INVESTIGATING INHERITED DISEASE RISKS USING COMPUTATIONAL GENETICS METHODOLOGY

Artomov M. [1, 2, 3, 4, 5]

[1]Almazov National Medical Research Center, Saint Petersburg, Russia
[2]World-Class Research Centre for Personalized Medicine, Saint Petersburg, Russia
[3]ITMO University, Saint Petersburg, Russia
[4]Broad Institute, Cambridge, USA
[5]Massachusetts General Hospital, Boston, USA

Corresponding author:
Artomov Mykyta,
Almazov National Medical Research Centre,
Akkuratova str. 2, Saint Petersburg,
Russia, 197341.
E-mail: artemov_nn@almazovcentre.ru

Received 03 September 2021; accepted 25 October 2021.

ABSTRACT

Large amount of genetic data accumulated over the recent years enabled the transition from association studies, aimed on the search for novel risk genes to the interpretation of personal risks and prognosis.

In this review the key milestones of computational biology are presented and the strengths and limitations of current genetic risk prediction methods are discussed.

Key words: bioinformatics, computational biology, genetics, GWAS, history of science, inherited risks.

For citation: Artomov N.N. Investigating inherited disease risks using computational genetics methodology. Russian Journal for Personalized Medicine. 2021;1(1):136-145.

INTRODUCTION

At the moment, the array of accumulated genomic data is becoming sufficient not only to understand the molecular causes of diseases, but also to assess individual congenital risks. Thus, there is a transition to the prediction of diseases and the personalization of medical interventions, especially in the field of preventive medicine.

The use of computational methods in biology has opened up opportunities for the development of a whole branch of research — bioinformatics. The development of methods for effective work with arrays of genomic and clinical data is the key to maximizing practical benefits from fundamental genetic research.

This review presents a brief history of the development of computational biology and the current state of the scientific field dedicated to the assessment of congenital disease risks.

DEVELOPMENT OF COMPUTATIONAL BIOLOGY AND BIOINFORMATICS

On September 19, 1957, Francis Crick outlined key ideas about the function of genes. In particular, he formulated the so-called "central dogma" of molecular biology, which states that genetic information can be transmitted from nucleic acids to protein, but not in the opposite direction [1, 2].

Despite the key role of DNA as a carrier of genetic information, the ability to establish the sequence of amino acids in a protein appeared much earlier than the methods of sequencing nucleic acids. For example, the Edman degradation, which became available in 1949 due to automation, resulted in sequences of more than 15 protein families in the following decade [3]. One of the important limitations of the Edman method is the inability to sequence more than 50-60 amino acids in one round [4]. In order to sequence proteins consisting of more amino acids, they had to be fragmented beforehand.

Thus, the main problem was not the sequencing process itself, but the assembly of a sequence of amino acids in a protein, based on hundreds of small polypeptides sequenced using the Edman method. This task gave the main impetus for the emergence of bioinformatics. The task of assembling a sequence of large proteins turned out to be extremely difficult for an analytical solution. This led to the fact that in the early 1960s, the first bioinformatics program was created, which allowed solving this problem automatically [2].

Margaret Dayhoff and Robert S. Ledley, considered the ancestors of modern computational biology, in 1962 created COMPROTEIN — the first computer program recorded on punch cards, which solved a problem that is relevant in present, *de novo* sequence assembly [5]. It should be noted that the introduction of software solutions and the existing limitations of computer performance led to innovations that we still use today, speaking about protein sequences. Thus, due to the complexity of storing three-letter names of amino acids in the memory of computers of that time, Margaret Deyhoff proposed a modern single-letter encoding of amino acids [6]. Following COMPROTEIN, the use of computational methods for the analysis of amino acid substitutions and sequences gave a significant impetus to the development of bioinformatics.

By 1970s, it became obvious that sequencing individual proteins was inefficient, since it required isolation and purification of each protein separately, while DNA sequencing would allow determining the amino acid sequence of all proteins simultaneously. The emergence of the Maxam-Gilbert [7] and Sanger [8] sequencing has led to the fact that the search for genes and the study of the effect of mutations in DNA on the occurrence of diseases have become one of the main directions in biology.

