Genomic approach to complex diseases

Szalai Csaba

Semmelweis University, Department of Genetics, Cell and Immunobiology and Heim Pál Hospital; Budapest, Hungary

Abstract- **The complex or multifactorial diseases are those which develop through interactions of often hundreds of genes and environmental factors. The complex diseases like cancer, asthma, hypertension, diabetes mellitus, cardiovascular diseases or Alzheimer disease are often very frequent, it can even be said that more or less everybody is affected by some of them. In this review it is discussed why it is important to study the genomic background of the complex diseases and the main genomic methods are summarized. Next, the difficulties of these studies are shown and discussed what the reason of the missing heritability of the complex diseases can be. In the end some developments are shown which try to cope with these problems.**

Keywords- **multifactorial diseases, genetic markers, genetic variants, genomic methods, missing heritability.**

I. Introduction

The complex or multifactorial diseases are those which develop through interactions of a few (oligogenic) or several (polygenic) genes and the environmental factors. The complex diseases, in contrast to the monogenic diseases, which affect only a small fraction of the population, are often very frequent, it can even be said that more or less everybody is affected by them. Complex diseases are the endemic, non-communicable diseases, or NCD, which are non-infectious and non-transmissible between persons, like cancer, asthma, hypertension, diabetes mellitus, cardiovascular diseases and Alzheimer disease, etc.

First, one can ask, why it is important to study the genomic background of the complex diseases? Perhaps the most important is that it helps to explore the molecular pathomechanism. In contrast to the traditional methods, the genomic methods are often hypothesis free, i.e. they do not require any knowledge about the pathogenesis. In this way novel pathways and mechanisms can be detected, which can offer new drug targets or new therapies. Otherwise, the genomic studies can reveal the genetic differences between people, offering novel possibilities for personal therapies, and connections can be found between the success of the therapy and the genetic background. Genomic studies can reveal genetic variations which influence the risk of developing a disease. In this way, right after the birth the genomic background and the risk to different diseases of a new-born can be determined, which offers the possibilities to change from “diagnose and treat” to “predict and prevent”. Earlier it was regarded as the most important task of the medical genomics, but later it turned out that in most cases the sum risk to a multifactorial disease is so complex that it is usually impossible to give a clinically relevant estimation.

As for both researchers and the whole society the significance of genomic results are widespread appreciated, this has led to a large-scale effort for the development of genomic methods and huge breakthroughs have been achieved.

II. Genomic studies

A. Genetic markers

A genetic marker is usually a sequence variation with a known location on a chromosome that can be used to identify individuals, with a relative high chance to differentiate between different alleles on homologous chromosomes. Genetic markers can be used to study the relationship between an inherited disease and its genetic cause (for example, a particular mutation of a gene that results in a defective protein). It is known that pieces of DNA that lie near each other on a chromosome tend to be inherited together (they are linked). This property enables the use of a marker, which can then be used to determine the precise inheritance pattern of the gene that has not yet been exactly localized. Genetic markers have to be easily identifiable, associated with a specific locus, and highly polymorphic, because homozygotes do not provide any information.

One of the most popular markers are the microsatellites, or simple sequence repeats (SSRs) or short tandem repeats (STRs), are repeating sequences of 2-6 base pairs of DNA. Often they are very polymorphic, meaning that individuals are often heterozygotic to them, which means that they differ in the number of repeats.

They are widely used in mapping disease genes or differentiate between individuals. The human genome is now mapped by approximately 30,000 highly polymorphic microsatellites. The average length of linkage disequilibrium (LD) for microsatellites is ~100 kb, which is considerably higher than that of SNPs. Therefore, a single microsatellite captures a larger genomic region than does a single SNP. Microsatellites also provide several other advantages, such as a higher information content (6–10 alleles as compared with 2 alleles for SNPs), and a smaller interpopulation variability. Most existing forensic DNA databases are STR-based. It has been demonstrated that 20–50 ascertained autosomal SNPs could reach match probabilities similar to those obtained with 10–15 forensically used STRs.

But the disadvantages of the STRs are that the detection methods are quite complicated relative to those of the SNPs,

they are much rarer than SNPs, and their mutation rates are 100,000 times higher.

