Tatsuhiko Tsunoda · Toshihiro Tanaka
Yusuke Nakamura *Editors*

Genome-Wide Association Studies

Springer

Genome-Wide Association Studies

Tatsuhiko Tsunoda • Toshihiro Tanaka
Yusuke Nakamura
Editors

Genome-Wide Association Studies

Springer

Editors
Tatsuhiko Tsunoda
Graduate School of Science
The University of Tokyo
Tokyo, Japan

Toshihiro Tanaka
BioResource Research Center
Tokyo Medical and Dental University
Tokyo, Japan

Yusuke Nakamura
Cancer Precision Medicine Center
Japanese Foundation for Cancer Research
Tokyo, Japan

ISBN 978-981-13-8176-8 ISBN 978-981-13-8177-5 (eBook)
https://doi.org/10.1007/978-981-13-8177-5

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Preface: History of Genome-Wide Association Study

As we understand from variations in individual's height, weight, character, or looks, human is diverse in every aspect. From the clinical point of view, it is one form of the expression of human diversity that every patient diagnosed with the same disease does not respond equally to the same therapy. Naturally, these diversities come from complex combination of different genetic and environmental factors. Genetic epidemiology focuses on revealing genetic backgrounds of clinical status including the disease itself, drug responses, or adverse effects of drugs. As a scientific backbone, common disease-common variant hypothesis (CDCV hypothesis) claims that genetic risk of common diseases would be due to variations in the genome with relatively high allele frequencies. The simplest and thorough way to investigate this hypothesis is to examine all the DNA variations, especially SNPs (single nucleotide polymorphisms) in the genome, and this approach is called genome-wide association study (GWAS). Nowadays, it has become one of the most powerful tools to understand genetic aspects of human/disease diversity. However, its potency was not proven true until 2002, when our team in RIKEN first in the world reported in *Nature Genetics* (Ozaki et al. 2002) the identification of functional variants through GWAS that are genetically associated with myocardial infarction, one of the common cardiovascular diseases. The success completely changed the mainstream of study for identifying disease genes/loci, from linkage analyses to GWAS, from rare to common diseases.

Several important mechanisms were indispensable for this first-in-the-world achievement that were established through the Japanese Millennium Genome Projects started in 2000. These projects were financially supported as a Japanese National Project by the Japanese Government led by the late ex-prime minister, Keizo Obuchi. One of the mechanisms is gene-based SNP discovery project, which aimed to identify 150,000 SNPs located within gene regions, because public SNP databases contained considerable "noises" that could not be found in the real world. The reason why we focused on gene regions was very simple; considering the complexity and uncovered significance of the variations in the genome which are outside gene regions, it should be much easier to interpret the links between genetic variations and phenotypes. As a first step toward personalized medicine that utilizes

genetic information, this SNP discovery project was successful by the identification of 190,000 genetic variations in the human genome (Haga et al. 2002).

Another concurrent project was raised by the question how we could examine large number of SNPs in a practical time that would be identified through the SNP discovery project. To solve this problem, high-throughput genotyping system was developed using combination of multiplex PCR and Invader assay (Ohnishi et al. 2001). With this system, 100 SNP loci for each of the 384 individuals could be examined simultaneously. Even today, the number of individuals that can be examined in one experiment seems to be among the largest, that is, this system is appropriate for replication study that should examine a limited number of loci for a large number of individuals.

In parallel with these achievements originated in Japan, an international collaborative effort named International HapMap Project started in 2001 to develop genetic information database served as an infrastructure for GWAS. This project was based on two previous findings. First, the genetic backgrounds seem to be different among ethnicities; therefore, allele frequency of some loci might be different, which might cause stratification of the sample population. This raised the need of knowing genetic information on ordinary individuals from each ethnicity to serve as control subjects. Second is linkage disequilibrium (LD). In the genome, crossover of chromosomes during meiosis does not occur at any point randomly but rather accumulate at specific loci, called "recombination hotspot." Therefore, recombination rate is not flat throughout the genome but has sharp spike-like form at the loci. Generally, the region surrounded by the two spikes is in LD (LD block). Within LD block, SNP loci are sometimes completely linked with another SNP in the same block. This phenomenon enabled the researchers to perform GWAS much more efficiently because they do not need to examine each of the two loci that are in absolute LD; just one out of two is enough. The achievements were published in *Nature* in 2005, where RIKEN made largest contribution among nine genotyping institutes in the world. We genotyped 269 DNA samples from four populations for 1,000,000 SNP loci throughout the genome and found 250,000 to 500,000 SNP loci (tag SNP) are enough to study whole genome variations, depending on the populations (The International HapMap Consortium 2005). The rapid progress in genotyping technology using DNA microarray also has accelerated GWAS, and now, more than 10,000 study results have been published, and the data can be browsed through web site.

The subject of this book is the discussion of the history, future, and beyond of the genome-wide association study (GWAS), which enabled exploration of unknown disease etiology in the whole human genome. In particular, it aims to show the current utility and limitation of GWAS and how to breakthrough that limitation. This book presents (1) analytic methodologies of GWAS, (2) results for disease and pharmacogenomic analysis, (3) GWAS in the era of next-generation sequencing (NGS), and big data. For typical common diseases, we focus on cardiovascular, autoimmune, diabetic, cancer, and infectious diseases. Important feature of this book is that it gives directions as to (1) which types of diseases/phenotypes are suited for GWAS, (2) future of GWAS, and (3) what is beyond GWAS. The readers

would expect to understand how a road map resulting from GWAS can lead to the realization of personalized/precision medicine: functional analysis, drug seeds, pathway analysis, disease mechanism, risk prediction, and diagnosis.

Tokyo, Japan Toshihiro Tanaka

References

Haga H, Yamada R, Ohnishi Y, Nakamura Y, Tanaka T (2002) Gene-based SNP discovery as part of the Japanese Millennium Genome Project: identification of 190,562 genetic variations in the human genome. J Hum Genet 47:605–610

Ohnishi Y, Tanaka T, Ozaki K, Yamada R, Suzuki H, Nakamura Y (2001) A high-throughput SNP typing system for genome-wide association studies. J Hum Genet 46:471–477

Ozaki K, Ohnishi Y, Iida A, Sekine A, Yamada R, Tsunoda T, Sato H, Sato H, Hori M, Nakamura Y, Tanaka T (2002) Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction. Nat Genet 32:650–654

The International HapMap Consortium (2005) A haplotype map of the human genome. Nature 437:1299–1320

Contents

Chapter 1
Genotyping and Statistical Analysis

Artem Lysenko, Keith A. Boroevich, and Tatsuhiko Tsunoda

Abstract Development of technologies for high-throughput profiling of DNA variation has led to rapid discovery of causal genetic mutations underlying complex phenotypic traits and diseases. These exciting advances were originally enabled by the results from the Human Genome project (1990–2003) that allowed the completion of the first genome-wide association study in 2002 and led to the development of haplotype maps of the human genome. Technological advances in microarray genotyping and next-generation sequencing have since made possible the widespread and cost-effective application of this approach and, in combination, have powered the new age of biomedical discovery. This chapter introduces the history and fundamental principles of genetic association analysis, and explains key concepts and current statistical methods for processing these data. In particular, discussed topics include experimental design of association studies, quality control procedures, approaches for dealing with the population stratification, statistical testing for genetic associations and more recent developments in detection of effects of rare variants and genetic interactions.