The first attempts to detect the position of genes in DNA and link mutations to the risk of disease began long before the start of the Human Genome Project. James Gusella, Nancy Wexler and colleagues in 1976-1983, during the study of Huntington's disease, a rare monogenic disease with autosomal dominant inheritance, were able to detect and describe the *HT gene*, in which mutations led to the onset of the disease. This was the first example of a gene associated with a genetic disease. The study analyzed the segregation of DNA markers with a phenotype in a large number of families with hereditary Huntingdon's disease [9]. Obviously, large-scale analysis of family trees is an extremely laborious task that can be successfully solved algorithmically, which led to the development of a number of methods for Linkage analysis) [10] and gene linkage mapping [11].

In the case of polygenic diseases, such an analysis is complicated by the need to monitor the joint segregation of many DNA markers simultaneously. The GENEHUNTER program made it possible to solve this problem and detect hundreds of DNA loci associated with diseases even before the completion of the Human Genome project [12].

With the advent of modern methods of high-performance sequencing, as well as genotyping using microchips, it became possible to accumulate a large array of data for associative studies. To date, the accumulated results of such studies have made it possible to detect thousands of genes and individual DNA variants associated with disease risks [13, 14].2 As a result, a deep-

er understanding of the molecular causes of impaired functioning of the body gradually leads to a shift in the focus of research towards the use of data on the effects of DNA variants on the phenotype to assess individual predispositions to diseases.

MONOGENIC DISEASES

For about 20% of the genes encoding proteins in the human genome, a connection with one or more phenotypes has been reliably established to date, which shows significant progress made over the past 40 years, but at the same time shows a huge amount of work that remains to be done.

Monogenic diseases can be caused by DNA variants of various types, from single-nucleotide polymorphisms to complex genomic rearrangements [15, 16]. In the case of a particular patient, the risk assessment of a monogenic disease consists in understanding the role and penetrance of individual DNA variants found in the gene associated with the disease. For example, hereditary breast cancer is often caused by individual DNA variants in the BRCA1 gene. The task of separating low-risk polymorphisms from high-risk rare DNA variants is one of the most relevant topics for studying the risks of hereditary diseases and formulating molecular diagnoses [17]. Currently, a large number of computational methods based on various sources of functional information (conservation of amino acids, functions of protein domains, etc.) are used for numerical risk assessment associated with a specific DNA variant [18—20]. However, checking the quality of such predictions based on computational models has remained difficult until recently.

Findlay and colleagues in 2018, using CRISPR technology, created all possible single-nucleotide mutants in the haploid HAP1 cell line sensitive to the normal functioning of the *BRCA1* gene [21]. By measuring the cellular viability for each mutant cell line, the functional importance of each of the DNA variants was assessed. This approach made it possible to reproduce the phenotype with high accuracy (~98%) for known clinically significant variants in *BRCA1* [22], as well as to evaluate the effectiveness of popular algorithms for functional annotation of DNA variants. Their low sensitivity remains the main problem for interpreting individual risks of monogenic diseases, but it is sufficient to prioritize DNA variants when searching for new risk genes.

In 2018, Cummings and co-authors were able to show that genomic sequencing analysis for making a molecular diagnosis in patients with muscular dystrophy is still insufficient for ~50% of patients. However, simultaneous analysis of genomic or exomic sequencing with RNA sequencing helps to identify the genetic cause of the disease for an additional 35% of patients [23].

DNA sequencing provides unique opportunities for making molecular diagnoses of monogenic diseases. The development of methods for prioritizing and classifying DNA variants, in particular in a non-coding DNA sequence, in the future will allow to reveal even more the value of genetic information for the medicine of monogenic diseases.

POLYGENIC DISEASES

Unlike monogenic diseases, the congenital component of the risk of complex (or polygenic) diseases is more associated with the cooperative action of a whole set of DNA variants, each of which has little effect. However, in some patients, despite the complex nature of the disease, the risk may be associated with rare variants that have a large phenotypic effect in one gene [24—26]. This kind of aggregation of risks from different groups of DNA variants makes it even more difficult to assess individual risks.

Information from one DNA variant is not enough to assess the risk of a polygenic disease. Instead of interpreting the genotypes of individual variants, an assessment of the genetic "load" is used using a value that includes all risky DNA variants. There are several approaches to combining information obtained from several DNA loci. The most common method is the polygenic risk score (PRS), a weighted sum of the number of risk alleles present in a patient [27]. In some cases, this criterion is sufficient to identify patients with a congenital polygenic risk comparable to the presence of highly penetrant mutations predisposing to monogenic diseases [28].