Nowadays, the advantages of the SNPs are much more significant, and mainly because of their number and simple detection techniques, they will replace STRs in most areas. E.g. forty-five unlinked autosomal SNPs were ascertained by screening more than 500 candidate SNPs in 44 worldwide populations. These 45 ascertained SNPs have high levels of heterozygosity and low levels of population differentiation and are therefore suitable for universal human identification purposes. Multiplex genotyping assays for these SNPs have been developed.

B. Study of genetic variants

Genetic variations play important roles in disease susceptibilities, differences between individuals or in responses to drugs, and the study of them is important in discovery of novel drug targets, personal therapies or pharmacogenetics, etc. The HGP and the subsequent different genome projects (Human Variome Project, HapMap, 1000 Genome project, etc.) detected millions of genetic variants [1,2]. Presently, there are more than 65 million short variants, and more than 10 million structural variants in the databases.

The simplest method for the study of the genetic background of a disease is the candidate gene association study. In these studies genes are selected, which are thought to play a role in the disease. Then, genetic variations are searched in these genes. Earlier the genes were sequenced in several individuals, now the databases contain practically all the common variants. The first one is often called wet laboratory method, the latter one *in silico* method. Then, the selected variants are genotyped, and their frequencies are compared in the population with and without the studied trait (disease). If the frequencies of the variants differ in a statistically significant way between the two populations, then they are suspected to play a role in the disease susceptibility. Several 10 thousand such investigations have been carried out in the last decades in different diseases. But, there were a lot of problems with these studies. One of the problems is the multiple testing problems, but in a different way than discussed in connection with GWAS (s. later). Because here, the same variants have been tested in different laboratories, and naturally only the positive results have been published; the negative ones have been discarded. And, if 100 laboratories study the same variants, there is a chance that one of them gets a positive association purely by chance. This is called publication bias. Because of this, hundreds of false positive results (and genes) have been published.

The other problem is that with this methods only those genes can be studied whose role was already known in the disease, and in this way no new mechanism could be detected.

The hypothesis-free genomic methods theoretically could solve this last problem. First, whole genome screenings were developed and carried out in several diseases. In this method families were screened with microsatellites. Those families were recruited where there were at least two affected siblings. These studies are also called affected sib pair (ASP) studies, or linkage studies. Here LOD scores were calculated. The LOD score (logarithm (base 10) of odds) is a statistical test often used for linkage analysis. The LOD score compares the likelihood of obtaining the test data if the two loci, or the disease phenotype and a locus are indeed linked, to the likelihood of observing the same data purely by chance. Positive LOD scores favor the presence of linkage, whereas negative LOD scores indicate that linkage is less likely. A LOD score greater than 3.0 is considered evidence for linkage. A LOD score of +3 indicates 1000 to 1 odds that the linkage being observed did not occur by chance. On the other hand, a LOD score of less than -2.0 is considered evidence to exclude linkage.

The method has given a lot of interesting results, but there have been several problems with it. First, it is difficult to collect families with two affected siblings, second, the genotyping of the microsatellies are very cumbersome and expensive. Because of this latter, the number of microsatellites in the studies was limited (usually not more than 400), thus the resolution was very low. This means that it was a great chance that disease associated loci, which were not in linkage with any of the microsatellites were lost. In addition, these studies could determine only genomic regions (because of the limited number of markers), and not genes. And often, these regions are large, several megabase long and contain several hundreds of genes. In this way, additional methods are needed for the determination of the genes.

C. GWAS

Presently, the most popular method for the study of the genomic background of complex diseases and traits is called GWAS (genome-wide association study), also known as whole genome association study (WGA study or WGAS). The method has become possible, when arrays and chips have been developed with which first 100 thousand, then several million SNP could be genotyped in one measurement, and the price of one chip has become relatively cheap, i.e. about $100. First, only SNPs were determined, later, when the significance of CNVs became apparent, they were involved as well. The CNVs were determined through their known linkage with SNPs. In 2007 this method was selected for the breakthrough of the year.