Keywords Genome-wide association study · High-throughput genotyping technologies · Genetic association testing · Genotype imputation · Haplotype mapping

A. Lysenko · K. A. Boroevich
Laboratory for Medical Science Mathematics, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan

T. Tsunoda (✉)
Tsunoda Laboratory (Medical Science Mathematics), Department of Biological Sciences, Graduate School of Science, The University of Tokyo, Tokyo, Japan

Department of Medical Science Mathematics, Medical Research Institute, Tokyo Medical and Dental University, Tokyo, Japan

Laboratory for Medical Science Mathematics, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan
e-mail: tsunoda@bs.s.u-tokyo.ac.jp

T. Tsunoda et al. (eds.), *Genome-Wide Association Studies*,
https://doi.org/10.1007/978-981-13-8177-5_1

1.1 Principles of Genetic Association Analysis

Genetic association is the co-occurrence of inherited genomic characteristics that exist with a frequency higher than would be expected by chance. The study of genetic associations therefore aims to identify such associations, most commonly for the purpose of establishing a link to an observable phenotype, like a disease, which can provide hints about the underlying genetic mechanism giving rise to the trait in question. Importantly, an association may also exist between particular genetic polymorphisms themselves, either due to physical proximity of their sites (genetic linkage) [1] or due to differences in frequencies of particular alleles (linkage disequilibrium). Similarly, according to this definition, genetic associations can be said to exist between particular phenotypes, even if exact genetic determinants are not known.

Physical proximity on the chromosome is an important factor underlying key principles of genetic association analysis. During prophase I of meiosis in eukaryotes, recombination occurs between different pairs of homologous chromosomes, which gives rise to new combinations of paternal alleles in the offspring. As recombination points are essentially random, greater distance between different alleles will increase the probability that they will be separated and *vice versa*. As some alleles can produce easily observable phenotypes, in combination with cross-over frequency information, they could be used to construct genetic linkage maps even in cases where the exact genomic locations of these causal alleles are still unknown [2]. Patterns of genetic linkage may make actual identification of exact causal variants more challenging, as the causal variant is usually embedded in a linkage disequilibrium block of its genomic region.

1.2 Common Disease-Common Variant Hypothesis, Linkage Disequilibrium, and SNPs

Determination of the first complete human genome sequence by the International Human Genome Project (1990–2003) has rapidly accelerated medical research and became a major turning point for its future direction. In particular, it has led to advances in genetic linkage analysis, where genotype markers in the genomes of patients' families were used to successfully identify the causes of several monogenic diseases. However, eventually it became clear that the relevant genes underlying many common diseases could not be as easily identified using this method. The reason for this is that common diseases are often multifactorial diseases, meaning that many factors with moderate penetrance are involved. However, given that a presence of a particular factor by itself only confers a moderate increase in risk, it also follows that these factors would have to be relatively common in affected populations (if the disease in question is also common). This interpretation of these early observations gave rise to the Common Disease-Common Variant (CDCV)

hypothesis [3]. CDCV proposes that the causes of common diseases are often high-frequency polymorphisms within the population that originated as DNA mutations of common ancestors and have been inherited by their many descendants. In case of these common multifactorial diseases, it was theoretically shown that an association study using unrelated individuals, which examines allele frequency differences between cases (disease carriers) and healthy controls, has higher detection power than the classical pedigree-based linkage analysis. Also, owing to linkage disequilibrium (LD) between polymorphisms based on the inheritance of common genomic fragments (haplotypes) from a small set of ancestor individuals, it would be possible to detect true genetic causes of the disease by looking at surrounding polymorphisms as their proxies. In this respect, among different types of genetic polymorphisms single nucleotide polymorphisms (SNPs) are considered to be particularly promising as very large number of them occur in the human genome with high population frequency. For the same reason, the number of SNPs that needed to be directly genotyped for association analysis was found to be relatively small because the experiments can be made more efficient by carefully selecting representative (tag) SNPs to cover the majority of all haplotype regions. This largely removes the need to profile a large number of redundant SNPs in linkage disequilibrium with each chosen tag SNP. In combination, these observations suggested the theoretical possibility of a genome-wide association study (GWAS) – a type of analysis that looks for genetic association using tag SNPs covering the entirety of the human genome. However, due to the technology available at the time, GWAS analysis only rose to prominence several years later.

1.3 First GWAS in the World and the Dawn of the High-Throughput Genomics Age

In order to identify genes related to common diseases using GWAS, it was first necessary to isolate SNPs in proximity to all genes in the human genome. The first project aiming to collect these necessary data was done by the Institute of Medical Science and the University of Tokyo with the support of the Japan Science and Technology Agency (2000–2002) [4]. During this work, the regions flanking exons and promoters for each gene were sequenced in genomic DNA of 24 Japanese individuals. The analysis identified 174,269 polymorphisms that were subsequently released for public use in the Japanese Single Nucleotide Polymorphisms (JSNP) database (http://snp.ims.u-tokyo.ac.jp). Using this information, a group of Japanese researchers from RIKEN Institute successfully developed the first pioneering GWAS. These early efforts also lead to the introduction of several notable technological advances, among them was a robotic system that enabled highly accurate SNP genotyping assay (Invader method) [4], which was instrumental in greatly facilitating necessary data collection. From a biomedical perspective, the most important outcome was the discovery of myocardial infarction-related genes in

2002 [5]. Additionally, the large-scale analysis of about 80,000 genotyped SNPs in 564 individuals led to the development of the first map of LD/haplotype blocks of all human chromosomes, which in turn allowed to greatly improve the efficiency of subsequent genotyping efforts by identifying a suitable representative set of (tag) SNPs that captured sufficient information about haplotypes of over 13,000 genes [6]. The establishment of this powerful approach paved the way for the rapid advances in discovery of disease-related genes for multiple human diseases.

1.4 The Rise of Commercial SNP Genotyping Assays

With the success of the first GWAS, many commercial platforms for high-throughput genotyping began to appear. Most of these protocols use a combination of DNA hybridization and DNA ligase, nuclease or polymerase, followed by a technique to visualize the alleles present, such as fluorescence [7].

Primer Extension Methods

A common method among early SNP genotyping techniques was primer extension. One, developed by SEQUENOM, is the homogenous MassEXTEND (hME) assay [8]. Sample DNA is hybridized with oligonucleotide primers based on the sequence adjacent to the SNP of interest.[1] These primers are then extended, using DNA polymerase with a mixture of terminator nucleotides, by a single base, into the polymorphism. This single base extension (SBE) results in two allele-specific extension products with different mass. The difference in mass is then quantified using matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) in a mass spectrometer. Later developments allowed for the multiplexing of reactions through the careful design of the expected products [8].

The AcycloPrime-FP assay developed by PerkinElmer uses template-directed dye-terminator incorporation with fluorescence-polarization (FP-TDI) [9, 10]. Similar to MassEXTEND, primers specific to the sequence adjacent to the SNP are hybridized with genomic DNA. The primers are then extended in the presence of dye-terminators specific for the SNP alleles, resulting in amplicons of different mass for each allele. However, rather than using mass spectrometry to quantify the products, FP is used. When excited by plane-polarized light, a fluorescent molecule will emit polarized light [11]. Under constant conditions, the degree of FP is proportional to the molecular volume, and therefore the weight, of the fluorescent molecule.

[1] https://www.ahc.umn.edu/img/assets/19726/Multiplexing_hME_App_Note.pdf.