Initially, PRS was used in "case-control" studies to detect genetic differences between cohorts, which made it possible to confirm the importance of genetic risk factors for a wide range of complex diseases. This has become particularly important for research on the genetics of psychiatric diseases, which require tens of thousands of participants to achieve sufficient statistical power [29].

The use of this approach to identify patients with an increased risk of polygenic diseases faces a number of difficulties on the way to real application in clinical practice. One of the important limitations is the lack of a standardized translation of the relative risk provided by the PRS metric to absolute risk [30].

The possibility of transfer between populations of GWAS results, which are used as a reference for building a polygenic risk model, is also limited. Methods that allow making adjustments to the population structure have started to appear quite recently [31].

Models for predicting the risk of complex diseases, including a combination of clinical and biochemical factors, as well as lifestyle-related factors, work quite well in real conditions [32, 33]. However, the addition of PRS to these models can significantly help identify people at higher risk earlier, long before clinical symptoms occur [34].

Thus, the assessment of polygenic risks is in the process of transformation from a method that allows detecting a difference in genetic "load" in case-control type of studies, to a method that needs standardization and deeper understanding on its way to translation into clinical practice.

CONCLUSION

The development of methods for working with "big data" has become quite common in economics and statistics. The application of this approach in medicine and genetics is a booming field of research. The accumulation of data on genomic diversity, structuring and integration of clinical data are key steps for creating a fundamental base for personalized medicine.

Conflict of interest

The authors declare no conflict of interest.

REFERENCES

1. Cobb M. 60 years ago, Francis Crick changed the logic of biology. PLoS Biol. 2017; 15 (9): e2003243. DOI: 10.1371/journal.pbio.2003243.
2. Gauthier J, Vincent AT, Charette SJ, et al. A brief history of bioinformatics. Brief Bioinform. 2019; 20(6):1981–1996. DOI: 10.1093/bib/bby063.
3. Hagen JB. The origins of bioinformatics. Nat Rev Genet. 2000;1(3):231–236. DOI: 10.1038/35042090.
4. Edman P, Begg G. A protein sequenator. Eur J Biochem. 1967;1(1):80–91. DOI: 10.1007/978-3-662-25813-2_14.
5. Dayhoff MO, Ledley RS. Comprotein: A computer program to aid primary protein structure determination. AFIPS. 1962;1:262–274. DOI: 10.1145/1461518.1461546.
6. IUPAC-IUB Commission on Biochemical Nomenclature. A one-letter notation for amino acid sequences. Tentative rules. Biochemistry. 1968;7(8):2703–2705. DOI: 10.1021/bi00848a001.
7. Maxam AM, Gilbert W. A new method for sequencing DNA. Proc Natl Acad Sci U S A. 1977 Feb;74(2):560–564. DOI: 10.1073/pnas.74.2.560.
8. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A. 1977;74(12):5463–5467. DOI: 10.1073/pnas.74.12.5463.
9. Gusella JF, Wexler NS, Conneally PM, et al. A polymorphic DNA marker genetically linked to Huntington's disease. Nature. 1983;306(5940):234–238. DOI: 10.1038/306234a0.
10. 10. Bryant SP. Software for genetic linkage analysis. In: Mathew CG. (eds) Protocols in human molecular genetics. Methods in Molecular Biology. Springer, Totowa, NJ, 1991;9:403–418. doi: 10.1385/0-89603-205-1:403.
11. Lander ES, Green P Abrahamson J, et al. MAPMAKER: An interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. Genomics. 1987;1(2):174–181. DOI: 10.1016/0888-7543(87)90010-3.
12. Kruglyak L, Daly MJ, Reeve-Daly MP, et al. Parametric and nonparametric linkage analysis: a unified multipoint approach. Am J Hum Genet. 1996;58(6):1347–1363.
13. Hamosh A, Scott AF, Amberger JS, et al. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res. 2005;33(Database issue):D514-7. DOI: 10.1093/nar/gki033.
14. MacArthur J, Bowler E, Cerezo M, et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res. 2017;45(D1):D896-D901. DOI: 10.1093/nar/gkw1133.
15. Posey JE. Genome sequencing and implications for rare disorders. Orphanet J Rare Dis. 2019;14(1):1–10. DOI: 10.1186/s13023-019-1127-0.
16. Lupski JR. Structural variation mutagenesis of the human genome: Impact on disease and evolution. Environ Mol Mutagen. 2015;56(5):419–436. DOI: 10.1002/em.21943.
17. Cooper DN, Krawczak M, Polychronakos C, et al. Where genotype is not predictive of phenotype: towards an understanding of the molecular basis of reduced penetrance in human inherited disease. Hum Genet. 2013;132(10):1077-130. DOI: 10.1007/s00439-013-1331-2.
18. Adzhubei IA, Schmidt S, Peshkin L, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010;7(4):248–249. DOI: 10.1038/nmeth0410-248.
19. Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res. 2003;31(13):3812–3814. DOI: 10.1093/nar/gkg509.
20. Samocha KE, Kosmicki JA, Karczewski KJ, et al. Regional missense constraint improves variant deleteriousness prediction. bioRxiv. 2017;148353. DOI: 10.1101/148353.
21. Findlay GM, Daza RM, Martin B, et al. Accurate classification of BRCA1 variants with saturation genome editing. Nature. 2018;562(7726):217–222. DOI: 10.1038/s41586-018-0461-z.
22. Landrum MJ, Lee JM, Benson M, et al. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res. 2018;46(D1):D1062-D1067. DOI: 10.1093/nar/gkx1153.
23. Cummings BB, Marshall JL, Tukiainen T, et al. Improving genetic diagnosis in Mendelian disease

