Nonlinear Phenomena in Complex Systems, vol. **23**, no. 2 (2020), pp. 113 - 120

Genetic Complexity of the Human Genome in Health and Disease: Basic Concepts

A. Athanassiadou*
Professor Emeritus, Department of Medicine, University of Patras, Patras, GREECE
(Received 20 November, 2019)

Determination of the DNA sequence of the human genome, revealing extensive genetic variation, and the mapping of the genes and the various regulatory elements of genome function within the genomic DNA, has revolutionized the way we view the states of health and disease in our time. Genetic complexity of the genome is manifested on different levels. The first level refers to the expression of protein coding genes, as regulated by their individual promoter in linear proximity. The next level of genetic complexity involves long distance action by far away enhancers, interacting with promoters through DNA looping. This 3-dimensional (3D) regulation is further developing by chromosome folding into the so called *transcription factories*, for fully physiological expression. Chromosome folding, mediated by specific genetic elements – insulators – is adding to the genetic complexity by facilitating movements of chromatin of specific genomic regions – the so-called *topologically associated domains* (TAD) in support of transcription and other cellular functions. Further genetic complexity has emerged with the finding that over 75% of the genome is transcribed and except of the coding genes, a plethora of RNA transcripts are produced – the non-coding RNA – that has important regulatory roles in the gene expression context. The great variation of genome sequence and regulatory elements of the genome architecture are exploited in studies of *genome-wide association* with disease, in the framework of Precision Medicine and in general of Genomic Medicine.

PACS numbers: 87.18.-h,87.14.Gg
Keywords: genetic complexity, human genome, transcription enhancers, chromatin insulators
DOI: https://doi.org/10.33581/1561-4085-2020-23-2-113-120

1. Introduction

The human genome refers to the DNA content of the human cells. Every human cell contains 22 pairs of autosomal chromosomes, numbered from 1 to 22 and 2 sex chromosomes, X and Y, that carry the genetic information in the form of DNA, residing in the nucleus. This is most of the genetic information needed by the cell, the organism and the species to survive and proliferate. The human genome refers to the haploid human genome of nuclear origin, that is the sum of one part of the 22 pairs of autosomal chromosomes and the sex chromosomes, plus the small mitochondrial DNA molecule. The human genome is comprised of 3.3 billion base pairs and the actual nucleotide sequence was determined and finalized in April 2003. The determination of the sequence of the human genome has made possible the study of its 3-dimensional structure and the nuclear architecture in a living cell and the way this is related to the regulation of the gene activity in cells. This, in turn, has lead to the elucidation of genetic processes governing crucial biological states and in the understanding how the genetic complexity of the genomic DNA, coding and non-coding, can affect the outcome of such processes, in health and disease.

The determination of the DNA sequence of the human genome was carried out by an international collaborative action, involving several participants all over the world and it is a great discovery in history, by researchers set to sequence and map all of the genes of our species: *Homo sapiens* [1].

*E-mail: athanass@med.upatras.gr

2. Elements of the human genome

2.1. Reference sequence

This is the sequence of the haploid human genome to be used as a reference for comparison of any sequence of part or whole genome of any individual. The reference sequence of the genome does not represent the genome of any particular individual. The genome of any given individual is unique, and it is not exactly the same among humans, among cells in the same body, even among the same cells of the same tissue at different ages of the organism. This is because the DNA of the genome within the respective cell, is subjected to various changes, such as: chromosome aberrations, gene duplications, insertions or deletions and translocations of DNA, mutations of one or a few nucleotides and more. As a consequence, the determination of the DNA sequence and the mapping of the various parts of DNA into the human genome involved the sequencing of a small number of individuals and the assembling of these data together to get the complete sequence for each chromosome. Therefore, the finished human genome is not representing any one individual as it is rather a mosaic. The completion of the human genome reference sequence was a milestone in modern biology.