Hybridization Methods

One of the first genotyping assays was the TaqMan assay, first devised by researchers at Cetus Corporation [12], and later developed by Roche Life Science and Applied Biosystems. This assay is a hybridization method built on the Taq polymerase, a thermostable DNA polymerase with 5′ exonuclease activity from the thermophilic bacteria *Thermus aquaticus*. TaqMan hybridization probes are designed for both alleles of the SNP of interest and hybridized with genomic DNA. The probes have a reporter fluorophore specific to each allele on the 5′-end and a quencher molecule on the 3′-end. When intact, the quencher molecule is in close enough proximity to quench the fluorescence emitted by the reporter. PCR using Taq polymerase is performed using primers flanking the SNP of interest. As the polymerase extends the sequence, if it encounters a perfectly hybridized probe, one that matches the genomic sequence, it cleaves both dyes off the probe, and the fluorescence or the reporter dye is observed. If the probe does not perfectly match, hybridization is greatly reduced, neither the reporter nor quencher is released, and no fluorescence is observed.

Another hybridization approach is that of the Invader assay [13], developed by Third Wave Technologies and mentioned in the previous section. The Invader assay is an isothermal probe-based method that utilizes the action of a flap endonuclease (FEN) named cleavase. FENs are a class of endonucleases that catalyze structure-specific cleavage [14]. In the most basic assay, an allele specific primary probe, containing a reporter fluorophore and quencher molecule, and an allele independent Invader probe are hybridized to target genomic DNA. If the probe is complementary to the target, an overlapping invader structure is formed, the 5′ end of the probe is cleaved off releasing the reporter, and fluorescence is observed. This assay was further developed to include two invasive cleavage reactions and a distinct fluorescent signal for each of the SNP alleles [13].

Multiplexing Methods

Multiplexing involves minimizing the number of times an assay has to be performed while maximizing the number of independent SNPs genotyped. Today's technologies can genotype close to one million SNPs in a single DNA microarray. Microarrays consist of a two-dimensional array of synthesized oligonucleotides bound to a substrate [15].

The Affymetrix GeneChip DNA microarray technology uses a photolithographic process to synthesize oligonucleotides directly onto a treated quartz wafer [16]. Nucleotides are added with protected terminal hydroxyl groups. Between each round of oligonucleotide extension (coupling), a UV mask is used to allow light through only at sites where the current nucleotide (A, T, C, or G) is to be added (deprotection). Through repeating rounds of deprotection and coupling, 25-mer

probe sequences are synthesized. For SNP genotyping, multiple probes for both alleles are generated. The location of the SNP within the probe varies ±4 nucleotides from the center. Sample genomic DNA is amplified and labeled and hybridized to the microarray, after which the fluorescence intensity of each site is measured, and genotypes are determined through relative intensity of all the SNPs' probes.

In comparison, Illumina's BeadArray technology[2] uses 3-μm silica beads covered with hundreds of thousands of copies of a unique address oligonucleotide. The beads are randomly placed into wells on a substrate. In the GoldenGate Assay,[3] three primers per SNP are required: two allele specific sequences, each with a different forward-facing universal primer sequence, and one locus specific sequence, attached to a bead address oligonucleotide, and a reverse facing universal primer. In the allele specific extension and ligation step, only allele specific primers present in the target genomic DNA are extended and ligated to the address sequence. The next step of PCR amplification uses Cy3 and Cy5 fluorophore-labeled universal primers to tag each allele. The amplicons are then hybridized to the beads on the array and genotypes can be determined by the fluorescence observed at locations of each address tagged bead.

ParAllele Bioscience developed the Molecular Inversion Probe (MIP) [17] assay as a multiplex genotyping solution. MIPs, also known as padlock probes, are linear oligonucleotides containing two sequences complimentary to the target sequence at the 5′ and 3′ ends, separated by a linker sequence [18]. When hybridized perfectly to the target sequence, the MIP's ends can be joined using DNA ligase, forming a circularized molecule. For genotyping, locus specific MIPs are designed with the complementary sequence flanking the SNP and a unique 20 base tag. After annealing to the target genomic DNA, the gaps are filled in four separate polymerization and ligation reactions, one for each of the four possible nucleotides. Each reaction is then amplified and a fluorescent label is added. The four nucleotide specific reactions are then hybridized to a separate array and visualized.

1.5 The International HapMap Project

In October 2002, the International HapMap project began, an international initiative to comprehensively examine the polymorphisms and LD patterns throughout the human genome [19, 20]. This project was a collaborative effort by the researchers from Japan, UK, Canada, China, Nigeria and the United States. The initiative aimed to profile the genetic diversity across several key human sub-populations by genotyping the DNA of 90 African (30 families), 90 Caucasian (30 families), 45 Chinese and 45 Japanese individuals. To achieve best possible efficiency of human labor use, lower genotyping expenses and effectively target common diseases, it was decided that in Phase I, the genotyping will be done for at least one SNP with high allele

[2] https://www.illumina.com/science/technology/beadarray-technology.html.

[3] http://barleyworld.org/sites/barleyworld.org/files/illuminasnpgenotyping.pdf.

frequency (minor allele frequencies $\geqq 0.05$) in each 5-kb window. More than 1 million SNPs were genotyped across ten centers, using five different genotyping platforms. Among all contributors, Japanese researchers from RIKEN were responsible for 24.3% (seven chromosomes) of all data, which was the largest contribution from any single institution. In Phase II, Perlegen Sciences, Inc. received funds from US National Institutes of Health (NIH) and aimed to genotype an even larger number of SNPs, no longer limited by the distance and allele frequency. This led to the successful genotyping of about 4.4 million SNPs using an oligonucleotide array technology. In addition, data for about 11,000 non-synonymous SNPs and 4500 SNPs located in the major histocompatibility complex (MHC) region, and data collected using Affymetrix GeneChip Mapping Array 500K set, Illumina HumanHap 100 chip, and Illumina HumanHap 300 chip platforms, were combined to create the most comprehensive dataset available at the time. Subsequently, sets of SNPs under linkage disequilibrium coefficients of $r^2 > 0.8$ were consolidated and merged in order to create a combined reference set of 500,000 tag SNPs. Based on this new data, Illumina, Inc. was able to develop probes to target these tag SNPs on microarray chips, which greatly improved speed and reduced costs of high-throughput genotyping. During the same period, Affymetrix commercialized a genotyping array equipped with SNP probes picked at random positions in the human genome (though on SNP Array 6.0, tag SNPs from the international HapMap project were also included). With these commercial chips and arrays, GWAS developed rapidly all over the world, and in 2007, the approach came into a wide-spread use [21].

