Biochimica et Biophysica Acta 1842 (2014) 1889–1895

ELSEVIER

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 rationale of the international HapMap project, proposed in 2002 [21]. In this project, SNPs were identified in large pedigrees to obtain phase information about which SNPs travel together over generations. Typically, the haplotype blocks present relatively short segments of DNA, up to 200–300 kb. The identification of these haplotype blocks can have an important implication in genetic analysis. It would be extremely expensive to directly genotype millions of SNPs. Since some SNPs form haplotype blocks, tagging SNPs (tagSNPs) could be identified to serve as a marker for a particular haplotype. These tagSNPs form the core set of 500,000 markers that is normally used on a DNA chip for detecting genotypes. In addition, imputation algorithms have been developed to predict the genotype of a SNP based on the known genotype of a SNP or SNPs in the same haplotype. In this way, it is possible to predict the genotypes for ~100 million SNPs, based on the genotype information of a limited set of tagSNPs (500,000) and the information on linkage disequilibrium. Imputation strategies have become more and more accurate since whole genome sequencing (WGS) has been performed in large numbers of samples. WGS has permitted the construction of reference genomes and haplotypes, for example, the 1000 Genomes project [22] and the Genome of the Netherlands (Go-NL) project [23]. In particular, WGS can now capture low-frequency variants more accurately. As a result, genotyping on a genome-wide scale has become not only affordable but also highly informative; it has completely changed the field of human genetics over the past ten years.

4. Genome-wide association studies

Analyses using genome-wide association studies (GWAS) represent one of the most important advances in genetics that emerged from the complete human genome sequence and the availability of affordable DNA chips and accurate imputation algorithms. These analyses compare allele frequencies between patients and controls. If a specific allele of one SNP shows a higher frequency in patients than in controls, the SNP is determined to be associated with that disease and the specific allele is the risk allele for that disease. The strategy is based on the common variant–common disease hypothesis, which states that common genetic variation must play a major role in common diseases [24–26]. The first large-scale GWAS were published by the Wellcome Trust Case–Control Consortium in 2007: they performed a chip-based SNP study on 17,000 individuals, testing association between seven diseases and 469,557 SNPs [27]. Now GWAS have been applied to hundreds of different diseases and phenotypes. At present, more than 12,000 disease-associated SNPs have been established; they are spread throughout the human genome (see the Catalog of Published Genome-Wide Association Studies at https://www.genome.gov/26525384). Follow-up replications have shown that most of these associations are robust [28,29]. Nevertheless, the effect size of each SNP is quite moderate, and odd ratios are rarely >1.2. However, when all the SNPs associated with a certain disease are combined, they can explain a substantial proportion of the heritability for that disease. For instance, 40 loci have now been established for celiac disease, explaining up to 53.7% of its heritability [30]. Some SNPs also show association with more than one disease, for example, SNPs associated with several different autoimmune diseases are either shared and/or function in the same molecular pathways, providing more insight into the disease etiology [31,32].

Despite the great success of GWAS in identifying common genetic variants associated with complex diseases, the translation of these findings into clinical applications is still limited. There are three main reasons why this translation to the clinic is proving difficult. First, a large proportion of the heritability for most complex diseases is still unexplained. This implies that the identified SNPs are not the whole answer. Much effort is being made to close the "heritability gap"; it might be explained by unknown epistatic mechanisms, and/or rare variants with strong effects, and/or a large number of common SNPs with very weak effects [33]. GWAS with the current sample sizes (thousands to tens of thousands individuals) are still often underpowered to detect these effects. Second, there is a lack of functional characterization of disease-associated SNPs. For the majority of disease-associated loci, it is still difficult to identify the causal genetic defect. So far, most reported genes that map close to the association signal have no obvious functional connection to the associated diseases. As the associated SNP may tag up to a hundred other SNPs at the same haplotype block (which can harbor several candidate genes), much work needs to be done to identify the true causal gene. Third, not all SNPs fall into groups with the tagSNPs that co-occur on a specific haplotype background. The causal variants might show very low linkage to the tagSNPs. Thus, if these SNPs are not genotyped directly, they cannot be imputed based on the tagSNPs, nor is it possible to test them systematically in GWAS. We therefore need to have an intensive genotyping and novel methods to impute these SNPs using very much larger cohorts.

5. Identification of disease-predisposing variants

To translate the association signals to disease etiology for clinic applications, the first important step is to identify the causal, disease-predisposing variants. The associated SNP serves as a proxy for the causal variant and is itself not necessarily the causal one. One strategy is to fine-map the association signal by intensively genotyping the common and rare SNPs at the associated regions. This approach has been successfully used to analyze SNPs that associate with celiac disease by employing the Immunochip, which contains high-density SNPs at immune-associated loci, including rare variants [30]. The size of the associated loci was greatly reduced and most of them contained only one candidate gene.