2.2. Variation of the human genome

Before even the completion of the sequencing phase of the HGP, studies on the DNA apparent variation had started by the Human Genome Organisation (HUGO) (www.hugo-international.org), through the Human Genome Variation Society (http://www.hgvs.org/). Any two copies of the human genome differ in about 0.1% of the nucleotide sites, meaning there is one variant per 1000 bases in the genome. The most common type of variant is a change in a single base nucleotide, the so-called single nucleotide polymorphism, or SNP (Figure 1). SNPs constitute 90% of the variation in the population. Groups of SNPs along a chromosome, that are close enough to each other to be transmitted together in the next generation, are called haplotypes (Figure 1). Genetic variants have been mapped on the genome mainly by the International HapMap Project (http://www.hapmap.org/) and sequences are constantly updated (ncbi.nlm.nih.gov/genome/guide/human). SNPs and haplotypes represent the first level of genetic complexity of the genome and their bioinformatics analysis provides data used in disease association studies [2, 3].

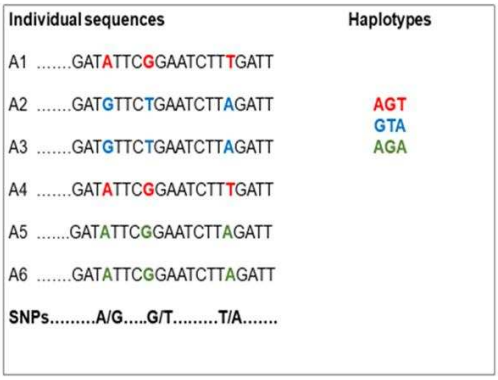

FIG. 1. (color online) SNPs and the derivation of haplotypes. Six individual sequences on a specific part of chromosome A within a given population that are similar except for the SNPs at three different positions. Haplotypes are defined by the group of the changed bases along the same chromosome.

2.3. Tools for the study of the genetic complexity of the genome

Next generation sequencing and Bioinformatics [4]. The analysis of the human genome sequence and the mapping of its elements was facilitated by the development and application of rapid and efficient methods for the determination of the DNA sequence and in particular of the Next Generation sequencing-NGS. This produced a great volume of sequence data that lead to the development of advanced bioinformatics for their analysis and identification of correlations and meaning among the various sequences generated by NGS.

Animal models of disease [5]. The conservation of gene structure and function in animals through evolution and the fact that experimentation in humans is restricted, lead to the development of animal models for the various human biological pathways, conditions and diseases, that offered an indispensable tool for these studies. The extrapolation of data and conclusions from e.g mouse to humans is mostly informative, but not always.

Chromosome Conformation Capture (3C) [6, 7]. This is a technology that determines chromosome folding and chromosome-chromosome interactions within the 3-dimensional structure of cell nucleus. Such interactions may be detected by the technique of Fluorescent in situ Hybridization (FISH) [8], but it is the 3C technology that has been pivotal in unraveling the mode of transcription in the 3D nucleus within Transcription Factories, whereby, gene transcription does not take place by the polymerase complex approaching the gene, but rather by the active genes migrating to preassembled transcription sites [6]. Additionally it shows that the genome of many species is organized into domains of preferential chromosomal interactions called Topologically Associated Domains (TAD), which have emerged as a key element for higher order genome organization and function [9].

Mathematical tools [10–14]. Studies on the proper execution of the gene expression program for each cell as well as on structural changes of the genome, reveal complex processes, which are demonstrated by a massive body of data regarding genetic information and networks. Mathematical tools such as bioinformatics, algorithms, mathematical modelling etc play a crucial role in interpreting this vast amount of biological data obtained by the application of mass analysis of the human genome function.

2.4. Landscape of the human genome

The various entities of the DNA of the human genome, as defined by their specific identity/function, are located in various places along the chromosomal DNA and, collectively, constitute the landscape of the human genome, which is comprised of:

- protein coding genes, approx. 23000, that encode for proteins;
- exons, deriving form the Open Reading Frames (ORF) which are the protein coding sequences of the genes (approx. 1.5% of the genome) – analysis of the specific sequences of the ORFs and their involvement in disease lead to the formulation of ORFs landscape across diseases [12, 13];
- introns – the non-coding parts of the genes (approx. 3.5%);
- repetitive DNA sequences, composed of tandem, repeated sequences of two to several thousand base pairs (and is estimated to constitute about 30% of the genome);
- non-coding RNA genes, deriving from transcription of DNA but not coding for proteins (have mostly regulatory function);
- RNA genes for the generation of functional RNAs such as the ribosomal and transfer RNA for protein synthesis and of unknown identity or function etc. (https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/regulatory-dna-sequence).