1.6 Next Generation Sequencing and the 1000 Genomes Project

After the International HapMap project, which predominantly focused on high-frequency polymorphisms (minor allele frequencies >5%) in several key populations, the focus gradually shifted to other types of potentially highly relevant variants. In particular, polymorphisms with moderate frequencies and penetrance, and variants with low frequencies and high penetrance were also believed to be promising for discovery of novel associations between specific genes and diseases. High-throughput genotyping of these SNPs became increasingly more tractable with the rise of a completely new family of sequencing methods, called “next generation sequencing”. Rapid development of these technologies coincided with the early efforts to establish the GWAS methodology. Once next generation sequencing platforms had sufficiently matured, it became possible to sequence the whole genome of each individual at a very rapid speed. Taking advantage of these technological advancements, in January 2008, the 1000 Genomes project was launched with the aim to take forward the efforts of HapMap project. One of the goals was to identify and comprehensively profile 1–5% minor allele frequency SNPs throughout the whole genome using the next generation sequencing [22]. This was a joint

research effort by the NHGRI in the United States, Wellcome Trust Sanger Institute in the United Kingdom, and BGI in China. As a pilot study, the following whole genomes were sequenced: 1 Caucasian and 1 African family (total of 6 people) with high coverage (depth = 42×) as well as 179 individuals from 4 populations with low coverage (3.6×). In addition, 8,140 exonic regions of 697 individuals from 7 populations were sequenced with high coverage (56×). From these results, 15 million SNPs, 1 million short insertions/deletions and 20,000 structural variations were identified and genotyped [22]. At the time, it was estimated that 95% of the human genomic variations were detected. Additional plans were then made to sequence a further 2500 genomes with low coverage (4×). New tag SNP sets that take into account this information about SNPs with minor allele frequencies as low as 2.5% were used to further extend the Illumina chip designs, starting from the Illumina HumanOmni1 microarray onwards. Although the number of probes on those new chips was 2.5 million to begin with, it was expected that it will increase to 5 million due to the new information to be generated by the 1000 Genomes project. In addition to advances in whole-genome coverage chips, many specialized chips were also made possible with this great wealth of information. Some examples include pathway specific chips, such as the metabochip [23] and immunochip [24] are enriched for SNPs that have been associated with metabolic diseases and immunogenetics studies, respectively. Exome chips, such as the Illumina HumanExome BeadChip, not only contains common exonic SNPs, but also known rare non-synonymous variants [25].

Following in the footsteps of the 1000 Genomes project, multiple population-specific projects have appeared in recent years. The UK10K Project performed the low coverage whole genome and high coverage exome sequencing of almost 10,000 normal and diseased individuals from the British population [26]. Over 24 million novel variants were discovered. Similarly, the Tohoku Medical Megabank Project sequenced the whole genomes of over 1000 Japanese individuals, identifying over 4 million autosomal SNPs with a MAF greater than 5% [27]. The project has since increased the number of individuals to over 3.5 thousand and 7 million SNPs with a MAF greater than 1%. The efforts of these and similar projects will continue to strengthen our knowledge of human variation.

1.7 Experimental Design of Genome-Wide Association Studies

While these efforts have been greatly increasing our knowledge about the structure and variation of the human genome, GWAS methods for associating these variations with diseases and other phenotypic traits likewise became increasingly more refined and standardized. As with many other types of biomedical experiments underpinned by statistics, appropriate experimental design plays a particularly important role in ensuring success of such studies. This is because hypothesized relative risks

attributable to particular factors, the number of samples, and the number of markers to be examined (in relation to multiple testing) directly affect statistical power (i.e. the probability to correctly reject the null hypothesis) – and therefore chances of successfully discovering robust novel associations between individual variants and traits. The estimation of statistical power and determination of the sample size necessary to detect a significant association effect is most commonly done by testing for the difference in the relevant population proportions [28].

Most commonly, an independent statistical test is performed to check for an association between each individual genetic locus and a phenotype of interest. Therefore, the number of such tests can easily number in the millions and adequately accounting for multiple testing is particularly important for controlling false positive findings. While more detailed overview of this topic will be introduced elsewhere in this chapter, from the experimental design perspective it is important to highlight the central the role that replication of results plays in ensuring robustness of all GWAS findings. For these reasons it is now usually expected that a GWAS will have at least two sets of samples, a "discovery" subset and a possibly smaller "replication" subset. Ideally, a replication subset would have been generated from a distinctive cohort of patients, with two sets of samples potentially being collected and processed at different sites. These sets of samples are then independently analyzed and any genome-wide significant results are compared, with the idea being that only variants that were well-replicated across these two datasets represent true associations. However, it is worth noting that interpretation of these replication results may not always be straightforward, as different top variants in the same LD regions may often be found across the two sets.

1.8 Fine-Mapping of Trait Associated Variants

As outlined above, even when a highly significant trait-associated variant is identified by GWAS analysis, it may not necessarily mean that the variant is mechanistically causal of that trait. A possibility always exists that it may be one of many other variants located in the same LD block is the true cause. For this reason, it is usually necessary to conduct additional *post-hoc* analysis to identify actual mechanisms from the raw GWAS association results. One possible strategy is to perform targeted resequencing around the identified markers in order to comprehensively map out the surrounding variations in LD patterns in relevant case/control samples. Due to effects of random variation and complexity of the haplotype structure, the true causal variant can be very far from the strongest association signal, therefore there is no definitive strategy to determine how large this surrounding region needs to be. Despite these potential complications, starting the search in an LD region of the strongest association signal is still a reasonable first step. Among all of these regions, it is common to first consider the variants with the most significant associations and then explore the wider haplotypes (consisting of multiple variants) more significantly associated with the disease than any single variants. It may then be possible

to further narrow down these candidates to more likely ones by performing meta-analyses – i.e. to combine the association statistics of equivalent GWAS performed in other ethnic groups in order to further increase power of the association tests. However, meta-analysis can be complicated as different studies often use different microarray platforms, which may profile very different sets of tag SNPs. It follows that if SNPs are not found in all of the platforms they cannot be meta-analyzed. Likewise, due to imperfections of the modern microarray technology, some of the SNPs may not be called at a 100% rate, leading to presence of missing values. When multiple studies are combined, these eventualities tend to increase by chance and for this reason whole-genome imputation is often essential for allowing GWAS meta-analysis to be done across these diverse microarray platforms.

Tagged haplotype blocks are often larger than a genomic region for a given gene and its non-coding regulatory elements (i.e. transcription factor binding, enhancer and promoter regions). Therefore, a situation can frequently arise were more than one underlying causal polymorphism may be present in a block. These types of independent signals can be recovered using conditional analysis, where an association test is repeated while including the SNP with a strongest signal as a co-variate. In this setup, SNPs that do not contribute an independent signal will tend to have their significance lowered, whereas any remaining highly significant SNPs may indicate an existence of an independent causal polymorphism in the tagged area. Despite being relatively simple, this approach can frequently yield new and more precisely localized association signals [29].

1.9 Identifying Single Nucleotide Variants in Next Generation Sequencing Data

Genetic variants found in a particular genome can be directly profiled by next generation sequencing technologies. The completion of the Human Genome Project has generated a first reference genome, which then allowed unambiguous locations to be defined for newly discovered variants by using this complete sequence as a reference. When a new sequence is determined using NGS technologies, it is then compared to the reference genome to identify potential differences in a process called “variant calling” [30]. Understanding function of these genotyped variants can facilitate disease diagnosis, suggest their driving mechanisms and improve our understanding about how complex phenotypes arise. Although high-throughput sequencing was made increasingly easy and cost-effective by recent technological advances, this process is still error-prone. For this reason, when interpreting NGS data, one important factor is sequencing depth, which refers to the number of times a particular fragment of the genome has been sequenced. As different errors are likely to occur each time a particular fragment is sequenced, it follows that if the process is repeated enough times it will be possible to derive a true sequence by consensus. However, sequencing errors are not always entirely random and

are frequently determined by the particular biases of technologies used, which also means that these patterns can be modelled statistically to correct those errors.