The second strategy is to investigate the downstream effect of SNPs. For instance, the SNPs that alter the protein coding are likely to be functional. Much effort went into sequencing the complete exome (the part of the genome formed by exons) in order to identify the causal coding variants. However, this strategy has proved largely unsuccessful [34]. Another possibility being given much attention is that the causal genetic variants are present in the non-coding regions, which do not alter protein coding but play a regulatory role [35]. For instance, they influence promoter activity and alter the expression level of a certain transcript, hence producing disease. The transcript can be either protein-coding (i.e. translated into a protein) [36] or non-coding, like microRNA (miRNAs) or long non-coding RNA (lncRNAs), which can play a regulatory role in biological processes [37,38]. Interestingly, it has been observed that disease-associated SNPs are more likely to have an effect on gene expression than randomly chosen SNPs [39–42]. In most cases, SNPs exert a similar effect in different cell- or tissue types, but some of them may affect a transcript in only one specific cell- or tissue type. It is estimated that approximately 30–60% of SNPs show a tissue-specific effect on transcripts [43–45]. Thus, when a SNP at a disease-associated locus shows an effect on the level of one or more transcripts, especially in disease-relevant cell- or tissue types, this SNP is likely to be (or very close to) a causal variant. This approach is being widely used to pinpoint the causal variant and genes for disease association. For example, it has successfully identified the causal effect of a non-coding variant at the 1p13 cholesterol locus. This non-coding variant is located at the binding site of a transcription factor and it affects the expression level of the *SORT1* gene specifically in the liver [46].

In conclusion, linking specific functions and mechanisms to the disease-associated loci is a very important step. Fine mapping may be required and subsequent steps could directly identify a downstream functional effect. From there, we can either zoom in on the predisposing genetic variant or focus on the SNPs that alter protein coding or show a regulatory effect.

6. Functional analysis of candidate genes

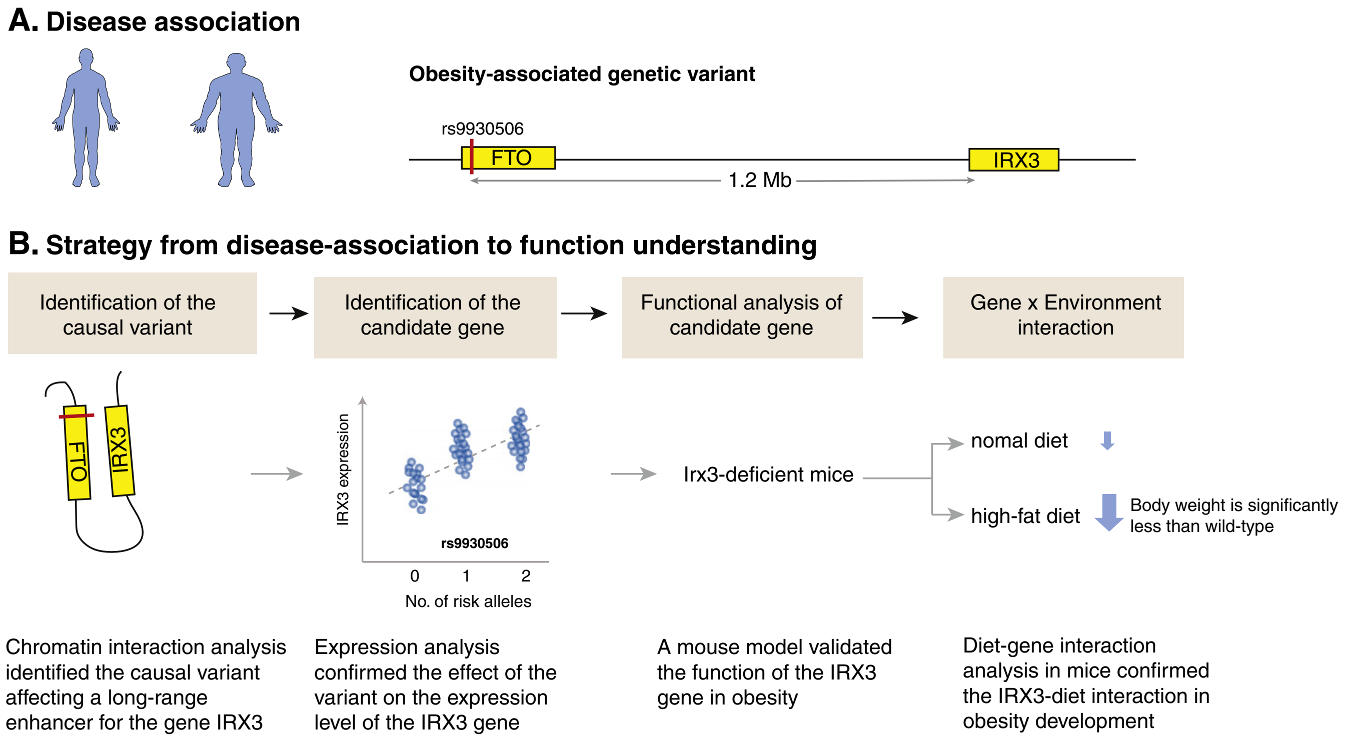
Understanding disease mechanisms requires the identification of causal variants and causal genes as well as functional analysis. Since the causal genes are not necessarily the ones closest to the associated genetic variants, pinpointing the closest gene as the candidate gene can sometimes yield misleading results. However, one successful example of functional analysis is a recent work on the obesity-associated variants at the *FTO* (fat mass and obesity associated) gene [47]. Since 2007, a variant at the first intron of the *FTO* has been known to show the strongest association with obesity in humans [48]. Since then, tremendous efforts have been made to try and understand the role of *FTO* in obesity [49], but no functional connection had been discovered. In 2014, Smemo et al. reported that the *FTO* genetic variant acts as a long-range enhancer using chromatin interaction analysis, affecting the expression of the *IRX3* gene at a distance of 1.2 mb (Fig. 1) [47]. *IRX3* is thus likely to be the true causal gene rather than *FTO*. Further experiments revealed that Irx3-deficient mice showed a 25–30% reduction in body weight, especially when on a high-fat diet. This study nicely demonstrates the multiple steps needed to move from pinpointing a genetic variant to understanding its functionality, including how identifying the causal variant can lead to a better understanding of the gene–environment interaction (Fig. 1).

