at a gene, a region, or a nucleotide position of the genome. Genes are twisted around histones, which are small DNA-binding proteins. And the DNA of the genes and the histones are wrapped into chromosomes.

Gene expression through RNA to the polypeptide chain

Those whose schooling days are beyond them for several decades might still remember the twenty-three pairs of chromosomes, including the sex chromosomes with one X and one Y for a male and two X for a female. How the information of the base sequences is finally bestowed into individual proteins is termed 'gene expression.' RNA is instrumental in reading and transferring information on the DNA chain. RNA is similarly composed as DNA in that the sugar is ribose, and instead of thymine, there is the base uracil (U). Most RNA molecules are single-stranded, have fewer nucleotides than the DNA strand, and are flexible.

The information transmitted by the RNA is the genetic code necessary to compose proteins consisting of amino acids. The nucleotide sequence determines the amino acid sequence within the polypeptide chain. The four letters of the code are the four nucleotides. Three letters of nucleotides stand for a particular amino acid and are called a codon. The four nucleotides can be arranged into four power three, that is, 64 possible triplets. There are not only codons for amino acids but also for regulating gene expression as such. There are codons to initiate the reading and stop the progress. Stop reading codons are TAA, TAG, and TGA. More than one codon might be available for a given amino acid, the so-called nonsense codons (13) (pages 258 – 267).

Transcription and translation

Gene transcription is more complex in eukaryotic cells than in procaryotic organisms, but the general scheme is comparable. Here, the transcription in eukaryotes is outlined, concentrating on the basic steps. The process starts with the transcription of DNA to RNA within the nucleus. The translation into a polypeptide chain takes place on so-called workbenches, the ribosome in the cytoplasm. The enzyme RNA polymerase is the key promoter of the process. While DNA regions are transcribed into mRNA, synthesized by RNA polymerase, so-called 'introns' are removed. What finally is transferred to the cytoplasm are the 'exons. The strands are separated, 'spliced,' for the mature tRNA from which translation into polypeptides occurs.

The RNA polymerase has to 'know' where to start the process upstream, at the 5' end of the DNA strand. An activator protein links to the enhancer sequence and to an additional mediator The mediator on the opposite site links to the RNA polymerase at the promotor sequence finally to start transcription and to loop the DNA strand around itself (see Figure 2 (14). The regulatory mechanism of the polymerase affects the chromatin structure. An open chromatin structure relates to active gene transcription, and a more compressed chromatin inhibits it.

How chromatin is organized is essential for gene expression within the cell. Each cell has the same set of genes, but considering the body's different organs, there are many types of cells. Transcription, therefore, is regulated by the cell according to its distinct needs by the interaction enhancer and promotor elements. The RNA polymerase activity is influenced by regulator proteins with different roles for different genes, and by this, the cell can regulate many genes at once (13) (pages 267 – 283).

Chromosome looping

For gene expression, in prokaryotes, such as bacteria, regulatory proteins are close to the transcription promotor sites, but this is not the case for eukaryote cells. Enhancers might be

considerably ‘far away’ from the genes that are supposed to be regulated. A combination of several regulator proteins might be active at the same time. To bring the genes and functional proteins together, the looping of chromosomes of three-dimensional (3D) genomes facilitates the process. As mentioned above, looping out the DNA strand brings the activator protein into the reach of the RNA polymerase and the other proteins necessary to initiate transcription. Chromosome looping and three-dimensional (3D) genome organization is a very active research field, as reviewed in a short review recently (15).

A little bit of history of genetics

Not so long ago, only chromosomes could be seen in a microscope. Characterized as the necessary basics to understand recent developments in ‘omics’ illustrates that advancement in methodology must be far beyond just only looking at chromosomes. A journey through the history of the science of genetics might start with Darwin, the father of the discipline of evolution, who missed reading Mendel’s results about pea traits and the fundamentals of inheritance. Mendel’s findings were ignored when published in 1865 and just rediscovered around 1900. Oswald Avery and coworkers in 1944 were able to establish that genes consist of DNA, and its structure, as mentioned above, was made known by Watson and Crick in 1953. Frederick Sanger got one of his two Nobel Prizes in 1980 for his contribution to DNA sequencing technology. The Sanger sequencing technology paved the way for machinery to be less expensive and more accurate, such as the whole genome shotgun sequencing and the ‘next-generation sequencing’ (16). To understand how DNA sequencing functions in principle, remember how gel electrophoresis works.

Sanger sequencing

Hardly anybody who sets foot into a lab in biology will miss coming across the gel electrophoresis technique. In an electric field, charged molecules migrate through the gel towards the opposite charge. The phosphate group of the DNA molecule is negatively charged and, on its way to the positive pole, is separated into bands. These could be made visible by various means, which could then be identified. This fundamental principle could be extensively modified finally into an automatic process.