If the sequencing depth is insufficient, difficulties can arise in distinguishing errors from true observed variants. In principle, sufficient sequencing depths could largely eliminate the need for more sophisticated statistical analysis to distinguish variants from errors. In cases where sequencing depth and quality are adequate, even simple heuristic approaches like majority call can be sufficient to produce an accurate result. In practice, sequencing is still relatively expensive and the resources spent on increasing sequencing depth in most cases can ultimately be better spent on sequencing additional samples. Therefore, statistical methods for variant calling often have a pivotal role in modern NGS analysis pipelines, where their use can deliver better value through more efficient use of the laboratory-based sequencing resources.

Most current methods for variant calling use Bayesian statistical approaches, which bring certain advantages of being able to incorporate various kinds of additional information to improve results. The simplest of such methods are single-site calling approaches, where each site for which alternative bases are called is considered independent of all others [31]. The base is called by combining a genotype likelihood (e.g. how often a particular call is observed in a sample) and some form of informative prior derived from a suitable reference panel. However, this approach cannot easily resolve the situation where different calls are made in different samples, which can frequently arise in cases where sequencing depth is insufficient. Therefore, more sophisticated methods combine the information from multiple samples [32], e.g. genotype frequency information and can optionally assume Hardy-Weinberg equilibrium as part of the estimation process. Due to the increased amount of information such methods need to reconcile, the problem is most commonly solved using expectation maximization (EM) algorithms. As newer technologies can produce longer reads containing multiple polymorphisms, more refined methods can now use information across these multiple sites in order to further improve call quality [33, 34]. Given the complex structure of multi-site likelihood and respective priors, the problem is most commonly solved using an MCMC approximation approach.

1.10 Quality Control Procedures in Genome-Wide Association Studies

Reliability of samples and markers can have a profound effect on downstream statistical analysis of genome-wide association studies [35]. Batch effects, population stratification, and sample relatedness are all major factors that need to be identified and potentially corrected to ensure the results are not biased and true associations are discovered.

Although a wide variety of quality control issues can arise depending on the nature of the study, technologies used, and data collection protocols, most widely applicable factors include population stratification, call rate profiling, sex consistency and sample relatedness. One of the most frequently employed ways to check for sample handling errors is to look at the gender recorded in annotation versus one that could be derived from the genomic data. This can be done by considering the X-chromosome heterozygosity rate. An additional advantage of this check is that it may also reveal some common types of chromosome abnormality syndromes, which can adversely affect downstream analysis. Sample relatedness can be checked using kinship coefficients and by looking at the distribution of alleles identical by decent (IBD). An IBD statistic refers to a piece of DNA inherited from a common ancestor where no recombination events have occurred. Long IBD regions can indicate the relatedness of samples and complete identity can be an indicator of a duplicate sample, whereas frequencies of IBD alleles can be used to deduce the degree of relatedness, which can then be checked against a pedigree, if available. Population substructure refers to systematic differences between much larger groups of individuals and are commonly associated with wider ethnically or geographically-linked groups. Global similarity of genomic sequences of such individuals can easily lead to spurious associations due to allele frequencies inherent in these different populations. Due to complexities involved in profiling and correcting these patterns, this topic is covered in detail in its own section further in this chapter. Lastly, it is important to look at inconsistencies of allele sets across independent markers, especially where variant calls could not be made due to insufficient confidence and exclude these cases from downstream analysis. These checks are usually done both with respect to individual samples and loci. If it is found that these types of errors are particularly prominent in specific sample(s), it can be an indication of poor quality of DNA material. For this type of quality control commonly used filtering thresholds are usually set at 98–99% call rate.

Batch effect analysis can be done by looking at the differences in quality between samples, like call rate differences, minor allele frequencies (MAF) and genomic inflation. Batch effects commonly originate from influence of practical aspects of laboratory analysis or data collection considerations, where it is often more efficient to process a set of samples at the same time and slight differences between these sets therefore unavoidably arise. One way to control for batch effects is at a stage of experimental design, e.g. by ensuring random allocation of samples to batches. If information about experimental processing of batches is retained, differences in quality can be identified by comparing call rates and minor allele frequencies between them, where strong differences can indicate incorrect calls made in one of the subsets. A more comprehensive diagnostic can also be done using an association test with batch label treated as a dependent variable. If detected, batch effects can be adjusted for using standard multivariate modelling techniques.

Lastly, the problem can be approached from the perspective of individual marker analysis. Here, quality control methods include evaluations of marker-specific call rates, comparison with established reference datasets to identify deviations and, if available, using duplicate samples to assess overall quality by concordance. Low

call rate can be a property of a particular marker as well as a particular sample. Recommended practice is to evaluate and filter low call rate markers prior to performing an analogous type of analysis for particular samples. Likewise, some experimental designs could incorporate control samples, i.e. duplicate samples which can be effectively leveraged to verify overall reproducibility of experimental profiling. To validate accuracy of experimental profiling, one common strategy is to conduct the genotyping of reference cell lines for which the true sequence is known with great confidence, like those originating from the HapMap or 1000 genomes projects. If accurate pedigree information is available, genotypes can be checked for Mendelian errors, defined as instances where alleles are found that could not be received from either of the parents. SNPs which have very low minor allele frequency (MAF), usually below 1 or even 5%, are also commonly excluded from analysis. The reason is that low occurrence of one allele can lead to violation of underlying distribution assumptions for most of the conventional parametric significance tests and, likewise, very low MAF also commonly arises as an artifact due to genotyping errors. Another common quality control test is to statistically examine evidence of selective pressure affecting the marker. This is usually done by computing the deviation from Hardy-Weinberg equilibrium using Pearson's Chi-squared test. However, it is worth pointing out that such departure from equilibrium can be both evidence of true association signal as well as evidence of genotyping error, so Hardy-Weinberg equilibrium is most commonly taken into consideration when interpreting the results and may not always be used at the quality control stage. In the cases where binary traits are investigated, like presence or absence of a given disease, one option could be to do the Hardy-Weinberg equilibrium-based filtering on the control samples only, where no deviation due to trait of interest would be expected.

1.11 Genotype Inference Methods

Despite recent advances, whole-genome sequencing technologies remain relatively expensive compare to genotyping arrays. One potential way to ameliorate the cost is to use arrays for profiling in combination with whole-genome imputation methods [36]. Imputation is a collective name for a family of statistical inference approaches that aim to predict untyped genotypes based on observed ones using prior knowledge about haplotype structure and frequencies in a given reference panel. Imputation is especially useful for identification of causal SNPs in a given genomic location, as a causal variant will likely be in linkage disequilibrium with the ones found significant by the association tests, even in cases where it is not directly observed.

Haplotype phasing can improve both accuracy and performance of imputation methods. Given that in most cases the human genome is profiled in a diploid state, modern profiling techniques will usually not be able to directly determine the haplotype to which a particular variant belongs. However, if the overall distribution of

different haplotypes in a population is known, this information can be reconstructed using statistical modelling approaches in a process called “haplotype phasing”. Possible strategies include simple multinomial brute force expectation-maximization algorithms, for example one of the early successful tools PHASE employed this strategy [37].

Once haplotypes in a sample are determined, this information can be further leveraged in combination with a suitable reference panel in order to estimate the unobserved makers. Accuracy of these whole genome imputation methods can be greatly improved by using a larger reference panel. If size of the reference is large enough, even very rare SNPs can be imputed very accurately using the current generation of these methods [22]. Therefore, efforts to expand available reference panels are currently underway for several major human population groups. HapMap was the first project to produce a reference set for multiple human populations, however this dataset is now largely superseded by the reference panel from 1000 Genomes project, which now covers several other major ethnic groups. As was already discussed earlier in this chapter, this information is also used to improve the quality of modern genotyping arrays by ensuring that the most informative markers for all haplotypes in a population of interest are included on the array.