Once candidate genes have been identified, there are many strategies to study the possible disease mechanisms, for example, cellular models and in vivo models (ranging from yeast and *Caenorhabditis elegans* to mammalian models in rodents and primates) [50–52]. A comprehensive discussion of the possible choices would depend on the disease area and the scientific questions, for instance, there is much interest for studying cardiac development in zebrafish [53,54]. But if a gene with exceptionally strong evolutionary conservation needs to be interrogated and the precise disease phenotype is not important, *C. elegans* or *Drosophila* maybe the model of choice, as studies in these lower animals can be performed efficiently.

The mouse is, however, the most widely used and versatile animal model. Most human diseases can be studied in the mouse, because it is also a mammal and it has an almost identical gene set to man. Furthermore, mouse models have a good track record in replicating human genetic diseases when challenged with appropriate environmental stimuli and/or when genetically modified at a homolog gene. One example is the SNP mapping close to the *TCF7L2* gene, which shows the strongest association with type 2 diabetes. Many studies in man predicted the gene's function to be a beta-cell controlling insulin secretion [55,56], but a gene knockout study provided evidence that, at least in mice, the gene functions in the hepatic glucose metabolism [57]. More work is needed to map the causal mutation, find further support that *TCF7L2* is the causal gene, and clarify its role in man.

The breakthrough for the mouse as a model for human gene function and genetic diseases came with the discovery – in the laboratory of Nobel Prize winner Mario Capecchi – that any gene of interest could be subject to homolog recombination in embryonic stem cells [58]. This finding, published in 1988, paved the way to making loss-of-function mouse models, which have been widely used ever since. Examples of well-established mouse models are the knockouts for the *APOE* gene for Alzheimer's disease and *LDLR* for dyslipidemia and CVD [59]. However, knockouts lead to a genetic deficiency in the entire animal and are often lethal. So it was an important refinement when the Cre–LoxP system was invented, allowing the generation of cell-type-specific loss-of-function models [60]. The Cre recombinase can be expressed in a cell-type-specific fashion, and the timing can be adjusted using an appropriate promoter or by making use of an induction system. The Cre–LoxP method is extremely useful in cases where the complete knockout of the gene cannot be studied due to lethality and it extended the applicability of the knockout strategy. Hence, the molecular genetic mouse toolbox has been instrumental in generating most of our current mechanistic insight. A caveat in mouse models, however, is the difficulty of generating mouse mutants with low gene expression levels, rather than full gene deficiencies. Such models are extremely useful in mimicking the effects of common genetic variation, which is also believed to act by affecting gene expression levels rather than by silencing them completely.