2.5. 3D function of the human genome

The human genome carries out the basic processes for the DNA function, namely: the DNA replication, by which the total DNA in a cell is duplicated before cell division; the transcription of DNA into RNA, in execution of the specific gene expression program in the particular cell; and the repair of DNA damage incurred by internal and external agents. All three processes involve the complex action of genetic elements, whether coding or non-coding for proteins. The following subsections are examples of such actions.

2.5.1. Genetic complexity of the regulation of gene expression

The physiological expression of a gene relies on proper transcription for the production of the message and the translation of the message into protein, in the correct cells, at the correct time and at the correct level. These requirements are ensured by the function of several regulatory, genetic elements, acting at different levels of complexity. The following paragraphs present, in brief, basic features of this genetic activity in ascending levels of complexity.

Long distance gene regulation [15]. The analysis of the human genome lead to the development of tools for the investigation and the understanding the gene expression in the 3-dimensional space of the nucleus (Figure 2). The function of promoters, as the site where the RNA polymerase complex binds in order to start transcription, is regulated by enhancers sites on the DNA usually close to the gene, but sometimes are located hundreds of kilobases (kb) away from the promoter of the gene. Enhancers are DNA sites upon which a set of specific transcription factors binds forming a functional complex capable of engaging with target promoters through long-range interactions. Recent studies into the 3D folding of chromosomes are now providing new insights on how enhancers participate in the regulation of a specific gene [14]. Sequence variation within a distal regulatory element might influence phenotypes or association with disease states through the action of causal mutation or the presence of polymorphic SNPs (Figure 3).

Transcription factories. Active genes, especially genes that are co-regulated (e.g. the haemoglobin gene families), are dynamically organized for transcription into nuclear sub-compartments, which they share according to their identity, localization and function. Evidently, active genes rather than recruiting and assembling transcription complexes in their position dictated by linearity, migrate to preassembled transcription sites, to attain physiological level of expression [8, 16]. These sites are called transcription factories. They are

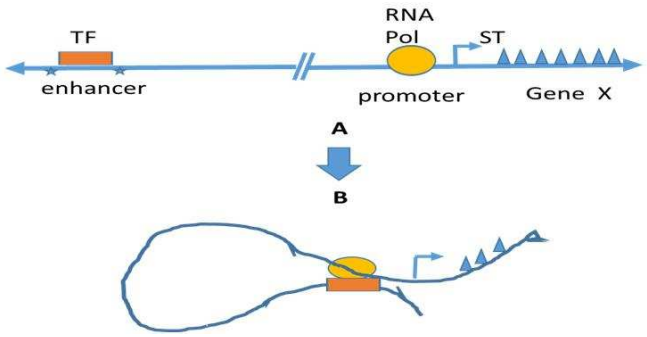

FIG. 2. (color online) A simple demonstration of the long distance regulation of gene expression. Structure A represents the linear configuration of Gene X and elements of the regulation of its transcription. ST is the site of transcription start; RNA Pol is the complex of RNA polymerase that transcribes the DNA into mRNA, after binding on the promoter site. TF is a transcription factor that can bind on an enhancer site on the DNA-part between the two stars – located a long distance away from the gene and its promoter. Structure B represents the looping of the DNA so that the enhancer comes opposite of the promoter site and the TF interacts with the RNA polymerase complex, promoting transcription.

formed by stretches of DNA tens of thousands of nucleotides long and rely on DNA regulatory elements such as *enhancers* that serve as binding platforms for transcription factors and form long range chromatin contacts through chromatin looping. Sub-nuclear positioning and chromatin loops that influence transcriptional activity may be involved in impaired movement to specific nuclear sites and this can cause changes in expression, leading to disease [17].