The Sanger sequencing copies the DNA replication as in the cell, using ingenious methods within the laboratory. The human genome is too large and complex, so only small pieces can be scrutinized at a time. For further examination, the DNA is split into fragments of identical copies, named recombinant DNA molecules. The process is called molecular cloning. Particular enzymes, such as EcoRI enzymes (pronounced eco R one), are used to cut the DNA double helix at specific sites. The enzyme is an endonuclease that originated from the bacterium E. coli. Many identical DNA fragments provide enough material to work with. Essential ingredients are plasmid vectors of bacteria, small molecules which can replicate themselves like viruses. To distinguish plasmids from viruses, they were called extrachromosomal DNA, separated from the bacterial chromosome. By splitting up the genome, many fragments are achieved with different recombinant DNA. Through another enzyme, DNA ligase, human DNA particles, and plasmid vectors are combined. One way to store the particles, in what is called a human genomic DNA library, is E. coli cells with recombinant DNAs.

Part of the recombinant plasmid with inserted human DNA serves as a template together with a primer for sequencing. The single-stranded primer contains deoxyribonucleotide triphosphates (dATP, dCTP, dGTP, and dTTP). To the template at the free 3' end, nucleotides complementary to the template started to be added due to their inbuild independent replication possibility. Each dideoxynucleotide has a different fluorescent dye. When the newly synthesized DNA strand is run through electrophoresis, the bands can then be scanned, and the sequence of the nucleotide read, and through the computer, the DNA sequence is identified (13) (page 313 – 320).

From Sanger to whole genome shotgun into the Human Genome Project

One sequencing run could at least result in 1000 bases. Genomes, as a whole, are much longer. On the 23 human chromosomes, there are more than three billion base pairs. There are repetitive sequences of the genome known as transposable elements. The problem was that at the 5' end, not only the 'real' sequences were added but also the transposable elements. The question remained where is the required 3' end to connect to the 5' end of the additional part of the genome? What was needed was paired-end sequencing. This was achieved through two primers which allowed the correct assembly of the sequences. For the human genome, the method of whole-genome shotgun sequencing was applied. The technique was called shotgun sequencing because randomly short sequences of the DNA were analyzed in a very time-consuming process. A complete genome sequence was published in 1981 from the cauliflower mosaic virus (17). The initial version of the human genome sequence was published more than 20 years ago by the company Celera Genomics and the Human Genome Project (HGP) (18, 19).

Further on, sequencing was rapidly improved. The Sanger method can only sequence short DNA pieces, is time-consuming and costly. To save time, electrophoresis was a major limiting factor. By improving different dNTP templates, DNA could be detected by light, enabling real-time sequencing and avoiding electrophoresis. Cloning and amplifying DNA was improved, thousands of clonal amplification products became available, and many clonal runs could be run simultaneously. The polymerase chain reaction (PCR) improved clonal amplification. Finally, in 2005, a high-throughput method sequenced whole bacterial genomes quickly and cheaply. Fluorescent signals of single bases and parallel sequencing techniques allowed the development of high-throughput sequencing machines and opened the age of 'next-generation sequencing' (20). The improvement in DNA sequencing went along with the computerization techniques without the ever-increasing multitude of results that were required to be interpreted.

With the improvement in sequencing technology and sophisticated computer programs, those in genetics became increasingly puzzled by the high amount of 'junk DNA' across the human genome. What is not common in bacterial genomes is present in higher organisms, such as humans. About 90% of the variants are non-coding DNA. It is now known that these regions have essential biological functions, such as RNA processing and translating but don't encode protein (21).

Genetics for hereditary and common diseases

The difference between coding and non-coding DNA segments became more evident after sequencing whole genomes was possible instead of just hunting for genes related to rare genetic diseases, such as Hunting disease, causing neurological disorders in one out of 25.000 Caucasians, or Cystic fibrosis occurring in one out of 2000 Caucasians. The latter is a lung disease with excessive mucus production causing early death. Formerly genetic investigation focused on hereditary diseases called Mendelian or Monogenic diseases, of specific clinical interest. In addition to rare Monogenic diseases, research is now more and more extended to discover genetic variations and phenotypes influencing human biology and very common diseases, such as asthma and diabetes (22, 23).

Next-generation sequencing will follow

The developments in the genetics of common diseases should interest Public Health. While curative medicine will benefit from targeting therapeutics, the advantage for Public Health will be to provide well-founded suggestions to particular groups in the population for altering diets and behavior as well as prophylactic intervention. An additional overview of human disease genetics during next-generation sequencing will be the topic of part two about genomics following soon.