In 2013, the CRISPR/Cas system was invented and it has proved to be even more versatile for generating mouse models [61,62]. It is based on a viral defense system found in bacteria. The CRISPR vector guides the Cas-cleavage enzyme to a specific site in the genome, which is then cleaved and repaired imprecisely. In this way, a series of mutations can be engineered. This new technique is also very efficient, since the vectors can be introduced into fertilized oocytes to generate the


A. Disease association
Obesity-associated genetic variant
rs9930506
FTO
IRX3
1.2 Mb
B. Strategy from disease-association to function understanding
Identification of the causal variant
Identification of the candidate gene
Functional analysis of candidate gene
Gene x Environment interaction
FTO
IRX3
IRX3 expression
rs9930506
0
1
2
No. of risk alleles
Irx3-deficient mice
nomal diet
high-fat diet
Body weight is significantly less than wild-type

Chromatin interaction analysis identified the causal variant affecting a long-range enhancer for the gene IRX3

Expression analysis confirmed the effect of the variant on the expression level of the IRX3 gene

A mouse model validated the function of the IRX3 gene in obesity

Diet-gene interaction analysis in mice confirmed the IRX3-diet interaction in obesity development

Fig. 1. Functional analysis of obesity-associated variants at *FTO* locus.

mouse models directly. Alternatively, the constructs can be introduced into other animals or cell lines to study their effects in vitro, prior to generating an animal model. Thus, we now have mature technologies to modify specific genes and there are now hardly any barriers to studying the mechanisms and genotype–phenotype relationships in model systems of human disease [63].

7. Combining datasets: systems genetics

Functional analysis of candidate genes can confirm which genes are causally involved in a disease and provide insights into their function. However, the gene knockout in model organisms does not really mimic the consequence of the genetic variant in humans. When we identify a genetic variant that alters protein coding or the expression level of a certain transcript, the next natural step is to decipher the further consequences of the altered protein and transcripts. Thus, a systems-based approach is essential to track the flow of biological information from DNA → transcripts → proteins → metabolites → phenotypes [64,65] and analyzes the interaction between the human genome and environmental factors, including gut microbial compositions [66]. This promising strategy is called systems genetics. By definition, this approach reveals the flow of genetic variation from genotype to phenotype, through multi-dimensional biomolecules and their interactions [67].

The rapid advancement of high-throughput technologies that allow the affordable profiling of transcripts, proteins, metabolites and even gut bacterial composition has made this systems approach more and more attractive in the post-GWAS era. It has been observed that the effect of genetic variants at intermediate molecular levels can be much more pronounced than that observed at end-of-point phenotypes (called "phenotype buffering") [68]. This implies that genetic variation can have dramatic effects on gene expression and strong effects on immediate downstream phenotypes, such as lipids [69] or other metabolites [70], while having very weak or no effect on the final disease presentation, like myocardial infarction. For instance, genetic variants at the *FADS1* gene (fatty acid desaturase 1) could explain up to 40% of the observed variation of phospholipids [71], but they have only a small or modest effect on several complex diseases, including inflammatory bowel diseases and metabolic disorders. Therefore, we should have the power to detect such genetic effects on intermediate molecules in reasonably sized samples (e.g. 1000–2000 samples) with genetic analysis of transcripts and metabolites [36,71,72]. Linking genetic loci to the changes in specific molecules can yield novel insights into the underlying disease etiology. The molecules that co-associate with diseases are likely to be mediators between the genotype and phenotypes. Moreover, recent progress of the Encyclopedia of DNA Elements (ENCODE) project has provided multiple levels of information on DNA sequences, including the open chromatin region, transcription factor-binding sites, enhancers and methylated sites [73]. The disease-associated SNPs are enriched for these functional elements, offering mechanistic insights into the regulatory role of genetic variants [74].

The clinical translation of knowledge from GWAS results to disease mechanism would be limited unless the discrete genetic variants and molecules converge into networks that can graphically describe the relationships among the genetic variants and various molecules and phenotypes [75]. Data-driven network construction, combining prior knowledge from metabolic pathways and protein–protein interactions, can result in a new and more complete view of the biological network [76]. Success in network modeling depends heavily on the completeness and quality of datasets, the sample sizes, and the use of proper mathematical models. With only a limited financial budget, experiment designers have two choices at the moment, depending on their research questions. One choice is to profile a dataset as deeply as possible on a limited number of samples, like the ENCODE project or Genome of the Netherlands project. The other option is to focus on a few dimensional data and to profile as many samples as possible, for instance, the Genotype-Tissue Expression (GTEx) project focuses on the gene expression across diverse human tissues [77]. One notable hallmark of network modeling has been its utilization of mathematical models and computational algorithms for integrating and analyzing large datasets. Various algorithms and causal inference modeling techniques have been proposed to establish networks and determine the causal inference of the underlying molecular pathways from genotype to phenotype [78–81]. However, biological systems are complex and no single method is best suited for every situation. Compared to generating the data, the computational analysis now demands a major effort. In the near future, we expect to see the advancement of computational algorithms and mathematical models, pushed by the huge flow of data now being generated.