Topologically associated domains. The genome is organized in transcriptionally active domains (euchromatin) and transcriptionally inactive-condensed-chromatin (heterochromatin). The two states are separated and defined by DNA sequences, the *insulators*, acting as boundary elements. Gene positioning within different nuclear environments is closely linked to gene activity [14]. Usually, the active parts of the genome occupy the inner part of the nucleus, while the inactive heterochromatin is found in the periphery of the nucleus. The actual genes that are active or inactive may

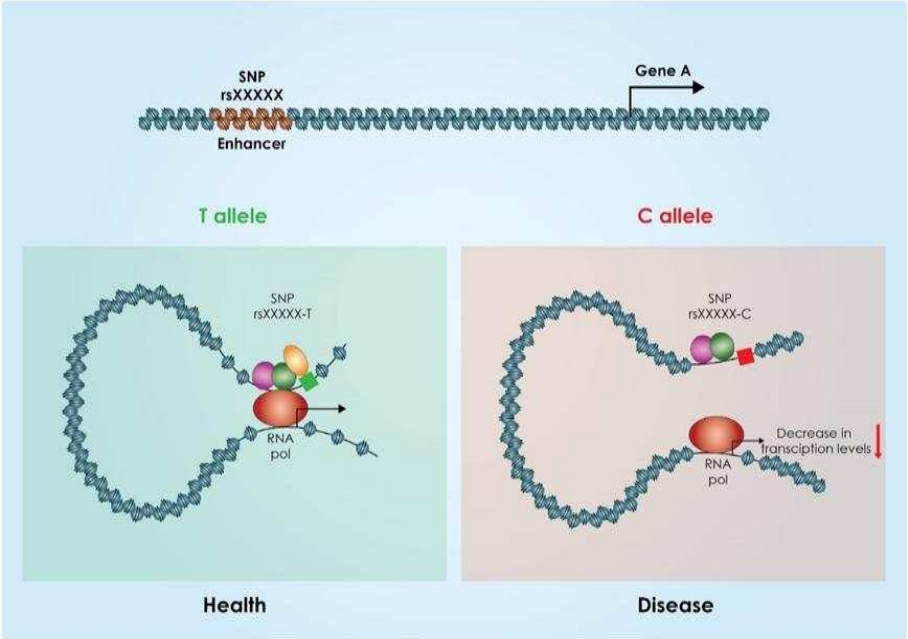

FIG. 3. (color online) Regulation of a gene by distal enhancer. In normal condition, defining Health, the enhancer can bind a set of transcription factors and by looping out can come into contact with the promoter site where RNA pol binds, thus the complex of the RNA polymerase and the transcription factors can start transcription of the Gene A.The DNA sequence of the enhancer is specific and must be intact, in order to be able to recruit the specific transcription factors. In the case that a change of bases from Thymidine (T allele), to Cytocine (C allele) has occurred at a site termed polymorphic site rsXXXXX, and defining Disease, then the enhancer sequence cannot bind a certain transcription factor, does not allow for chromatin loop formation and its efficiency in activating Gene A is halted. As a result, the level of the product of the Gene A drops accordingly and may cause disease. Adjusted from [15].

differ in the different stages of cell cycle or in development and this necessitates gene and chromatin movements within the nucleus. The active part of the genome is organized in defined domains of folded chromatin [18] that facilitate chromosome movement within the nucleus, as identified by 3C (chromosome conformation capture) and similar technologies with the application of bioinformatics and mathematical modelling. Such genetic complexes are the so called topologically associated domains (TADs), formed by the specific sites in the genome, insulators, that act as a barrier between non-transcribed and transcribed individual genes within euchromatin and may affect the function of an enhancer (Figure 4). Insulators recruit architectural proteins, can form self-interactions between two insulator sites, and can facilitate the formation of loops and greater chromatin fording. Such long-range gene regulation is crucial for the hierarchical functions of TADs in the development of an organism. Their dysregulation, e.g. by inappropriate promoter-enhancer communication and genomic alterations, may change gene expression patterns, e.g. overexpression of oncogenes and downregulation of tumor suppressor genes can lead to cancer [19].

