Part I

Genome-wide association studies

1 Introduction to genome-wide association studies and personalized medicine

Krishnarao Appasani and Raghu K. Appasani

My scientific studies have afforded me great gratification; and I am convinced that it will not be long before the whole world acknowledges the results of my work.

Gregor J. Mendel, Austrian botanist/geneticist (1822–1884)

Ronald A. Fisher, an English statistical geneticist, showed for the first time that a complex quantitative trait can be explained by Mendelian inheritance if multiple genes affect the trait (Fisher, 1918). Thus, one can infer accurate statistical predictions of a complex trait requiring the identification of many small-effect variants, which, in combination, can explain a large fraction of variance in the phenotype. Before 1990, a number of examples of pharmacogenetic traits, usually binary, were published and reviewed (Nebert *et al.*, 2008). Most of them adhere to simple Mendelian inheritance and are controlled by one or a very small number of large-effecter genes. These breakthroughs in genotype–phenotype associations helped to establish expectations of individualized genetic risk prediction. In the pre-genomic era, the genetic dissection of complex diseases was done through classical linkage studies (Lander and Botstein, 1989), and candidate gene-based association studies (Cousin *et al.*, 2003; Patnala *et al.*, 2013). The classical linkage study is a powerful approach to identify rare and high penetrant disease variants or genes, whereas the candidate gene approach was limited to a few genetic markers that are involved in the pathogenesis of complex diseases.

The genome-wide association studies (GWAS) approach was first proposed by Risch and Merikangas in 1996 as a statistical method to detect common variants with modest genetic effects compared to linkage studies. A global collaborative effort called the "HapMap project" was initiated in 2003 to characterize the haplotype patterns in the human genome and subsequently to identify single nucleotide polymorphism (SNPs or snips; The International HapMap

Genome-Wide Association Studies: From Polymorphism to Personalized Medicine, ed. K. Appasani. Published by Cambridge University Press.

Consortium, 2005). HapMap resources guided the design and analysis of genetic association studies that shed light on structural variation and recombination, and identified loci that are involved in natural selection during human evolution. Since the discovery of GWAS and SNPs, association studies have been increasingly employed to reveal the relationships between SNPs and particular disease conditions. The first large set of results from a GWAS were reported in 2005 and 2006 (Klein *et al.*, 2005; Dewan *et al.*, 2006). In 2007, the Wellcome Trust Case Control Consortium published the first large, well-designed GWAS for complex diseases to employ a SNP chip that had good coverage of the genome (Wellcome Trust Case Control Consortium, 2007). The GWAS technique has already supplied useful information about the genetic basis of various diseases including: asthma, cancer, diabetes, heart disease, and mental disorders. The GWAS technique seeks out statistical differences among sets of polymorphic genetic markers between sufferers of a given disease and a control group. The use of GWAS also helped to extend the field of human genetics further using expression quantitative trait loci studies, individualized drug therapy, and personalized medicine strategies.

Scope of this book

This text consists of 23 chapters, grouped into 5 parts, and many of the aforementioned applications are described within various sections of the book, which are summarized as follows.

Part I: Genome-wide association studies

This section consists of four chapters. Phenotypes are composites of the observable traits of organisms and living individuals that originate from the expression of the instructions recorded in the organism's DNA under the influence of environmental factors. Heritability is a global measure that quantifies the overall contribution of genetic factors to a phenotype. Researchers working on such disparate fields such as livestock selection, medical genetics, behavioral economics, or evolutionary biology need to understand the genetic basis of phenotypes. Arcadi Navarro's colleagues describe the study of the genetics of polygenic traits and its importance to include GWAS in Chapter 2. Hongyu Zhao and his colleagues in Chapter 3 provide recent progress in GWAS with respect to statistics, including heritability estimation, association mapping and risk prediction. The limited availability of GWAS performed on populations of non-European ancestry has not precluded a widespread replication of risk variants across populations being observed. Indeed, large inter-population replicability rates allow researchers to reject a scenario of susceptibility variants being continent-specific, and rather favor a scenario of causal variants being common and shared across human populations. These results are shadowed by a poor and unbalanced availability of cross-population data, but the extension of GWAS across ancestries remains one of the best tactics to finally unveil the location of causal variants in susceptibility loci. In Chapter 4, Martinez-Marigorta and colleagues

detail that GWAS replicability varies with time and space across human populations.