There are two main companies in the markets, Affymetrix and Illumina. The Affymetrix Genome-Wide Human SNP Array 6.0 features 1.8 million genetic markers, including more than 906,600 SNPs and more than 946,000 probes for the detection of CNVs.

The Illumina HumanOmni5-Quad (Omni5) BeadChip can detect 4.3 million tagSNPs selected from the International HapMap and 1000 Genomes Projects that target genetic variation down to 1% minor allele frequency (MAF).

In GWAS the distribution (frequencies) of the variants is compared in the different populations; usually one of them is affected with the trait, the other is not. But, with the development of the statistical methods GWAS has become capable of studying the genomic background of continuous

traits (like fasting glucose levels or blood pressure) as well. In this latter case there are no different groups.

GWAS has been offering a great chance for the investigation of the genomic background of the diseases, which have been utilized by a lot of research groups and consortia. Because of the strict statistical conditions and the large investigated populations, the results of GWAS may contain only few false results; and because this is a hypothesis-free method, there is a possibility that it reveals new aspects of the disease. To make these important results public, a web page was established on 25 November 2008 (A Catalog of Published Genome-Wide Association Studies) [3], and it includes only those publications which investigate at least 100,000 SNPs in the initial stage. Publications are organized from most to least recent date of publication, indexing from online publication if available. Studies focusing only on candidate genes are excluded from this catalog. Studies are identified through weekly PubMed literature searches, daily NIH-distributed compilations of news and media reports, and occasional comparisons with an existing database of GWAS. SNP-trait associations listed here are limited to those with p-values < 1.0 × 10-5. In 2013 the catalog contained about 1,800 curated publications of 12,000 SNPs [4]. In March 2015, the GWAS Catalog infrastructure was migrating to the European Bioinformatics Institute (EMBL-EBI).

D. Evaluation of GWAS results

The evaluation and handling of GWAS data are a great challenge for the bioinformaticians. One of the main problems is the multiple testing problem. If the p value of a SNP corresponds to the Bonferroni corrected value, then it is said that it reached the level of genome wide significance. It is, e.g. in case of 1 million SNPs 5×10^{-8}. As the main characteristics of the complex diseases are variants with weak effects, this low p value often can only be achieved through involving large populations. Often the number of participants must be >100,000, which is very difficult and expensive to collect, and which is in case of rarer diseases even impossible. Because of this, GWAS are often carried out by large international consortia.

A method to attenuate this problem can be, if several smaller populations are investigated independently. In this way the p values in the independent studies for each SNP are multiplied, and it is easier to achieve the low values (e.g. $10^{-3} \times 10^{-3} = 10^{-6}$). Usually, a discovery GWAS is carried out in a smaller population (discovery cohort). Then, SNPs are selected with a not so strict p value (e.g. cut off value < 5×10^{-2}), then several independent populations are collected (replication cohorts), and only the selected SNPs are studied. The SNPs which are confirmed in the replication cohorts can be those which are associated with the disease.

New statistical methods are also under development, such as Bayesian statistics and pathway analysis. For this latter, several databases are available like Gene Ontology (GO) [5] or KEGG (Kyoto Encyclopedia of Genes and Genomes [6].

Gene Set Enrichment Analysis (GSEA) is a computational method, which was originally developed for gene expression studies and can be applied in GWAS as well. This determines whether different a priori defined sets of genes show statistically significant, concordant differences between two biological states (e.g. phenotypes). Then the sets of genes are ranked according to their associations.

With these methods several new disease associated pathways have been detected.

E. DNA sequencing

DNA sequencing is the process of reading the nucleotide bases in a DNA molecule. Since the beginning of the HGP it has been developing continuously. In HGP the DNA was sequenced with Sanger method, i.e. with dideoxy or chain termination sequencing. In 2001 the sequencing of one human genome took a minimum of 1 year. It was obvious that both the price and the time were not appropriate for routine investigations, or even for sequencing several human genomes. It became clear that the Sanger method could not be developed much further to become much cheaper and faster. But it was also obvious that much cheaper and faster sequencing would have an immense leap in pharmaceutical research, personal medicine, but it could be used for countless aims. The high demand for low-cost sequencing has driven the development of high-throughput sequencing (also called as next-generation sequencing, or NGS) technologies that parallelize the sequencing process, producing thousands or millions of sequences at once. The methods were so successful that in 2007 the new generation sequencing (NGS) became the method of the year [7] in *Nature Methods* magazine. In 2007 the genome of James Watson was sequence with the 454 technology in 2 months and for $1 million. It was still far away from the aim, but it was a big step ahead. Since then, the price has been lower and lower, and the time shorter and shorter (Figure 2). E.g. in June 2009, Illumina announced that they were launching their own Personal Full Genome Sequencing Service at a depth of 30× for US $48,000 per genome.