with transcriptome sequencing. Sci Transl Med. 2017;9(386):eaal5209. DOI: 10.1126/scitranslmed.aal5209.

24. Saint Pierre A, Génin E. How important are rare variants in common disease? Brief Funct Genomics. 2014;13(5):353–361. DOI: 10.1093/bfgp/elu025.

25. Rivas MA, Beaudoin M, Gardet A, et al. Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. Nat Genet. 2011;43(11):1066–1073. DOI: 10.1038/ng.952.

26. Singh T, Neale BM, Daly MJ. Exome sequenc26. ing identifies rare coding variants in 10 genes which confer substantial risk for schizophrenia. medRxiv. 2020.09.18.20192815. DOI: 10.1101/2020.09.18.20192815.

27. Lewis CM, Vassos E. Polygenic risk scores: from research tools to clinical instruments. Genome Med. 2020;12(1):44. DOI: 10.1186/s13073-020-00742-5.

28. Khera AV, Chaffin M, Aragam KG, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nat Genet. 2018;50(9):1219–1224. DOI: 10.1038/s41588-018-0183-z.

29. Purcell SM, Wray NR, Stone JL, et al. Common polygenic variation contributes to risk of schizophrenia thatoverlaps with bipolar disorder. Nature. 2009;460(7256):748–752. DOI: 10.1038/nature08185.

30. Torkamani A, Wineinger NE, Topol EJ. The personal and clinical utility of polygenic risk scores. Nat Rev Genet. 2018;19(9):581–590. DOI: 10.1038/s41576-018-0018-x.

31. Martin AR, Kanai M, Kamatani Y, et al. Clinical use of current polygenic risk scores may exacerbate health disparities. Nat Genet. 2019;51(4):584–591. DOI: 10.1038/s41588-019-0379-x.

32. Assmann G, Cullen P, Schulte H. Simple scoring scheme for calculating the risk of acute coronary events based on the 10-year follow-up of the prospective cardiovascular Münster (PROCAM) study. Circulation. 2002;105(3):310–305. DOI: 10.1161/hc0302.102575.

33. Wilson PWF, Meigs JB, Sullivan L, et al. Prediction of incident diabetes mellitus in middle-aged adults: the Framingham Offspring Study. Arch Intern Med. 2007;167(10):1068–1074. DOI: 10.1001/archinte.167.10.1068.

34. Mars N, Koskela JT, Ripatti P, et al. Polygenic and clinical risk scores and their impact on age at onset and prediction of cardiometabolic diseases and common cancers. Nat Med. 2020;26(4):549–557. DOI: 10.1038/s41591-020-0800-0.

Author information:

Artomov Mykyta, Head of Research Laboratory of Population Genetics, Almazov National Medical Research Center, Research-professor, ITMO University, Researcher, Broad Institute, Instructor in Investigation, Massachusetts General Hospital.