Biochimica et Biophysica Acta 1842 (2014) 1889–1895

Contents lists available at ScienceDirect

Biochimica et Biophysica Acta

journal homepage: www.elsevier.com/locate/bbadis

Review

The genome revolution and its role in understanding complex diseases☆

Marten H. Hofker [a,*], Jingyuan Fu [b], Cisca Wijmenga [b]

[a] University of Groningen, University Medical Center Groningen, Department of Molecular Genetics, Groningen, The Netherlands
[b] University of Groningen, University Medical Center Groningen, Department of Genetics, Groningen, The Netherlands

ARTICLE INFO

Article history:
Received 4 February 2014
Received in revised form 30 April 2014
Accepted 6 May 2014
Available online 13 May 2014

Keywords:
Genome
Genetic variation
Complex disease
Disease etiology
Functional analysis

ABSTRACT

The completion of the human genome sequence in 2003 clearly marked the beginning of a new era for biomedical research. It spurred technological progress that was unprecedented in the life sciences, including the development of high-throughput technologies to detect genetic variation and gene expression. The study of genetics has become “big data science”. One of the current goals of genetic research is to use genomic information to further our understanding of common complex diseases. An essential first step made towards this goal was by the identification of thousands of single nucleotide polymorphisms showing robust association with hundreds of different traits and diseases. As insight into common genetic variation has expanded enormously and the technology to identify more rare variation has become available, we can utilize these advances to gain a better understanding of disease etiology. This will lead to developments in personalized medicine and P4 healthcare. Here, we review some of the historical events and perspectives before and after the completion of the human genome sequence. We also describe the success of large-scale genetic association studies and how these are expected to yield more insight into complex disorders. We show how we can now combine gene-oriented research and systems-based approaches to develop more complex models to help explain the etiology of common diseases. This article is part of a Special Issue entitled: From Genome to Function.

1. Introduction

The main aim of the Human Genome Project was to provide a complete and accurate sequence of the 3 billion DNA base pairs that make up the human genome, aiding a better understanding of the biology of man. With the completion of the human genome sequence, many insights have been obtained into the genetic variation among individuals and the genetic architecture of common complex diseases. Complex diseases are those that are caused by a combination of multiple genetic and environmental factors; they include cardiovascular disease (CVD), type 2 diabetes, autoimmune diseases, cancer and Alzheimer's disease. These diseases place an enormous burden on modern societies, particularly with an aging population. Much of our medical care in the future is expected to deal with these common complex diseases. However, the prevention and treatment of these diseases are still largely ineffective and their management does not take sufficient account of personal factors, such as genetic background and environmental conditions. In addition, these common complex disorders are often treated as if they were more simple disorders that are caused by one or a few risk factors. For instance, high plasma cholesterol level is considered as a risk factor for CVD. To reduce CVD risk, the most common drugs being prescribed are statins, which lower plasma cholesterol. Despite the success of statins in reducing CVD risk, a considerable number of problems remain unsolved [1]. A standardized treatment does not work in all patients. Thus, to develop personalized medicine, it is essential to have a better estimation of an individual's susceptibility to diseases based on his/her personal genetic and environmental factors and to decide on the most effective intervention steps for disease prevention and treatment. This concept represents “P4 healthcare”, which stands for a predictive, preventive, personalized and participatory healthcare system [2].

Before the completion of the Human Genome Project, progress in understanding complex disorders was slow and particularly the insight into their causes and mechanisms remained limited. Complex diseases often show a non-Mendelian inheritance pattern due to the interaction of multiple factors. Despite the success in studying Mendelian disorders, family-based linkage analysis showed little power and low resolution in identifying risk genes for complex diseases, often yielding inconsistent or ambiguous linkage signals that cannot be validated [3]. The completion of the human genome sequence has revealed millions of genetic variants in the human genome. This has generated an unprecedented explosion of innovative analysis techniques that can take full advantage of the full sequence data and the corresponding functionality of the genome. As a result, genome-wide association studies (in which the frequency of genetic variants is compared between patients and healthy

☆ This article is part of a Special Issue entitled: From Genome to Function.

* Corresponding author at: Department of Molecular Genetics, University Medical Center Groningen, Antonius Deusinglaan 1, 9713 AV Groningen, The Netherlands. Tel.: +31 50 363 5777; fax: +31 50 363 8971.
E-mail address: m.h.hofker@umcg.nl (M.H. Hofker).

http://dx.doi.org/10.1016/j.bbadis.2014.05.002
0925-4439/© 2014 Elsevier B.V. All rights reserved.

controls) have revolutionized the search for genetic risk variants underlying complex diseases. This review will highlight some of the steps that were needed to reach this point and will describe the success of the large-scale genetic association studies. It will show how the results from these studies are expected to contribute to insights into complex disorders that will help drive P4 healthcare.

2. Historic perspective

What was known before the human genome sequence was completed? In a landmark paper in 1979 [4], Jeffreys described the first DNA sequence polymorphisms and estimated their occurrence throughout the genome at a frequency of approximately 1:100 nucleotides. This finding led rapidly to the realization that this genetic information could be deployed to generate a complete genetic linkage map of the human genome [5]. This led to the mapping of inherited diseases (with the aim of identifying the disease-causing genes), the determination of genetic defects and the development of genetic counseling. Indeed, by the mid-1990s, most Mendelian disorders had been mapped and defined, including the identification of the locus for Huntington's disease, and the cloning of the genes for Duchenne muscular dystrophy and for cystic fibrosis.