In November 2009, Complete Genomics published a peer-reviewed paper in *Science* demonstrating its ability to sequence a complete human genome for US$1,700. If true, this would mean the cost of full genome sequencing has come down exponentially within just a single year from around US$100,000 to US$50,000 and now to US$1,700.

In 2011 Complete Genomics charges approximately US$10,000 to sequence a complete human genome (less for large orders).

In May 2011, Illumina lowered its Full Genome Sequencing service to US$5,000 per human genome, or US$4,000 if ordering 50 or more.

In January 2012, Life Technologies introduced a sequencer to decode a human genome in one day for $1,000 and now several examples of other equipment are also capable for this.

F. Difficulties in the studies of the genomic background of complex diseases

At the beginning of the genomic era, even right after the completion of the HGP, it was generally thought that

genomic would revolutionize the medicine, and in a few years the era of personal therapy would come. But now we know that it did not come true, and even it would not in the next years. What can be the reason for this?
According to the general opinion, one of the main reasons for this failure is due to the very complex regulation of the genome, and the multifactorial nature of the diseases and traits. In Table 1 there are some characteristics which make the determination of the genetic background of the multifactorial diseases difficult.

Table 1 Factors, which make the determination of the genetic backgrounds of the complex diseases difficult

Problems	Explanation
Genetic heterogeneity	Different allelic combinations lead to similar phenotypes.
Phenocopy	Environmental factors lead to the same clinical phenotype as do the genetic factors. In other words, the environmental condition mimics the phenotype produced by a gene.
Pleiotropy	The genetic variation can lead to different phenotypes.
Incomplete penetrance	Some individuals fail to express the trait, even though they carry the trait associated alleles.
The exact diagnosis is difficult	Often in complex diseases there are no standard diagnoses. There are subtypes of the diseases that cannot be differentiated with standard methods. The symptoms can change with the time, or manifest in episodes. Different diseases with similar symptoms. Concordance of different diseases.

As for both researchers and the whole society the significance of genomic results are widespread appreciated, this has led to a large-scale effort for the development of genomic methods and huge breakthroughs have been achieved.
But there is no reason for the total satisfaction, since most of the aims have not been achieved. In 2009, Manolio et al. published a widespread cited table in a paper, which summarizes the results of studies aiming at determining the genomic background of multifactorial diseases and traits [8]. These results show that the GWAS, which were thought to be the very method for determining the genomic background of complex traits, could determine only a small fraction of the heritability proportion of the majority of the traits. It means that most variants identified until then conferred relatively small increments in risk, and explained only a small proportion of familial clustering, leading many to question how the remaining, 'missing' heritability can be explained. And the situation has not improved considerably since then. E.g. height is one of the QTs which is easy to determine, and it is known that the heritability of it is about 80%. In several studies, large populations were collected and several GWAS were carried out. In one study, 44 loci were determined, which were responsible only for 5% of the heritability. Later, 180 loci could be determined, but they were still responsible only for 10% of the heritability. This is true for the majority of the diseases. E.g. this value for T2DM is 6%, for fasting glucose level is 1.5%, for early myocardial infarction is 2.8%. The exceptions are diseases, where there are only a couple of mutations with strong impact, like in the case of macular degeneration. In contrast, the determination of the genetic background of monogenic diseases is a great success; it has been clarified for about 4000 such diseases so far.
What can be the reason for this situation, which is often called the dark matter of heritability? Previously, some explanations have been already mentioned and below some additional ones will be given.