2.5.2. Genetic complexity of the non-coding RNA

An important source of variation within the human genome sequence comes from the presence of non-coding, regulatory RNA [22, 23] in the genome. The analysis of the human genome has revealed that more than 75% of the genome is actively transcribed into RNA, which is not mRNA or other structural RNA. More detailed

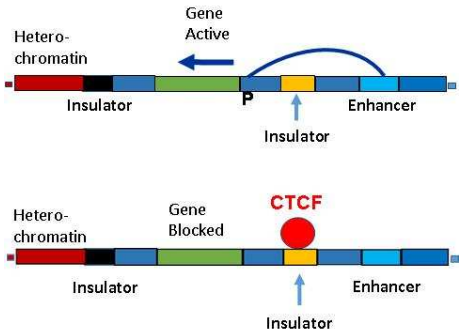

FIG. 4. (color online) A simplified model for the function of insulators. An insulator – black box – keeps apart the two basic chromatin states, heterochromatin and active genes domain. Another insulator – yellow box – is located between the promoter (P) of an active gene (beginning of transcription at the start of the arrow) and an enhancer of the gene's transcription (light blue box). The upper part depicts the interaction, through a loop, of the enhancer with the promoter P, for the start of transcription. The lower part depicts the binding of the architectural protein, insulator factor CTCF, the commonest and most important such protein, that binds to CCCTC sites residing in insulator consensus sequences, dispersed all over the genome. This results in the disruption of the formed P–enhancer contact and the blocking of the transcription of the gene.

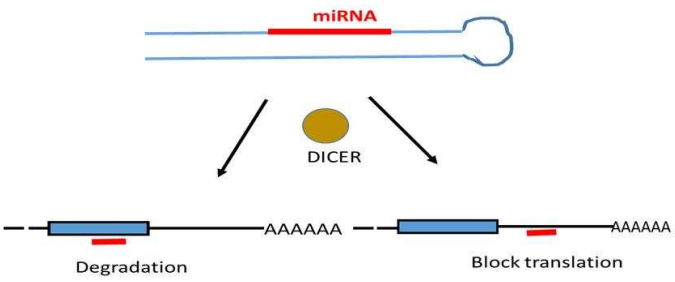

FIG. 5. (color online) miRNA represses/attenuates translation of mRNA, for the fine tuning of gene expression. An enzymatic system called DICER extracts from the hairpin and frees the miRNA, which then finds its matching place within a specific mRNA and pairs with it. The result depends on the position of the pairing site along the mRNA. If this site is within the coding sequencing, this mRNA is subjected to degradation; if it is within the non-coding 3' end of the message, then it blocks the translation of this message. These effects can be partial or complete, depending on the level of miRNA available.

studies of the transcriptome led to the discovery of a new, heterogeneous group of transcripts that do not code for proteins, the non-coding RNAs. These are mainly the "microRNA", miRNA, about 20 nucleotides long, and the "long non-coding RNA", lncRNA, about 200 nucleotides long, which are involved in a variety of functions within the framework of the regulation of gene expression and genome architecture, conferring robustness to biological processes [24].

Non-coding miRNA. Over 60% of genes are targets of miRNAs, that affect the fine tuning of their expression. The miRNAs are generated from their respective gene, as part of a hairpin structure (Figure 5).

Long non-coding RNA. The other major kind of regulatory RNAs, the lncRNA, has many regulatory functions in gene expression (Figure 6) and they also affect nuclear architecture by defining how specific protein-coding genes are organized in functional neighborhoods in the three-dimensional space of the living cell nucleus in healthy and diseased cells [25].

The non-coding RNA networks [26, 27] are currently under intense investigation, both for evolutionary studies as well as for the elucidation of their multiple roles in the development of disease.

2.6. Genome-wide association studies

Genome studies have revealed an extensive map of genetic variation within the human genome, with changes of all possible kinds, predominantly single nucleotide variants (polymorphisms) (SNP). Additionally, a plethora of non-coding RNA is being reported [27] that form the machinery for the fine tuning of gene expression and affect disease outcome. This wealth of data has been exploited by the use of computer software used to assemble the enormous sequence data in regions of high variability and to produce databases of human mapped variation (e.g. dbSNP). Such maps are used in the investigation of possible associations between a particular genome makeup, of specific haplotypes, with specific disease and/or specific phase. This

kind of studies refer to genome-wide association, whereby the whole genome is examined for disease associations. Alternatively, the analysis may be restricted, usually as a first approach, to the examination of exons only of the genome. Genome-wide association studies (GWAS) and exon-wide association studies (EWAS) form currently the main tools for determining the genetic complexity of a person/family/societal group for a given disease [28], and may enable better diagnosis and/or treatment.