1.12 Population Stratification and Its Implications

Population stratification is defined as systematic differences in allelic frequencies arising due to differences in ancestry of sub-populations considered in a given study. Subpopulations with low intra-mating frequencies can be subject to differential genetic drift, where frequencies of alleles not under selective pressure can diverge by chance, given sufficient time. These ancestry differences can therefore confound the true genetic determinants underlying the phenotype of interest [38]. For this reason, it is important to control for population stratification in order to identify true association. One obvious way to control for stratification would be to ensure complete population homogeneity during the experimental design stage, e.g. through use of ethnicity or family ancestry information during recruitment into the study. Though this is still one of the most important ways for controlling stratification, this information is usually subject to considerable inaccuracies and is often found to be insufficient to fully reflect full complexity of possible population structure. Alternatively, a family-based design can be used, where data is collected from individuals known to be related and therefore guaranteed to be unaffected by issues of population stratification.

Detection and quantification of population stratification is possible using the genomic control method proposed by Devlin and Roeder [39]. Their approach uses a Cochran-Armitage trend test to compute the inflation factor, which can then be used to adjust relevant association test statistics. However, one disadvantage is that possible differences between individual alleles are not taken into account, as adjustment is applied in a uniform way. To allow for greater flexibility, structured

association tests (e.g. [40]) were proposed that seek to identify sub-groups or clusters of individuals and therefore allows for greater flexibility, but these are computationally costly to apply and depend on additional parameters, like the number of clusters.

To address these limitations, yet another alternative method was developed, which uses principal components analysis to capture the population structure [41]. Principal component analysis identifies major axes of variation within the data, and was shown to accurately reflect self-reported ethnicity or even geographic distance between the samples. The amount of variation attributed to particular axes can then be directly used to adjust for effects of population stratification by incorporating them as co-variates in a regression model used for association test at a level of individual samples. Owing to its great computational efficiency and flexibility, principal component-based stratification analysis is now the most commonly used method to control for population stratification. Typically, a subset of reference markers is used to perform the analysis and identify any highly divergent outlier samples, which are then excluded. If the remaining main dataset is still determined to be subject to substantial stratification, top principal components are added to the model as a simple and efficient way to adjust for those effects.

1.13 Statistical Testing for Genetic Association

Once adequate preparatory and quality control steps have been completed, the next step is to investigate the genetic association that can explain the observed phenotype of interest. Most commonly, a trait linked to particular locus can be binary (a "case-control" design), like affliction with a particular disease or quantitative, like height or cholesterol levels. Of particular note among quantitative associations is the one linking genetic variation to expression patterns of particular gene(s), called expression Quantitative Trait Locus (eQTL). Based on the design of a particular study, recruited individuals can be from particular families or considered unrelated. For brevity, this section will only deal with the by far most common study design where recruited individuals are not related – a "population-based" study design. This section will describe most typical strategies for identifying associations of common variants whereas some of the alternative techniques for rare variants will be covered separately in the last section of this chapter.

In by far the most typical scenario, particular alleles do not necessarily lead to certain manifestation of a binary trait, but rather alter the probability or risk of such an occurrence. The probability of an individual in a population to display a trait is formally called "penetrance". Given that in a diploid human genome two possible copies of each allele are normally present, the correct statistical model of the relationship between genotype and phenotype will depend on the type of genetic dominance in effect at a given locus. Likewise, number of alleles can have additive or multiplicative effect – and this is equally applicable both for magnitude of quantitative traits and penetrance of binary traits. In case of binary traits, the strength of

association can be quantified as an "odds ratio", which is a ratio of odds for a trait of interest given particular alternative genotypes.

Given computationally intensive nature of the analysis at a whole-genome scale, each locus is usually tested independently of all others. In a simpler scenario, exact genotypes will be called and therefore (in case of a binary test) data can be represented as a contingency table where counts in each cell would be numbers of individuals with a particular genotype-trait combination category. The type of the model determines how the table is constructed, e.g. a two-by-two table in case of a dominant or recessive model or a two-by-three one if no particular model is assumed. As usually the correct model is not known, it is common to assume an additive model, which can be represented with a two-by-three contingency table that is also considered to have an ordered relationship to the trait. If there is an assumption of trend or ordering, this relationship can be captured using Cochran-Armitage trend test, otherwise a Chi-squared test can be used if independence between all categories is judged more appropriate. However, in practice it is often highly desirable to incorporate additional covariates into the model, which these types of simple tests cannot accommodate. For example, probability of developing particular diseases often increases with age or may be affected by individual's gender. This information can only be incorporated by using more sophisticated multivariate models and, in the case of GWAS analysis, logistic or linear regression models are most commonly used. Most typically, logistic regression models for binary traits include covariates for age, gender and, when correcting for population stratification, the first few principal component values for each sample.

Given the complexity of GWAS experiments and very large number of factors that can potentially lead to bias, it is vitally important to check and identify the presence of these potential problems. One commonly used generic way to verify the results is using the quantile-quantile (QQ) plots of the final association significance values. Given that the number of true signals in a GWAS is usually expected to be small, the patterns of unrelated SNPs are expected to be effectively random, i.e. an expectation to observe a particularly high significance value by chance is only influenced by the number of samples in a dataset. A QQ-plot is a scatter plot of expected versus observed significance values that can be used to verify this pattern. If all sources of bias have been accounted for, most of the points would fall on a 45-degree line, with a handful of highly significant points above this line if a true association signal is present. This analysis is often also summarized as a genomic inflation factor (λ) statistic. Genomic inflation factor is formally defined as a ratio of an actual over expected Chi-squared distribution medians, with λ close to 1 meaning no inflation. Though most typically this analysis is used to check for the presence of population substructure, other types of artifacts like block effects, may also be detected.

Given that the number of loci that can be profiled using whole-genome sequencing or array technologies supported by genome imputation can be in the millions, it is particularly important to correct the significance values for the number of tests performed. However, standard procedures to correct for family-wise error rate, like Bonferroni correction assume independence between individual tests. Due to linkage disequilibrium patterns, this assumption does not hold true in the case of GWAS

and therefore such methods are likely to be too conservative [42]. Previous estimates determined that an appropriate number of assumed independent signals is roughly in the region of 1,000,000. Insights from this work where used to derive a widely accepted GWAS significance cut-off of 5×10^{-8}, though it must be noted that this estimate is most applicable for the European population and the true correct value would depend on the diversity of the population being studied. Another alternative is to use a permutation test to compute the adjusted significance values. By permuting the response labels and calculating significance, an empirical distribution of probabilities can be computed. For this reason, the permutation test is considered to be the best method of correction, however this approach is very computationally intensive which can make it infeasible to apply in practice – though efficiency can be improved by using approximate methods [43].

To ensure the voracity of reported associations, the last step in the analysis usually involves replication of the result in an independent dataset. Replication is particularly important in the context of GWAS as it has been found that genomic patterns underlying polygenic phenotypes tend to be highly complex and it is common to identify large numbers of loci which individually explain only very small amount of total heritability. The size of observed effect can mean that studies are often underpowered to robustly confirm the true effect. Likewise, replication can help to identify and discount spurious associations arising due to bias and can also serve to confirm the existence of the effect under different sets of conditions and derive a more accurate estimate of a true effect size.