G. Problems of the rare variants

GWAS work with pre-made chips, which could determine known variations with a population frequency of >5% (MAF = minor allele frequency). There is a theory named common disease - common variants or CD/CV, which says that common diseases are caused by several common (frequent) variants with weak effects. The weak effects of these variants are accumulated causing higher susceptibility to a disease. If the environmental factors are unfavourable, then the disease can develop. It proved to be true for a lot of traits, like Alzheimer disease, where the roles of the common apoE4 variants or the obesity where the roles of variations in the FTO genes were verified. But, there are also proofs for the so-called common disease rare variants hypothesis (CD/RV), which states that the common diseases are caused by rare variants with strong effects [9]. Example is the breast cancer where thousands of rare variants with strong effects have been found. The rare variants cannot be determined with GWAS, and the traditional statistical methods are not suitable for their detection. It is suggested that even in diseases, where common variations are known, there are also rare variations with strong effect.
The rare variants can also cause another statistical problem called synthetic associations. In this case rare variants at the locus create multiple independent association signals captured by common tagging SNPs causing that variants which do not participate in the given phenotype, will be falsely named.

H. The random behavior of the genome

In September 2010 researchers published in Nature that genetic circuits that regulate cellular functions are subject to stochastic fluctuations, or ‘noise’, in the levels of their components [10]. It means that the behavior of the genome is sometimes random and thus cannot be predicted in 100%. It means that it is theoretically impossible even with more developed genomic and informatic methods to exactly forecast the future traits (phenotypes) of a newborn.

I. Statistical problems

The next problem originates from the evaluation methods, i.e. from the statistics. The most variations associated with increased risk to complex diseases, increase the risk with only 10-20%. It means that the chance in the carriers for the

development of the disease is only 1.1-1.2 times higher than in non-carriers. Detecting variations with such weak effects is very difficult. In addition, as the population is genetically heterogeneous, and interactions between these variants are needed, the possible number of genetic backgrounds associated with increased risk is practically infinite. In statistical point of view it is advantageous if the population is larger, but the larger population is genetically more heterogeneous, thus the effect of each genetic variant is diluted, becoming less significant and may be lost.

The other problem is the lack of proper statistical methods. One problem is called the multiple testing problem.

If in a GWAS 100 thousand genetic variations are measured, in a statistical point of view it means that 100 thousand independent measurements are carried out. In this case the probabilities of the false results are summed up. In statistics, $p < 0.05$ is used as a significance threshold. It means that the probability of the false statement is 5% (we can make a false statement 5 times in 100 independent investigations). One of the methods to correct this is called Bonferroni correction. In this case, 0.05 is divided by the number of the measurements (in this case with 100 thousand; $p = 0.05/100.000 = 5\times10^{-7}$). But the number of the independent investigations depends not only on the number of the measurements, but on several other factors, like the number of the samples, the clinical parameters and the type of tests, etc. But the Bonferroni correction is too conservative, i.e. if the correction is applied, only the strongest effects can be detected. In contrast, according to the CD/CV hypothesis the complex diseases develop through interactions between multiple genetic variants with weak effects and the environment. In addition, as the genetic factors interact with each other, if we want to calculate this interaction as well, it would increase the number of independent questions to a very large number. It means that the Bonferroni corrections and the similar other methods are not capable of detecting the variants of weak effects, i.e. other methods are needed.

J. Possible solutions

There are several developments which try to cope with the above mentioned problems. E.g. utilizing the results of the 1000 Genome Project, new chips are under development, which can measure rarer (MAF < 0.05) variants as well (e.g. Illumina 5M chip). Furthermore, next to genotyping based methods, the new generation sequencing (NGS) may be soon suitable for population based studies. With the NGS, all type of variations can be detected. It must be added, however, that the statistical problems are even larger with this method, since it can give terabit size of data and hundreds of thousands of variations, many of which can be sequencing mistakes, or unknown variations whose functional characterizations are immensely difficult.