In this review we have discussed several steps leading from genetic association to causal mechanism that are needed to realize P4 healthcare. These steps include the identification of disease-predisposing variants, functional analysis on candidate genes, and the systems-based approach to illustrate the molecular circuitry from genotype to phenotype. The established causal variants can be incorporated into risk prediction models, resulting in a better prediction of an individual's susceptibility for complex diseases than the use of tagSNPs. This information can aid a targeted, personalized strategy for disease prevention. At the same time, functional studies and systems-based analyses will yield better insights into disease mechanisms and aid the development of new therapeutic targets. Moreover, there is an increasing evidence that reveals the genetic basis of drug efficacy in some patients [82,83]. Some rare and common genetic variants can have a large impact on severe drug reactions, the optimal dosage and efficacy. These observations should stimulate the adoption of a more personalized approach to healthcare and more patient participation. Within the next decade, many more outcomes for clinical application can be expected. Thus, the completion of the human genome sequence and the subsequent technological advances mean that we can now pave the way to a broad uptake of P4 healthcare.

Acknowledgement

We thank Jackie Senior for editing this manuscript.

Funding

This work was supported by the Systems Biology Centre for Metabolism and Ageing (SBC-EMA) and an NWO VENI grant 863.09.007 (J.F.).

References

[1] P. Libby, P.M. Ridker, G.K. Hansson, Progress and challenges in translating the biology of atherosclerosis, Nature 473 (2011) 317–325.
[2] L. Hood, A personal journey of discovery: developing technology and changing biology, Annu. Rev. Anal. Chem. 1 (2008) 1–43.
[3] J. Altmüller, L.J. Palmer, G. Fischer, H. Scherb, M. Wjst, Genomewide scans of complex human diseases: true linkage is hard to find, Am. J. Hum. Genet. 69 (2001) 936–950.
[4] A.J. Jeffreys, DNA sequence variants in the G gamma-, A gamma-, delta- and beta-globin genes of man, Cell 18 (1979) 1–10.
[5] D. Botstein, R.L. White, M. Skolnick, R.W. Davis, Construction of a genetic linkage map in man using restriction fragment length polymorphisms, Am. J. Hum. Genet. 32 (1980) 314–331.
[6] S.S. Rich, J.M. Norris, J.I. Rotter, Genes associated with risk of type 2 diabetes identified by a candidate-wide association scan: as a trickle becomes a flood, Diabetes 57 (2008) 2915–2917.
[7] P. de Knijff, L.M. Havekes, Apolipoprotein E as a risk factor for coronary heart disease: a genetic and molecular biology approach, Curr. Opin. Lipidol. 7 (1996) 59–63.
[8] E.H. Corder, A.M. Saunders, W.J. Strittmatter, D.E. Schmechel, P.C. Gaskell, G.W. Small, et al., Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families, Science 261 (1993) 921–923.
[9] D.M. Altshuler, R.A. Gibbs, L. Peltonen, E. Dermitzakis, S.F. Schaffner, F. Yu, et al., Integrating common and rare genetic variation in diverse human populations, Nature 467 (2010) 52–58.
[10] H. Donis-Keller, P. Green, C. Helms, S. Cartinhour, B. Weiffenbach, K. Stephens, et al., A genetic linkage map of the human genome, Cell 51 (1987) 319–337.

[11] K. Osoegawa, A.G. Mammoser, C. Wu, E. Frengen, C. Zeng, J.J. Catanese, et al., A bacterial artificial chromosome library for sequencing the complete human genome, Genome Res. 11 (2001) 483–496.

[12] E.S. Lander, L.M. Linton, B. Birren, C. Nusbaum, M.C. Zody, J. Baldwin, et al., Initial sequencing and analysis of the human genome, Nature 409 (2001) 860–921.

[13] J.C. Venter, M.D. Adams, E.W. Myers, P.W. Li, R.J. Mural, G.G. Sutton, et al., The sequence of the human genome, Science 291 (2001) 1304–1351.

[14] S. Schuster, Next-generation sequencing transforms today's biology, Nature 5 (2007) 16–18.

[15] H.P.J. Buermans, J.T. den Dunnen, Next generation sequencing technology: Advances and applications, Biochim. Biophys. Acta (2014) 1932–1941 (in this issue).

[16] M. Margulies, M. Egholm, W.E. Altman, S. Attiya, J.S. Bader, L. a Bemben, et al., Genome sequencing in microfabricated high-density picolitre reactors, Nature 437 (2005) 376–380.

[17] D.R. Bentley, S. Balasubramanian, H.P. Swerdlow, G.P. Smith, J. Milton, C.G. Brown, et al., Accurate whole human genome sequencing using reversible terminator chemistry, Nature 456 (2008) 53–59.

[18] M.W. Nachman, S.L. Crowell, Estimate of the mutation rate per nucleotide in humans, Genetics 156 (2000) 297–304.

[19] J.C. Roach, G. Glusman, A.F.A. Smit, C.D. Huff, R. Hubley, P.T. Shannon, et al., Analysis of genetic inheritance in a family quartet by whole-genome sequencing, Science 328 (2010) 636–639.