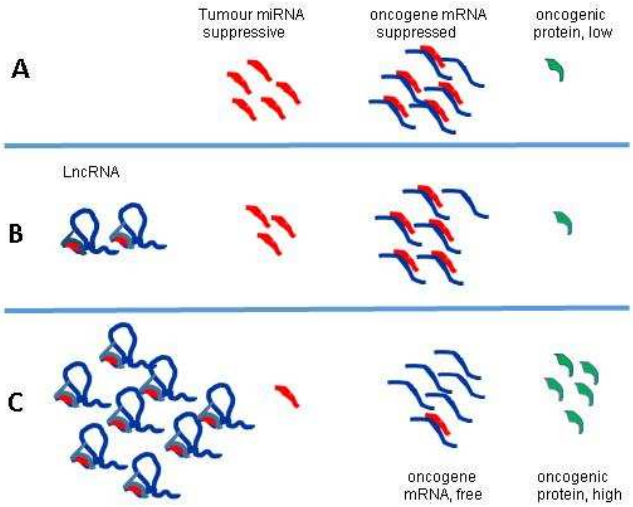

FIG. 6. (color online) The regulatory function of non-coding LncRNA. **A**: Oncogenic protein (green) is regulated (after transcription) by the binding of a tumour suppressive miRNA on the gene's mRNA, thus decreasing oncogenic protein to low level. **B**: Normal condition is ensured by the production of another regulatory non-coding RNA, the LncRNA Neat124, which binds to excess tumour suppressive miRNA, acting as a sponge and keeping the cell homeostasis normal. **C**: cancer condition appears when the LncRNA Neat1 is increased, and thus it absorbs more suppressive miRNAs and then more oncogene mRNA remains free to produce more oncogenic protein, leading to cancer. Adjusted from [25].

3. Precision medicine

The analysis of the genetic complexity of the human genome has provided medicine with molecular information which allows for more precise classification of a disease into smaller, defined entities and so improves the precision of patients categorization and treatment [29, 30]. This information derives from the genome variation among individuals, and, basically, reveals the changes that constantly occur in the genome with time, indicating that the genome as a whole of an individual is unique. The variation of the genome sequence refers to changes within the genes coding for proteins as well as to changes that are found to enhancers and insulators involved in the genome architecture and affecting gene regulation and chromatin localization within the nuclear architecture.

4. Concluding remarks

The sequencing and mapping of the human genome has revealed the genetic complexity of its content and has facilitated studies that advance our understanding of this complexity in its various manifestations. These studies have formed the basis for the generation of a new medical field namely the precision medicine, within the general personalized medicine framework and have important contribution to understanding of the modalities of gene transfer in cells in gene therapy applications.

References

[1] The Human Genome Project (HGP) from the Public Sector (https://www.genome.gov/human-genome-project)