Whereas the Mendelian disorders were initially revealed using linkage analysis in affected families followed by positional cloning strategies, the complex diseases were much harder to comprehend due to their non-Mendelian inheritance pattern. For example, the identification of genes defining type 2 diabetes and dyslipidemias initially relied on strategies based on candidate gene sequencing [6]. As the function of only 1% of man's ~22,000 genes was known in the 1990s, it was highly unlikely that the candidate gene strategy would be successful. Nevertheless, in the lipid field, many mutations were discovered in genes with known functions in lipid metabolism by sequencing cohorts of patients and controls (e.g. *LDLR*, *APOE*, *APOB*, *LPL*). Some genes were found by linkage analysis in the families that are affected by rare forms of diabetes, such as in mature onset of diabetes of the young (MODY). However, the mutations in such genes are relatively rare and do not explain the majority of patients with metabolic abnormalities or those suffering from common type 2 diabetes.

Potential disease genes were also widely deployed when using the candidate gene approaches to study a common disease. Several potential functional polymorphisms were frequently tested in genetic association studies using a case–control design. These studies were done on a gene-by-gene basis and were therefore extremely laborious. Despite the limited designs of these early studies, some of the gene polymorphisms showed very robust effects and were replicated successfully in many subsequent studies. These include the APOE polymorphisms that were associated with dyslipidemias [7] and Alzheimer's disease [8], while polymorphisms of PPARgamma were robustly associated with the risk of developing type 2 diabetes [9]. However, in many other studies, the use of relatively small cohorts (often fewer than 1000 cases) and the lack of sufficient knowledge of the human genome and of gene functions resulted in many spurious observations that could not be replicated.

Thus, prior to the complete sequencing of the whole human genome in 2003, only the specific regions of human genome (referred as loci), genes and mutations for most of the Mendelian diseases that segregate in large families had been discovered. Linkage analysis had had some success in linking candidate genes to complex diseases; it was particularly successful when focusing on some extreme and rare cases that resembled Mendelian disorders. But new methods and techniques were needed to understand the genetic basis of common complex diseases.

3. Completing the human genome sequence

The need to determine the whole human genome sequence was foreseen by the end of the 1980s, when the Human Genome Project was boldly proposed to initiate a massive, international, sequencing effort. The initial steps included generating the complete linkage map [10] and cloning the human genome using yeast and bacterial artificial chromosomes [11]. These formed the basis for sequencing the human genome, which was essentially completed after 30 years of effort and using automated machines based on shotgun Sanger sequencing strategies. Two landmark papers in 2001 [12,13] described the initial sequence of the human genome. At that time, the number of genes was estimated to be around 30,000–40,000. This number was subsequently adjusted to ~22,000 genes, which is much less than originally calculated largely because most genes appear in different alternative splice forms, which made a proper estimate very hard. In shotgun Sanger sequencing, the DNA fragments of 200 kb or longer are first cloned into appropriate bacterial and yeast vectors. The clones are then sequenced and produce reads of some 300 bases in length. The main challenge in shotgun sequencing is to assemble these short reads in the correct order and to form a contiguous sequence for each of the chromosomes. Thus, the cloned DNA fragments must have a high overlap and redundancy in order to generate a contiguous sequence with more than 99% accuracy. This greatly limits the sequencing efficiency. From 2005, the development of new, next-generation sequencing (NGS) technologies, which were not based on Sanger sequencing and cloning, greatly facilitated the fast sequencing of DNA [14,15]. Currently, the most popular NGS technologies include the Roche 454 [16] and Illumina sequencing platforms [17]. The principle underlying both technologies lies in sequencing-by-synthesis, using pyrosequencing (Roche 454) or fluorescent labeling of nucleotides (Illumina). Their huge advantage is their high-throughput capacity: they can sequence 30–60 million reads per run, thus increasing throughput many hundred-fold over Sanger sequencing techniques. These NGS technologies are also referred to as deep sequencing.

With the advent of NGS, we have witnessed a dramatic drop in the cost of sequencing and the accompanying exponential growth in the amount of sequence data generated in large numbers of individuals. The data have revealed many genetic differences between individuals, referred as genetic variation. Single nucleotide polymorphisms (SNPs) are one of the most studied types of genetic variation. The initial draft sequence from the HGP identified around 1.4 million SNPs in 2000, while now more than 50 million SNPs have been identified and this number is expected to increase further as more genomes are sequenced. These SNPs show the different frequencies in human population. Some SNPs can be common (with a minor allele frequency MAF $\geq$ 5%), whereas some SNPs have a low frequency (1% $\leq$ MAF < 5%) or are rare (MAF < 1%).

The spectrum of genetic variation in human population is shaped by the age of genetic mutations, natural selection, and random genetic drift. During the DNA replication procedure, some random mistakes can occur. These mistakes in genetics are called mutations and the altered nuclear acids are called alleles. The mutation rate was estimated to range from 1.1 to 3×10^{-8} per base per generation [18,19], and a recent analysis has shown that mutation rates can be higher in males than in females and that this effect increases with paternal age [20]. As these mutations are transmitted to the following generations, they become more and more frequent in a population over time, unless there is a subsequent loss of alleles from the population by natural selection or random genetic drift. This may lead to some alleles showing a lower frequency in human populations than anticipated based on the age of the mutations.

The alleles of different SNPs that near each other on the same chromosome can show non-random combinations. If you observe one specific allele at the first SNP position, you are more likely to observe another specific allele at the second SNP position than anticipated by chance. It is because these SNPs mostly remain linked during the chromosomal recombination at meiosis and travel together between generations. This phenomenon is called linkage disequilibrium and the region with such linked SNPs is called a “haplotype” block. This formed the