There are a couple of new solutions for the statistical problems as well. E.g. to overcome several of the limitations, probabilistic graphical models (PGMs) were proposed. Thanks to their ability to efficiently and accurately represent complex networks, PGMs represent powerful tools to dissect the genetic susceptibility of complex diseases. Bayesian networks are a popular class of PGMs, its graphical representation presents a crucial advantage and is able to efficiently deal with SNP–SNP interactions impacting the phenotype, a situation that is called epistasis. As Bayes statistics can evaluate networks, it is a suitable evaluation method for systems biology [11-13].

It is assumed that with better statistics significantly more information can be extracted even from the present results. E.g. in a paper it has been stated that from the old results but with better statistics they could explain 67% of the heritability of height, in contrast the 5% in the original paper. In this paper rather than considering SNPs one by one, the new statistical analysis considers what effect all the SNPs together have on height [14].

In another paper the genetic background of hypertension was studied. They reevaluated the results of a metaanalysis of several GWAS, which did not find any associated variants (owing to the too conservative Bonferroni correction, and the heterogeneous nature of this disease). In the new statistics the authors did not consider individual SNPs, but examined whether there are pathways where the distribution of the variations are statistically different in the hypertensive population relative to the controls. In this paper several pathways were found associated with the disease [15].

It is also a great challenge that the majority (~93%) of disease- and trait-associated variants emerging from these studies lie within non-coding sequence. It is therefore very difficult to explain how these variants influence the trait. In a study of the ENCODE project it was found that in a given cell line, 76.6% of all non-coding GWAS SNPs either lie within a DNase I hypersensitive site (DHS) (57.1% or 2931 SNPs), or are in complete linkage disequilibrium (LD) with SNPs in a nearby DHS [16]. DHSs show remarkable concordance with experimentally determined and computationally predicted binding sites of transcription factors and enhancers. With the help of the results of the ENCODE and similar other projects it will be much easier to determine the function of a variant lying in non-coding region of the genome.

ACKNOWLEDGMENT

This study was supported by OTKA (Hungarian Scientific Research Fund): K112872

CONFLICT OF INTEREST

The author declares that he has no conflict of interest.

REFERENCES

1. Venter JC, Adams MD, Myeers EW et al (2001) The sequence of the Human Genome. Science 291:1304-51

2. International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome Nature 431: 931 - 945
3. http://www.genome.gov/gwastudies/
4. Welter D, MacArthur J, Morales J et al. (2014) The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res 42 (Database issue):D1001-6
5. http://www.geneontology.org
6. http://www.genome.jp/kegg/
7. http://www.nature.com/nmeth/journal/v5/n1/full/nmeth1157.html
8. Manolio TA, Collins FS, Cox NJ, et al. (2009) Finding the missing heritability of complex diseases. Nature 461(7265):747-53
9. McClellan J, King MC (2010) Genetic heterogeneity in human disease. Cell 141(2):210-7.
10. Eldar A, Elowitz MB (2010) Functional roles for noise in genetic circuits. Nature 467(7312):167-73.
11. Ungvári I, Hullám G, Antal P et al. (2012) Evaluation of a partial genome screening of two asthma susceptibility regions using Bayesian network based Bayesian multilevel analysis of relevance. PLoS One 7(3):e33573.
12. Lautner-Csorba O, Gézsi A, Semsei AF, et al. (2012) Candidate gene association study in pediatric acute lymphoblastic leukemia evaluated by Bayesian network based Bayesian multilevel analysis of relevance. BMC Med Genomics 5(1):42.
13. Lautner-Csorba O, Gézsi A, Erdélyi DJ et al. (2013) Roles of genetic polymorphisms in the folate pathway in childhood acute lymphoblastic leukemia evaluated by bayesian relevance and effect size analysis. PLoS One. 8(8):e69843.
14. Yang J, Benyamin B, McEvoy BP et al. (2010) Common SNPs explain a large proportion of the heritability for human height. Nat Genet 42(7):565-9.
15. Torkamani A, Topol EJ, Schork NJ (2008) Pathway analysis of seven common diseases assessed by genome-wide association. Genomics 92(5):265-72.
16. ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489(7414):57-74.

Author: Csaba Szalai
Institute: Semmelweis University, Department of Genetics, Cell and Immunobiology
Street: Nagyvárad tér 4.
City: Budapest
Country: Hungary
Email: szalaics@gmail.com