Part II: Genome-wide studies in disease biology

This section consists of five chapters. Applications of GWAS in diseases such as myocardial infarction, cancer, and neurodevelopmental disorders are described in this section. Body mass index is the ratio of body weight to height squared, and is the most commonly used index of adiposity and the diagnostic criterion for obesity, a leading risk factor for type 2 diabetes, cardiovascular disease, cancer, and premature death. Approximately half of the factors contributing to body mass index are heritable, and unraveling the specific genetic variation that contributes to this heritability is vital for understanding the biological mechanisms that regulate adiposity. The implementation of the GWAS approach has dramatically increased the speed of gene discovery for body mass index, which is detailed in Chapter 5 by Kilpelainen. Myocardial infarction is a common disease and among the leading causes of death in the world. Pathogenesis of myocardial infarction depends on complex interactions of environmental and genetic factors. Kouichi Ozaki and Toshihiro Tanaka's group adopted GWAS for the first time in human disease biology especially for myocardial infarction and identified a disease susceptibility gene in the Japanese population, and those studies are detailed in Chapter 6. Genome-wide studies have been carried out for decades in familial studies of Mendelian diseases with high penetrance. As genotyping technology continued to advance, genotyping chips became available with increasing numbers of SNPs. The first SNP chip with genome-wide coverage was made available by the mid-2000s, with the first studies using this technology published in 2005 (Klein *et al.*, 2005). Yeager's group applied GWAS studies in the study of cancer, which is described in Chapter 7.

Schizophrenia is a severe mental disorder with a typical onset in adolescence. Symptoms include: delusions, hallucinations, anhedonia, blunted affect, and disorganized speech. In addition, the majority of patients with schizophrenia show some level of cognitive dysfunction. In general, patients with schizophrenia have deficits in most cognitive domains (e.g., attention, memory, and executive functioning). In 2009, the first GWAS study on schizophrenia was published. In Chapter 8, Derks and his colleagues describe the details of such genome-wide association analysis in this psychiatric disease. Advances in methodologies for epigenetic analysis, such as Beadchip microarrays and next-generation sequencing, enable the investigation of the epigenetic status at individual loci, multiple loci, or the whole genome (Appasani, 2012). These new approaches also enable epigenome-wide association studies (EWAS). Several lines of evidence suggest that epigenetic abnormalities can be induced by environmental factors. Thus, clinical epigenetic research not only needs to target congenital disorders, but also needs to investigate acquired chronic diseases including common mental and neurodevelopmental disorders, in which epigenetic abnormalities may reside at multiple genomic loci. In recent years, the numbers of patients suffering from such chronic

diseases have been reported to be increasing in Asian countries (for example, Japan). EWAS studies have mainly been performed for cancer and diabetes mellitus. However, for the first time, Kubota and his colleagues performed and compared EWAS in both brain tissue and blood samples for neurodevelopmental disorders, and the results are described in Chapter 9.

Part III: Single nucleotide polymorphisms, copy number variants, haplotypes and eQTLs

This section consists of five chapters describing the details of the inter-relationship of microRNAs and SNPs, copy number variants (CNVs), and expression quantitative trait loci (eQTLs). MicroRNAs are small, single-stranded RNAs of about 22 nucleotides involved in gene regulation by binding to 3′ untranslated regions of messenger RNAs (Bartel, 2004). Gene silencing by microRNAs is an important mechanism in physiological processes, and its deregulation can lead to complex diseases such as cancer (Garzon *et al.*, 2006). SNPs in the coding sequence of messenger RNAs have been well studied, but not much work has been done in the non-coding regions of the genome, such as the 3′ untranslated regions of messenger RNAs, which harbor many functional sequence elements involved in gene regulation. One type of functional element that can be disrupted by SNPs is the microRNAs target site. SNPs in microRNAs target sites (miRSNPs) can change the affinity between the miRNA seed sequence and its target messenger RNA, resulting in deregulation of gene expression, and possibly in phenotype differences and diseases (Sethupathy and Collins, 2008). In Chapter 10, Thomas and Sætrom summarize the details of the SNPs in microRNAs target sites and map them on the breast cancer susceptibility genes.