1.14 Recent Methodological Advances in Genotype Association Analysis

Genetic heritability refers to the rate at which a particular phenotype is inherited by an offspring from its parent. By comparing known heritability (e.g. how often siblings inherit a disease from their parents) with what can be predicted by existing models based on genomic data it is possible to determine how much of the variation in a phenotype is accounted for by currently identified genetic polymorphisms. Conventional GWAS analysis considers individual effects of genetic polymorphisms on the trait of interest, however it has now become evident that entirety of such variations still explains only small part of all known heritability. This phenomenon is referred to as the problem of “missing heritability” [44]. Several explanations for this problem have been proposed, including possible methodological limitations of estimating true heritability, accurately measuring or defining phenotypes and possible epigenetic effects. Other possible explanations attribute missing heritability to genomic effects which are not adequately captured by the classical GWAS analysis methods, like interactions, influence of rare polymorphisms or highly polygenic effects. If a phenotype is determined by additive effects of a very large number of polymorphisms with very small individual effects, simply

increasing the number of samples will eventually sufficiently increase the statistical power to detect all of these small associations, though this strategy will inevitably be subject to considerable diminishing returns. On the contrary, other possibilities imply that the missing heritability problem may eventually be solved by further improvements in methodology and several novel approaches have already been put forward to explore these avenues.

Considerable advances have been made in detecting the effects of rare SNPs. Detection of rare polymorphism by conventional GWAS statistical tests leads to inflated risk of false positive detections, due to highly unbalanced frequencies of alleles which violate distribution assumptions of commonly used significance tests. In most current analysis pipelines this risk is mitigated by not considering any polymorphisms where minor allele frequency is below particular threshold, most commonly below 1% or 5% of all samples profiled in the study. To capture these effects, rare SNPs can be pooled and considered as a group, where a test would then consider the overall effect of a set of polymorphisms [45], usually in the context of some form of a burden model. Consequently, such tests require additional inputs about how to group different SNPs into meaningful sets, with some common strategies being to group SNPs around particular genes or even pathways.

Interaction between polymorphisms occurs when an effect of one allele is conditionally dependent on the effect of another, a phenomenon also referred to as epistasis. Detection of interactions is challenging due to their combinatorial nature, which means that very large number of individual tests would be required to exhaustively check all possibilities [46]. As well as being computationally infeasible, this also leads to loss of statistical power due to multiple testing. Therefore, epistasis detection methods commonly involve development of strategies to reduce the number of tests performed by using some form of prior knowledge, e.g. for example by looking at interactions between polymorphisms found to be individually significant.

Ultimately it is most likely that some combination of these possible explanations underlies the problem of missing heritability and some evidence has been found to suggest influence of all of these factors in particular cases. It is also likely that different factors are prominent for different types of phenotypes. Given this diversity of possible hypotheses and the absence of a definitive solution, at present the question about the causes of missing heritability and best strategies to address it still remain the subject of active debate.

References

1. Morgan TH (1911) Random segregation versus coupling in Mendelian inheritance. Science 34:384
2. Sturtevant AH (1913) The linear arrangement of six sex-linked factors in Drosophila, as shown by their mode of association. J Exp Zool A Ecol Genet Physiol 14:43
3. Lyamichev V et al (1999) Polymorphism identification and quantitative detection of genomic DNA by invasive cleavage of oligonucleotide probes. Nat Biotechnol 17:292

4. Haga H, Yamada R, Ohnishi Y, Nakamura Y, Tanaka T (2002) Gene-based SNP discovery as part of the Japanese Millennium Genome Project: identification of 190 562 genetic variations in the human genome. J Hum Genet 47:605
5. Ozaki K et al (2002) Functional SNPs in the lymphotoxin-α gene that are associated with susceptibility to myocardial infarction. Nat Genet 32:650
6. Tsunoda T et al (2004) Variation of gene-based SNPs and linkage disequilibrium patterns in the human genome. Hum Mol Genet 13:1623
7. Dearlove AM (2002) High throughput genotyping technologies. Brief Funct Genomic Proteomic 1:139
8. Gabriel S, Ziaugra L (2004) SNP genotyping using Sequenom MassARRAY 7K platform. In: Current protocols in human genetics. Chapter 2, Unit 2 12
9. Chen X, Levine L, Kwok PY (1999) Fluorescence polarization in homogeneous nucleic acid analysis. Genome Res 9:492
10. Hsu TM, Chen X, Duan S, Miller RD, Kwok PY (2001) Universal SNP genotyping assay with fluorescence polarization detection. BioTechniques 31:560
11. Kwok PY (2002) SNP genotyping with fluorescence polarization detection. Hum Mutat 19:315
12. Holland PM, Abramson RD, Watson R, Gelfand DH (1991) Detection of specific polymerase chain reaction product by utilizing the 5′–3′ exonuclease activity of Thermus aquaticus DNA polymerase. Proc Natl Acad Sci U S A 88:7276
13. Olivier M (2005) The invader assay for SNP genotyping. Mutat Res 573:103
14. Mast A, de Arruda M (2006) Invader assay for single-nucleotide polymorphism genotyping and gene copy number evaluation. Methods Mol Biol 335:173
15. Bumgarner R (2013) Overview of DNA microarrays: types, applications, and their future. In: Current protocols in molecular biology. Chapter 22, Unit 22 1
16. Dalma-Weiszhausz DD, Warrington J, Tanimoto EY, Miyada CG (2006) The affymetrix GeneChip platform: an overview. Methods Enzymol 410:3
17. Hardenbol P et al (2003) Multiplexed genotyping with sequence-tagged molecular inversion probes. Nat Biotechnol 21:673
18. Nilsson M et al (1994) Padlock probes: circularizing oligonucleotides for localized DNA detection. Science 265:2085
19. International HapMap Consortium (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449:851
20. International HapMap Consortium (2005) A haplotype map of the human genome. Nature 437:1299
21. Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447:661
22. 1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. Nature 467:1061
23. Voight BF et al (2012) The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. PLoS Genet 8:e1002793
24. Cortes A, Brown MA (2011) Promise and pitfalls of the Immunochip. Arthritis Res Ther 13:101
25. Igo RP Jr, Cooke Bailey JN, Romm J, Haines JL, Wiggs JL (2016) Quality control for the illumina HumanExome BeadChip. Curr Protocols Hum Genet 90:2.14.1
26. Walter K et al (2015) The UK10K project identifies rare variants in health and disease. Nature 526:82
27. Yamaguchi-Kabata Y et al (2015) iJGVD: an integrative Japanese genome variation database based on whole-genome sequencing. Hum Genome Var 2:15050
28. Hong EP, Park JW (2012) Sample size and statistical power calculation in genetic association studies. Genomics Inform 10:117
29. Yang J et al (2012) Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. Nat Genet 44:369