* X-Ray diffraction (XRD) uses the phenomenon that atoms of a crystal cause a pattern of waves in a beam of X-rays similar to a beam of light. From the waves the chemical constitutes, the material's physics can be judged.

References:

1. Muktabhant B, Schelp FP, Kraiklang R, Chupanit P, Sanchaisuriya P. Improved control of non-communicable diseases (NCDs) requires an additional advanced concept for public health - a perspective from a middle-income country. F1000Res. 2019;8:286.
2. CRISPR: The biotechnological revolution - risks and achievements public health should know Khon Kaen Thailand: Faculty of Public Health, Khon Kaen University; 2023 [Available from: https://ph.kku.ac.th/eng/index.php/research/journal-club-phkku/206-270466.
3. Carey N. The epigenetics revolution. London: Icon Books Ltd.; 2012. 329 p.
4. Research - System biology and individual cancer therapy Journal Club: Faculty of Public Health; 2021 [Available from: https://ph.kku.ac.th/eng/index.php/research/journal-club-phkku/180-230264.
5. Bruggeman FJ, Westerhoff HV. The nature of systems biology. Trends Microbiol. 2007;15(1):45-50.
6. Horgan RPK, L.C. 'Omic' technologies: genomics, transcriptomics, proteomics and metabolomics. The Obstetrician and Gynaecologist. 2011;13(3):7.
7. Institute NC. Cell 2023 [Available from: https://www.cancer.gov/publications/dictionaries/cancer-terms/def/cell.
8. MedlinePlus. Cells and DNA 2023 [Available from: https://medlineplus.gov/download/genetics/understanding/basics.pdf.

9. Evolution and public health. Part 1: Paleogenetic - a new scientific discipline being of interest for public health might not be as absurd as it seems Khon Kaen, Thailand: Faculty of Public Health, Khon Kaen University; 2021 [Available from: https://ph.kku.ac.th/eng/index.php/research/journal-club-phkku/187-190864.
10. Evolution and public health. Part II - Humans' evolutionary past influences health and diseases Khon Kaen, Thailand: Faculty of Public Health, Khon Kaen University; 2021 [Available from: https://ph.kku.ac.th/eng/index.php/research/journal-club-phkku/188-200864.
11. Pietzsch J. Speed read: Deciphering Life's Enigma Code 2023 [Available from: https://www.nobelprize.org/prizes/medicine/1962/speedread/.
12. Hartwell LH, Hood, L., Goldberg, M.L., Reynolds, A. E., Silver, L.M. Genetics From Genes to Genomes. Fourth Edition ed. New York USA: McGraw-Hill; 2011.
13. Goldberg ML. Genetics. From Genes to Genomes. Seventh Edition ed. New York: Mc Graw Hill; 2021.
14. Scitable. Gene expression [Available from: https://www.nature.com/scitable/topicpage/gene-expression-14121669/.
15. Gaskill M, Harrison M. Tethering gene regulation to chromatin organization. Science. 2022;375(6580):491-2.
16. Portifolio N. Milestone Genomic Sequencing 2021 [Available from: https://www.nature.com/collections/dbieeeeeed.
17. Gardner RC, Howarth AJ, Hahn P, Brown-Luedi M, Shepherd RJ, Messing J. The complete nucleotide sequence of an infectious clone of cauliflower mosaic virus by M13mp7 shotgun sequencing. Nucleic Acids Res. 1981;9(12):2871-88.
18. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409(6822):860-921.
19. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome. Science. 2001;291(5507):1304-51.
20. Willson J. Sequencing - the next generation: Nature Reviews Cross-Journal Team 2021 [Available from: https://www.nature.com/articles/d42859-020-00103-7.
21. Shanmugam A, Nagarajan, A., Pramanayagam, S. Non-coding DNA - a brief rview. Journal of Applied Biology & Biotechnology. 2017;5 (5):6.
22. Claussnitzer M, Cho JH, Collins R, Cox NJ, Dermitzakis ET, Hurles ME, et al. A brief history of human disease genetics. Nature. 2020;577(7789):179-89.
23. Lappalainen T, MacArthur DG. From variant to function in human disease genetics. Science. 2021;373(6562):1464-8.

The textbook from Goldberg et al.: 'Genetics. From genes to genomes can be examined at the Faculty of Public Health. Please contact the email address below.

Frank P. Schelp is responsible for the manuscript's content, and the points of view expressed might not reflect the stance and policy of the Faculty of Public Health, Khon Kaen University, Thailand.

For comments and questions, please contact <awuso11@gmail.com>