[20] D.F. Conrad, J.E.M. Keebler, M.A. DePristo, S.J. Lindsay, Y. Zhang, F. Casals, et al., Variation in genome-wide mutation rates within and between human families, Nat. Genet. 43 (2011) 712–714.

[21] R.A. Gibbs, J.W. Belmont, P. Hardenbol, T.D. Willis, F. Yu, H. Yang, et al., The International HapMap Project, Nature 426 (2003) 789–796.

[22] G.R. Abecasis, A. Auton, L.D. Brooks, M.A. DePristo, R.M. Durbin, R.E. Handsaker, et al., An integrated map of genetic variation from 1092 human genomes, Nature 491 (2012) 56–65.

[23] D.I.D.I. Boomsma, C. Wijmenga, E.P. Slagboom, M.A. Swertz, L.C. Karssen, A. Abdellaoui, et al., The Genome of the Netherlands: design, and project goals, Eur. J. Hum. Genet. 22 (2014) 221–227.

[24] D.A. Hinds, L.L. Stuve, G.B. Nilsen, E. Halperin, E. Eskin, D.G. Ballinger, et al., Whole-genome patterns of common DNA variation in three human populations, Science 307 (2005) 1072–1079.

[25] J.K. Pritchard, N.J. Cox, The allelic architecture of human disease genes: common disease–common variant…or not? Hum. Mol. Genet. 11 (2002) 2417–2423.

[26] G. Gibson, Rare and common variants: twenty arguments, Nat. Rev. Genet. 13 (2011) 135–145.

[27] T. Wellcome, T. Case, C. Consortium, Genome-wide association study of 14,000 cases of seven common diseases and 3000 shared controls, Nature 447 (2007) 661–678.

[28] A. Herbert, N.P. Gerry, M.B. McQueen, I.M. Heid, A. Pfeufer, T. Illig, et al., A common genetic variant is associated with adult and childhood obesity, Science 312 (2006) 279–283.

[29] R.J.F. Loos, I. Barroso, S. O'rahilly, N.J. Wareham, Comment on "A common genetic variant is associated with adult and childhood obesity", Science 315 (2007) 187.

[30] G. Trynka, K.A. Hunt, N.A. Bockett, J. Romanos, V. Mistry, A. Szperl, et al., Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease, Nat. Genet. 43 (2011) 1193–1201.

[31] A. Zhernakova, C.C. van Diemen, C. Wijmenga, Detecting shared pathogenesis from the shared genetics of immune-related diseases, Nat. Rev. Genet. 10 (2009) 43–55.

[32] M. Parkes, A. Cortes, D.A. van Heel, M.A. Brown, Genetic insights into common pathways and complex relationships among immune-mediated diseases, Nat. Rev. Genet. 14 (2013) 661–673.

[33] E.E. Eichler, J. Flint, G. Gibson, A. Kong, S.M. Leal, J.H. Moore, et al., Missing heritability and strategies for finding the underlying causes of complex disease, Nat. Rev. Genet. 11 (2010) 446–450.

[34] K.A. Hunt, V. Mistry, N.A. Bockett, T. Ahmad, M. Ban, J.N. Barker, et al., Negligible impact of rare autoimmune-locus coding-region variants on missing heritability, Nature 498 (2013) 232–235.

[35] H.-J. Westra, L. Franke, From genome to function by studying eQTLs, Biochim. Biophys. Acta (2014) 1896–1902 (in this issue).

[36] H.J. Westra, M.J. Peters, T. Esko, H. Yaghootkar, C. Schurmann, J. Kettunen, et al., Systematic identification of trans eQTLs as putative drivers of known disease associations, Nat. Genet. 45 (2013) 1238–1243.

[37] V. Kumar, H.J. Westra, J. Karjalainen, D.V. Zhernakova, T. Esko, B. Hrdlickova, et al., Human disease-associated genetic variation impacts large intergenic non-coding RNA expression, PLoS Genet. 9 (2013) e1003201.

[38] B. Hrdlickova, R.C. de Almeida, Z. Borek, S. Withoff, Genetic variation in the non-coding genome: Involvement ofmicro-RNAs and long non-coding RNAs in disease, Biochim. Biophys. Acta (2014) 1910–1922 (in this issue).

[39] D.L. Nicolae, E. Gamazon, W. Zhang, S. Duan, M.E. Dolan, N.J. Cox, Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS, PLoS Genet. 6 (2010) e1000888.

[40] H. Zhong, J. Beaulaurier, P.Y. Lum, C. Molony, X. Yang, D.J. Macneil, et al., Liver and adipose expression associated SNPs are enriched for association to type 2 diabetes, PLoS Genet. 6 (2010) e1000932.

[41] R.S.N. Fehrmann, R.C. Jansen, J.H. Veldink, H.J. Westra, D. Arends, M.J. Bonder, et al., Trans-eQTLs reveal that independent genetic variants associated with a complex phenotype converge on intermediate genes, with a major role for the HLA, PLoS Genet. 7 (2011) e1002197.