and CELERA GENOMICS from the Private Sector.
[2] C. Lo *et al.* Strobe sequence design for haplotype assembly. BMC Bioinformatics. **12**, S24 (2011).
[3] M. J. P. Chaisson *et al.* Multi-platform discovery of haplotype-resolved structural variation in human genomes. Nat. Commun. **10**, 1784 (2019).
[4] X. Y. Woo *et al.* Genomic data analysis workflows for tumors from patient-derived xenografts (PDXs): challenges and guidelines. Med. Genomics. **12**(1): 92 (2019).
[5] P. McGonigle, B. Ruggeri. Animal models of human disease: challenges in enabling translation. Biochemical Pharmacology. **87**(1), 162-171 (2014).
[6] J. Dekker *et al.* Capturing chromosome conformation. Science. **295**, 1306-11 (2002).
[7] J. Dekker, M.A. Marti-Renom, L.A. Mirny. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. Nat. Rev. Genet. **14**, 390-403 (2013).
[8] C. S. Osborne *et al.* Active genes dynamically colocalize to shared sites of ongoing transcription. Nature Genetics. **36**, 1065-1971 (2004).
[9] J. Cairns *et al.* CHiCAGO: robust detection of DNA looping interactions in capture Hi-C data. Genome Biology. **17**, 127 (2016).
[10] B. Yang *et al.* MICRAT: a novel algorithm for inferring gene regulatory networks using time series gene expression data. BMC Systems Biology. **12** (Suppl 7), 115 (2018).
[11] D. Fehr *et al.* Classification-Based Inference of Dynamical Models of Gene Regulatory Networks G3: GENES, GENOMES, GENETICS Early online October 17, 2019. https://doi.org/10.1534/g3.119.400603.
[12] A. P. Delgado et al. Open reading frames associated with cancer in the dark matter of the human genome. Cancer Genomics Proteomics. 2014 Jul-Aug;11(4): 201–13.
[13] D.G. Lupiáñez *et al.* Disruptions of Topological Chromatin Domains Cause Pathogenic Rewiring of Gene-Enhancer Interactions. Cell. **161**(5), 1012-1025 (2015). doi: 10.1016/j.cell.2015.04.004.
[14] M. W. Vermunt et al. The interdependence of gene-regulatory elements and the 3D genome J Cell Biol. 2019 Jan 7; 218(1): 12–26.
[15] R.J. Palstra. F. Grosveld. Transcription factor binding at enhancers: shaping a genomic regulatory landscape in flux. Review, Frontiers in Genetics. **3**, 195 (2012)
[16] M. Kellis *et al.* Defining functional DNA elements in the human genome PNAS. **111**, 6131-6138 (2014).
[17] P.H.L. Krijger, W. de Laat. Regulation of disease-associated gene expression in the 3D genome. Nat. Rev. Mol. Cell Biology. **17**, 771-782 (2016). doi: 10.1038/nrm.2016.138.
[18] Q. Szabo, F. Bantignies, G. Cavalli. Principles of genome folding into topologically associating domains. Sci. Adv. **5**, eaaw1668 (2019).
[19] T. Ali, R. Renkawitz, M. Bartkuhn. Insulators and domains of gene expression. Current Opinion in Genetics & Development. **37**, 17-26 (2016).
[20] A. Sivakumar, J.I. de Las Heras, E.C. Schirmer. Spatial Genome Organization: From Development to Disease. Front. Cell Dev. Biol. **21**, 18 (2019)
[21] A. L. Valton, J. Dekker. TAD disruption as oncogenic driver. Curr. Opin. Genet. Dev. **36**, 34–40 (2016).
[22] I. R. Konig *et al.* What is precision medicine? Eur. Respir. J. **50**(4), pii: 1700391 (2017).
[23] Genya Dana. Precision Medicine Project Head, Precision Medicine Programme Center for the Fourth Industrial Revolution HE_PrecisionMedicine_4IROverview2017
[24] M.S. Ebert, P.A. Sharp. Roles on microRNAs in conferring robustness to biological processes. Cell. **149**, 27 (2012).
[25] C. Klec, F. Prinz, M. Pichler. Involvement of the long noncoding RNA NEAT1 in carcinogenesis Mol Oncol. **13**(1), 46–60 (2019).
[26] M. Sherafatian. Tree-based machine learning algorithms identified minimal set of miRNA biomarkers for breast cancer diagnosis and molecular subtyping. Gene. **677**, 111-118 (2018).
[27] M. Sherafatian, S. J. Mowla. The origins and evolutionary history of human non-coding RNA regulatory networks. J. Bioinform. Comput. Biol. **15**(2): 1750005 (2017).
[28] V. Tam et al. Benenefits and limitations of genome-wide association studies. Nat. Rev. Genet. **20**(8), 467-484 (2019). doi: 10.1038/s41576-019-0127-1.
[29] L. Cheng, B.P. Schneider, L. Li. Bioinformatics approach for precision medicine off-label drug drug selection among triple negative breast cancer patients. J. Am. Med. Inform. Assoc. **23**(4), 741-749 (2016). doi: 10.1093/jamia/ocw004.
[30] R. Nussinov *et al.* Review: Precision medicine and driver mutations: Computational methods, functional assays and conformational principles for interpreting cancer drivers. PLoS Comput Biol. **15**(3), e1006658 (2019).