Alcohol use disorder is under partial genetic control as a result of common variants in several genes, each with a small effect. Additionally, environmental exposures also play an important role in the development of alcohol use disorder. In recent years, it has been increasingly recognized that genetic and environmental factors are inter-dependent, suggesting that the expression of genetic liability depends on environmental factors. Furthermore, risk factors known as intermediate phenotypes are under the same gene–environment influence and may provide important clues to further understand this heterogeneous disorder. Therefore, Burmeister and colleagues studied the linkage complex associations in alcohol use disorder, and these are described in Chapter 11 of this volume. As we described earlier, GWAS is a method that can link human diseases and traits to specific haplotypes using SNPs. Using GWAS and genotyping approaches, traits and diseases can be associated with large DNA gains or losses. These are referred to as copy number polymorphisms or copy number variations, and can either be inherited as an Mendelian trait or arise as a *de novo* event. In Chapter 12, Brosens *et al.* describe the copy number variations observed in monozygotic twins, their timing and effect, and their impact on human traits and diseases.

Recent findings have suggested that links between the epigenetic status and genetic variants may underline the functionality of SNPs. Most importantly, DNA

methylation has been most frequently linked to several human diseases such as cardiovascular diseases, diabetes, obesity, dyslipidemia, and cancer. CpG-related SNPs (CGSs) constitute a group of SNPs with a particular relationship to DNA methylation. By definition, CGSs refer to those SNPs which can change the formation of CpG dinucleotides that have been established as the primary target site of DNA methylation. CGSs have been found to contribute a significant fraction of allele-specific methylation regions in the human genome (Shoemaker *et al.*, 2010; Gertz *et al.*, 2011), and more than 80% of CGSs were shown to play a regulatory role in DNA methylation (Zhi *et al.*, 2013). In Chapter 13, Ma *et al.* describe their group's first contributions to studying CGSs and their relationship with DNA methylation through genome-wide scale and integrated bioinformatics analysis. These studies help not only in providing a candidate functional mechanism to link SNPs and DNA methylation, but also in its potential contributions to personalized medicine.

Expression quantitative trait loci (eQTL) mapping is a powerful approach to detecting transcriptional regulatory relationships at the scale of the genome. In eQTL studies, gene expression levels measured by high-throughput technologies, such as microarrays and RNA-Seq, are treated as quantitative traits. By simultaneously capturing many regulatory interactions, expression quantitative trait loci offer valuable insights into the genetic architecture of expression regulation (Rockman and Kruglyak, 2006). The ultimate goal of eQTL studies is to elucidate how genetic variations affect phenotypes by using gene expression levels as intermediate molecular phenotypes (Nica and Dermitzakis, 2008). Integrative analysis of variations in transcriptome and next-generation sequencing will result in unprecedented accuracy in eQTL detection and interpretation. These studies will hopefully bring insights into the molecular pathogenesis of complex traits. Thus, in Chapter 14, Chen *et al.* summarize the details of the eQTL analysis and mapping studies.

Part IV: Next-generation sequencing technology and pharmaco-genomics

This section consists of six chapters. Most rare diseases have a genetic basis, and are inherited in a Mendelian fashion. They are usually monogenic disorders segregating in families or are sporadic, being autosomal or sex-linked, and dominant or recessive (McKusick, 1994). Before the advent of the new next-generation sequencing technologies, positional cloning was the most commonly used technique for the analysis of the genetic basis of Mendelian diseases. Linkage and homozygosity mapping have benefitted from the increasing densities of genetic and physical mapping. The new method of exome sequencing helps to screen for known or new functional variants in previously discovered candidate genes and genes functionally related to them. In several cases, exome sequencing has resolved or improved misdiagnoses (Ku *et al.*, 2012). A combination of next-generation sequencing and *in silico* approaches have helped to prioritize and catalog a number of rare variants in both complex traits and disease

phenotypes. In Chapters 15 and 16, Casals and Bosch describe the use of next-generation sequencing as a method to discover and understand the biology of several human rare and complex diseases, respectively.