[42] W. Cookson, L. Liang, G. Abecasis, M. Moffatt, M. Lathrop, Mapping complex disease traits with global gene expression, Nat. Rev. Genet. 10 (2009) 184–194.

[43] J. Fu, M.G.M. Wolfs, P. Deelen, H.J. Westra, R.S.N. Fehrmann, G.J. Te Meerman, et al., Unraveling the regulatory mechanisms underlying tissue-dependent genetic variation of gene expression, PLoS Genet. 8 (2012) e1002431.

[44] E. Grundberg, K.S. Small, Å.K. Hedman, A.C. Nica, A. Buil, S. Keildson, et al., Mapping cis- and trans-regulatory effects across multiple tissues in twins, Nat. Genet. 44 (2012) 1084–1089.

[45] P.K. Gregersen, Cell type-specific eQTLs in the human immune system, Nat. Genet. 44 (2012) 478–480.

[46] K. Musunuru, A. Strong, M. Frank-Kamenetsky, N.E. Lee, T. Ahfeldt, K.V. Sachs, et al., From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus, Nature 466 (2010) 714–719.

[47] S. Smemo, J.J. Tena, K.H. Kim, E.R. Gamazon, N.J. Sakabe, C. Gómez-Marín, et al., Obesity-associated variants within FTO form long-range functional connections with IRX3, Nature 507 (2014) 371–375.

[48] C. Dina, D. Meyre, S. Gallina, E. Durand, A. Körner, P. Jacobson, et al., Variation in FTO contributes to childhood obesity and severe adult obesity, Nat. Genet. 39 (2007) 724–726.

[49] M.E. Hess, J.C. Brüning, The fat mass and obesity-associated (*FTO*) gene: obesity and beyond? Biochim. Biophys. Acta (2014) 2039–2047 (in this issue).

[50] T.J. Aitman, C. Boone, G.A. Churchill, M.O. Hengartner, T.F.C. Mackay, D.L. Stemple, The future of model organisms in human disease research, Nat. Rev. Genet. 12 (2011) 575–582.

[51] O. Sin, H. Michels, E.A.A. Nollen, Genetic screens in *Caenorhabditis elegans* models for neurodegenerative diseases, Biochim. Biophys. Acta (2014) 1951–1959 (in this issue).

[52] C.J.L.M. Smeets, D.S. Verbeek, Cerebellar ataxia and functional genomics: Identifying the routes to cerebellar neurodegeneration, Biochim. Biophys. Acta (2014) 2030–2038 (in this issue).

[53] J. Bakkers, Zebrafish as a model to study cardiac development and human cardiac disease, Cardiovasc. Res. 91 (2011) 279–288.

[54] E.E. Davis, S. Frangakis, N. Katsanis, Interpreting human genetic variation with *in vivo* zebrafish assay, Biochim. Biophys. Acta (2014) 1960–1970 (in this issue).

[55] L. Shu, K. Zien, G. Gutjahr, J. Oberholzer, F. Pattou, J. Kerr-Conte, et al., TCF7L2 promotes beta cell regeneration in human and mouse pancreas, Diabetologia 55 (2012) 3296–3307.

[56] O. Le Bacquer, L. Shu, M. Marchand, B. Neve, F. Paroni, J. Kerr Conte, et al., TCF7L2 splice variants have distinct effects on beta-cell turnover and function, Hum. Mol. Genet. 20 (2011) 1906–1915.

[57] S.F. Boj, J.H. van Es, M. Huch, V.S.W. Li, A. José, P. Hatzis, et al., Diabetes risk gene and Wnt effector Tcf7l2/TCF4 controls hepatic response to perinatal and adult metabolic demand, Cell 151 (2012) 1595–1607.

[58] S.L. Mansour, K.R. Thomas, M.R. Capecchi, Disruption of the proto-oncogene int-2 in mouse embryo-derived stem cells: a general strategy for targeting mutations to non-selectable genes, Nature 336 (1988) 348–352.

[59] K. Wouters, R. Shiri-Sverdlov, P.J. van Gorp, M. van Bilsen, M.H. Hofker, Understanding hyperlipidemia and atherosclerosis: lessons from genetically modified APOE and LDLR mice, Clin. Chem. Lab. Med. 43 (2005) 470–479.

[60] A.J. Smith, M.A. De Sousa, B. Kwabi-Addo, A. Heppell-Parton, H. Impey, P. Rabbitts, A site-directed chromosomal translocation induced in embryonic stem cells by Cre–loxP recombination, Nat. Genet. 9 (1995) 376–385.

[61] H. Wang, H. Yang, C.S. Shivalila, M.M. Dawlaty, A.W. Cheng, F. Zhang, et al., One-step generation of mice carrying mutations in multiple genes by CRISPR/Cas-mediated genome engineering, Cell 153 (2013) 910–918.

[62] T. Wijshake, D.J. Baker, B. van de Sluis, Endonucleases: new tools to edit the mouse genome, Biochim. Biophys. Acta (2014) 1942–1950 (in this issue).

[63] T. Wang, J.J. Wei, D.M. Sabatini, E.S. Lander, Genetic screens in human cells using the CRISPR–Cas9 system, Science 343 (2014) 80–84.

[64] J.H. Nadeau, A.M. Dudley, Genetics. Systems genetics, Science 331 (2011) 1015–1016.

[65] M. Civelek, A.J. Lusis, Systems genetics approaches to understand complex traits, Nat. Rev. Genet. 15 (2014) 34–48.

[66] M.C. Cénit, V. Matzaraki, E.F. Tigchelaar, A. Zhernakova, Rapidly expanding knowledge on the role of the gut microbiome in health and disease, Biochim. Biophys. Acta (2014) 1981–1992 (in this issue).

[67] M.R. van der Sijde, A. Ng, J. Fu, Systems genetics: From GWAS to disease pathways, Biochim. Biophys. Acta (2014) 1903–1909 (in this issue).

[68] J. Fu, J.J.B. Keurentjes, H. Bouwmeester, T. America, F.W.A. Verstappen, J.L. Ward, et al., System-wide molecular evidence for phenotypic buffering in *Arabidopsis*, Nat. Genet. 41 (2009) 166–167.

[69] J.A. Kuivenhoven, R.A. Hegele, Mining the genome for lipid genes, Biochim. Biophys. Acta (2014) 1993–2009 (in this issue).

[70] C.M.L. Touw, T.G.J. Derks, B.M. Bakker, A.K. Groen, G.P.A. Smit, D.J. Reijngoud, From genome to phenome—Simple inborn errors of metabolism as complex traits, Biochim. Biophys. Acta (2014) 2021–2029 (in this issue).

[71] K. Suhre, S.Y. Shin, A.K. Petersen, R.P. Mohney, D. Meredith, B. Wägele, et al., Human metabolic individuality in biomedical and pharmaceutical research, Nature 477 (2011) 54–60.

[72] H. Dharuri, A. Demirkan, J.B. van Klinken, D.O. Mook-Kanamori, C.M. van Duijn, P.A.C. 't Hoen, et al., Genetics of the human metabolome, what is next? Biochim. Biophys. Acta (2014) 1923–1931 (in this issue).

[73] B.E. Bernstein, E. Birney, I. Dunham, E.D. Green, C. Gunter, M. Snyder, An integrated encyclopedia of DNA elements in the human genome, Nature 489 (2012) 57–74.

[74] M.A. Schaub, A.P. Boyle, A. Kundaje, S. Batzoglou, M. Snyder, Linking disease associations with regulatory information in the human genome, Genome Res. 22 (2012) 1748–1759.

[75] S.A. Khetarpal, D.J. Rader, Genetics of lipid traits: Genome-wide approaches yield new biology and clues to causality in coronary artery disease, Biochim. Biophys. Acta (2014) 2010–2020 (in this issue).

[76] K. Lage, Protein–protein interactions and genetic diseases: The interactome, Biochim. Biophys. Acta (2014) 1971–1980 (in this issue).
[77] GTEx Consortium, The Genotype-Tissue Expression (GTEx) project, Nat. Genet. 45 (2013) 580–585.
[78] E.E. Schadt, J. Lamb, X. Yang, J. Zhu, S. Edwards, D. Guhathakurta, et al., An integrative genomics approach to infer causal associations between gene expression and disease, Nat. Genet. 37 (2005) 710–717.
[79] D. Marbach, J.C. Costello, R. Küffner, N.M. Vega, R.J. Prill, D.M. Camacho, et al., Wisdom of crowds for robust gene network inference, Nat. Methods 9 (2012) 796–804.
[80] E.C. Neto, A.T. Broman, M.P. Keller, A.D. Attie, B. Zhang, J. Zhu, et al., Modeling causality for pairs of phenotypes in system genetics, Genetics 193 (2013) 1003–1013.
[81] A. Bordbar, M.L. Mo, E.S. Nakayasu, A.C. Schrimpe-Rutledge, Y.M. Kim, T.O. Metz, et al., Model-driven multi-omic data analysis elucidates metabolic immunomodulators of macrophage activation, Mol. Syst. Biol. 8 (2012) 558.
[82] P. Sanseau, P. Agarwal, M.R. Barnes, T. Pastinen, J.B. Richards, L.R. Cardon, et al., Use of genome-wide association studies for drug repositioning, Nat. Biotechnol. 30 (2012) 317–320.
[83] K. Zhou, E.R. Pearson, Insights from genome-wide association studies of drug response, Annu. Rev. Pharmacol. Toxicol. 53 (2013) 299–310.