Breast cancer is the most common type of malignancy among women in many countries around the world. It is well established that multiple genetic and epigenetic factors play an important role in breast cancer. Epigenetic characterization of cancer using DNA methylation profiling of tumors and their corresponding normal profiles has shown that the methylation landscapes are quite disrupted in cancer. Therefore, it is important to identify the mechanistic cross-talk between epigenetic modifications, genome instability, and transcriptional programs within breast cancers. Existing methodologies provide little insight into the mechanisms that drive these epigenetic and genetic changes on a genome-wide scale. In order to investigate these mechanisms Dimitrova and her colleagues and collaborators have performed systematic identification of differentially methylated regions and Alu-enriched loci in relation to major genome rearrangements and breakpoint enriched regions, and they are summarized in Chapter 17. A network analysis approach helps to elucidate networks underlying signaling mechanisms that govern cancer cell survival and proliferation and imply selective pressures for the evolutionary convergence of cancer genomic alterations. To study the genomic alteration in breast cancer, in Chapter 18, Zaman and Wang describe the details of a signaling network analysis approach used in breast cancer and implications for personalized medicine.

Acute myeloid leukemia is a clonal disorder and a very heterogeneous disease with various subtypes classified based on morphology, immunophenotype, and cytogenetics. The nucleoside analog, *cytarabine*, has been the mainstay of acute myeloid leukemia chemotherapy for more than 40 years. However, extensive inter-patient variation in treatment response, development of resistance, and inadequate response to first-line therapy remain the major hurdles to effective chemotherapy. The integration of pharmaco-genetic markers with prognostic markers in larger clinical cohorts would advance our ability to design personalized therapy in patients to achieve the greatest therapeutic benefit. In Chapter 19, Lamba and colleagues summarize the recent advances in pediatric acute myeloid leukemia pharmaco-genomics. Over the past decade, a vast amount of literature has been published about the significant pharmaco-genetic associations on the use of antiretroviral drugs, growing our knowledge base. Genotype–phenotype associations will require the use of advanced technologies for genome-wide screening and statistical analysis in order to draw a definitive comprehensive association. In Chapter 20, Sukasem *et al.* describe the combination of genome technology, computational, and bioinformatics used to study multi-ethnic pharmaco-genomics.

Part V: Population genetics and personalized medicine

This section consists of three chapters. It is a well-established fact that the anatomically modern human originated in Africa about 160,000 years ago and migrated

out-of-Africa. In this process, several populations arose and each one of them has their own evolutionary history. Genetic drift, endogamy practices, admixture, and natural selection are a few examples of evolutionary phenomena leading to genetic diversity among populations around the world, including susceptibility and resistance to genetic diseases, infectious diseases, therapeutic response to drugs, etc. Understanding these phenomena is much more relevant and important in a country like India, which has the richest ethnic, cultural, linguistic, and social diversity in the world. In Chapter 21, Thangaraj and colleagues clearly describe the diverse population genomics and specific variations in India and draw conclusions on the development of "ethnicity-based genomic medicine" strategies.

Personalized medicine is an emerging field that holds promise for major advances in prevention and care at the patient level, and major reductions in the cost of healthcare at the societal level. In personalized medicine, genomics generally plays the role of the ultimate molecular diagnostic. Solving disease at a population scale involves aligning very powerful stakeholders, which often involves obtaining the support of kings, princes, presidents, prime ministers or billionaires. In Chapter 22, Merriman details the "genomic medicine" approach by highlighting the various "ethnicity-based genome sequencing" efforts. According to Merriman, this genomic medicine strategy will allow countries to de-risk and accelerate its execution, removing barriers of competency and capacity, and focusing on integrating the benefits of genomics into their national healthcare systems for the benefit of their people.

Use of the term "personalized medicine" is now ubiquitous (Hamburg and Collins, 2010). However, there is little consistency in how personalized medicine is either defined or used in clinical practice. Various authors have attempted to bring together the different definitions of personalized medicine. Redekop and Mladsi have defined personalized medicine as: "the use of combined knowledge (genetic or otherwise) about a person to predict disease susceptibility, disease prognosis, or treatment response and thereby improve that person's health" (Redekop and Mladsi, 2013). Producing economic evidence within the emerging field of personalized medicine is vital to support the timely and evidence-based introduction of technologies which have a potentially beneficial role in healthcare systems worldwide. Therefore, in Chapter 23, health economists Payne and Eden give an overview of the issues to be considered when evaluating the economic impact of the technologies used to personalize medicine.

Future perspectives

According to McClellan and King, the main advantage of GWAS research results has helped us to discover hundreds of common variants whose allele frequencies are statistically correlated with various illnesses and traits, but these studies did not establish any significant biological or clinical relevance in terms of prognosis and/or treatment (McClellan and King, 2010). We mention throughout this volume that patients' genomes were compared (with those of healthy genomes)

using SNP chips in the ($100 million) "HapMap" project. However, these studies did not provide the molecular etiology of a given disease (Wade, 2011; Visscher *et al.*, 2012). The current generation of GWAS have contributed to identifying novel genes associated with common complex diseases. On the other hand, another approach called "Whole-Genome Sequencing" provides the ability to identify rare alleles with larger effects that are not detected by GWAS. Ideally, a combination of both these approaches will be utilized to provide a more complete view of human genetic variation (Burnham and Hayden, 2012). It is important to mention that although we can obtain GWAS and whole-genome sequencing data from a large number of patients, effect sizes for the majority are of small-effect variants which are simply too miniscule to be detected, even with a practicably attainable sample size. Future GWAS will have to explore structural variations, gene–gene interactions, epigenetic and gene–environment interactions. In a nutshell, a holistic approach is needed to contribute to the future of "personalized medicine."

In conclusion, we agree with the comments of Nebert and Zhang that the idealistic goal of personalized medicine and individualized drug therapy, which needs a holistic understanding of each individual patient's unique -omics readout, is most likely unattainable for the vast majority of complex traits (Nebert and Zhang, 2012). In the "human genome era," we have achieved several breakthroughs in our understanding of the biology, genetics and pathogenesis of human diseases. In the coming "human circuitry era" (Nybo, 2014; understanding the gene–protein networks of various tissues/organs, and all the neural circuits in brain) using "genomic medicine" approaches, we hope to unravel more mysteries of disease biology which will open up new research avenues and lay a strong foundation for the development of the new field of "personalized medicine."

Due to space constraints very few relevant papers were cited here. We apologize for not citing all other contributions.

REFERENCES

Appasani, K. (2012). Epigenomics: From chromatin biology to therapeutics. Cambridge University Press, Cambridge.

Bartel, D. P. (2004). MicroRNAs: genomics, biogenesis, mechanism and function. *Cell*, **116**, 281–297.

Burnham, L.R. and Hayden, M.R. (2012). Whole-genome sequencing: the new standard of care? *Science*, **336**, 1112–1113.

Cousin, E., Genin, M., Mace, S., *et al.* (2003). Association studies in candidate genes: strategies to select SNPs to be tested. *Hum. Hered.*, **56**, 151–159.

Dewan, A., Liu, M., Hartman, S., *et al.* (2006). HTRA1 promoter polymorphism in wet age-related macular degeneration. *Science*, **314**, 989–992.

Fisher, R.A. (1918). The correlation between relatives on the supposition of Mendelian inheritance. *Trans. R. Soc. Edinb.*, **52**, 399–433.

Garzon, R., Fabbri, M., Simmino, A., *et al.* (2006). MicroRNA expression and function in cancer. *Trends Mol. Med.*, **12**, 580–587.

Gertz, J., Varley, K.E., Reddy, T.E., *et al.* (2011). Analysis of DNA methylation in a three-generation family reveals widespread genetic influence on epigenetic regulation. *Publ. Libr. Sci. Genet.*, **7**, e1002228.