30. Nielsen R, Paul JS, Albrechtsen A, Song YS (2011) Genotype and SNP calling from next-generation sequencing data. Nat Rev Genet 12:443
31. Li H, Ruan J, Durbin R (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res 18:1851
32. Martin ER et al (2010) SeqEM: an adaptive genotype-calling approach for next-generation sequencing studies. Bioinformatics 26:2803
33. Scheet P, Stephens M (2006) A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am J Hum Genet 78:629
34. Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR (2010) MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet Epidemiol 34:816
35. Yang X, Chockalingam SP, Aluru S (2012) A survey of error-correction methods for next-generation sequencing. Brief Bioinform 14:56
36. Browning SR, Browning BL (2007) Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet 81:1084
37. Stephens M, Smith NJ, Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. Am J Hum Genet 68:978
38. Cardon LR, Palmer LJ (2003) Population stratification and spurious allelic association. Lancet 361:598
39. Devlin B, Roeder K (1999) Genomic control for association studies. Biometrics 55:997
40. Yang BZ, Zhao H, Kranzler HR, Gelernter J (2005) Practical population group assignment with selected informative markers: characteristics and properties of Bayesian clustering via STRUCTURE. Genet Epidemiol 28:302
41. Price AL et al (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38:904
42. Sham PC, Purcell SM (2014) Statistical power and significance testing in large-scale genetic studies. Nat Rev Genet 15:335
43. Gao X, Becker LC, Becker DM, Starmer JD, Province MA (2010) Avoiding the high Bonferroni penalty in genome-wide association studies. Genet Epidemiol 34:100
44. Manolio TA et al (2009) Finding the missing heritability of complex diseases. Nature 461:747
45. Lee S, Wu MC, Lin X (2012) Optimal tests for rare variant effects in sequencing association studies. Biostatistics 13:762
46. Cordell HJ (2002) Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. Hum Mol Genet 11:2463

Chapter 2
Genetics of Coronary Disease

Kouichi Ozaki and Toshihiro Tanaka

Abstract Coronary artery disease (CAD), including its severe form, myocardial infarction (MI), is a common serious disorder, and a leading cause of death in industrial countries. The pathogenesis depends on multiple interactions on an environmental and genetic basis. As genetic heritability of CAD comprises ~ 50% of the pathogenesis, elucidating the detailed genetic architecture of CAD would facilitate development of a future precision medicine. Initially, we started a genome-wide association study (GWAS) for MI with about 100,000 single nucleotide polymorphisms (SNP) in Japanese from early 2000, and identified the SNPs in lymphotoxin-α gene (*LTA*) associated with the increased risk of MI. As far as we know, this study is the first GWAS for common disease worldwide. This hypothesis-free GWAS ultimately led to identification of a possible MI pathological condition by mediating an inflammatory cascade including IKK signalosome and BRAP, encoded by the gene that was robustly associated with an increased risk of MI in Asian population. On the other hand, recent mega-GWASs for more than 200 traits have collectively revealed many genetic risk factors for common diseases. To date, GWASs from around the world have shown 98 genetic risk factors for CAD.

Keywords Coronary artery diseases · Myocardial infarction · Genetic heritability · Genome-wide association study · Susceptibility loci · IKK signalosome and BRAP · Inflammation

K. Ozaki (✉)
Laboratory for Cardiovascular Diseases, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan

Division for Genomic medicine, Medical Genome Center, National Center for Geriatrics and Gerontology, Obu, Japan
e-mail: ozakikk@riken.jp; ozakikk@ncgg.go.jp

T. Tanaka
Department of Human Genetics and Disease Diversity, Tokyo Medical and Dental University Graduate School of Medical and Dental Sciences, Tokyo, Japan

Bioresource Research Center, Tokyo Medical and Dental University, Tokyo, Japan

T. Tsunoda et al. (eds.), *Genome-Wide Association Studies*,
https://doi.org/10.1007/978-981-13-8177-5_2

2.1 Introduction

Coronary artery disease (CAD) and its severe form myocardial infarction (MI) are leading causes of death worldwide. CAD results from complicated interactions of multiple genetic and environmental factors. Life quality for CAD patients has been greatly improved by current pharmaceutical and diagnostic approaches, whereas the high morbidity still exists. In particular, MI often occurs without any preceding clinical signs and is followed by severe complications, especially ventricular fibrillation and cardiac rupture, which may result in sudden death. MI is a disease of the vessel that feeds the cardiac muscle, called the coronary artery. Irreversible damage to the cardiac muscle is incurred by abrupt occlusion of the coronary artery. The detailed CAD/MI pathogenesis is largely unknown; however, studies on epidemiology show that the risk factors for CAD include dyslipidemia, type 2 diabetes mellitus, obesity, hypertension, smoking, inflammation, and inheritance [4, 5]. The contribution of heredity to CAD seems to be relatively large because of ~50% genetic heritability [32, 37]. A hypothesis for common diseases – common genetic variants and much of the contribution of genetic variants to the increased/higher risk of common/sporadic disorders were proposed and considered respectively [6, 18, 36]. In 2000, we launched genome-wide association studies (GWASs) for CAD in Japanese, with nearly 100,000 SNPs selected in genes based [11] on a high-throughput multiplex PCR invader assay system [24], and found genetic loci associated with risk for CAD, including *LTA* [2, 10, 14, 17, 26, 31]. These observations showed the power of the GWAS, hypothesis-free, to find the clue for important novel pathogenesis of disease to further figure out the pathway of the disorder and to explore new diagnostic and therapeutic methods for precision medicine. For integrative genetic and biological analyses of the LTA-related pathway, we clarified further CAD molecules [27, 29, 30]. Improvement of the genetic infrastructure such as haplotype and linkage disequilibrium structure and genomic architecture (http://www.internationalgenome.org/) [40, 43] and the construction of large-scale genotyping array and informatics, statistical technologies, in addition to computing environment for big/large data allow global studies to clarify the genetic structure of common disorders. Many comprehensive GWASs and meta-analyses for CAD have been conducted worldwide, especially in Europe and the USA, and have identified a total of 98 loci with statistical GWAS significance [13, 31, 45]. These loci show a relatively modest effect with an increased relative risk from 1.04 to 1.92. We estimate that heritability accounts for only ~10% of these loci; however, these results could contribute several important biological and pathological pathways to CAD and reveal new insights for future precision medicine. In this book, we view and discuss the genetic architecture of CAD and its functional role that improves the establishment of future precision medicine for CAD.

2.2 The First Hypothesis-Free GWAS with a Japanese Population Connects the BRAP Inflammatory Cascade Strongly Associated with an Increased Risk of MI

Initially, we started a GWAS with a Japanese population using a high-throughput multiplex PCR-invader assay method developed by us with gene-based SNPs (approximately 100,000) as an initial stage of a global case–control association study. As far as we know, the study would be the first worldwide GWAS with comprehensive SNP markers demonstrating a disease-associated gene. From the GWAS, as a candidate risk locus for MI, we identified one SNP in the lymphotoxin-α gene (*LTA*) on chromosome 6p21.3, encoding a cytokine that is secreted at an early stage of inflammation [26, 28, 31]. Haplotype/linkage disequilibrium mapping analysis and further functional analyses showed that two SNPs (rs909253; *LTA* intron 1252A > G and rs1041981; exon 3804C > A) with functional annotations were in linkage disequilibrium in the chromosome 6 locus and associated with an increased risk of MI in the Japanese population. In the Precocious Coronary Artery Disease [PROCARDIS] study (white Europeans), a transmission disequilibrium test analysis of large trio families (447) with CAD revealed statistically significant excessive transmission ($\chi^2 = 8.44$, $P = 0.002$, recessive association model) to affected offspring for the *LTA* 804C allele (26 N-LTA) [33].

We next explored the molecules that bind to LTA to totally comprehend the role of LTA in the pathogenesis of CAD. We have therefore identified a protein, galectin-2, as a possible interaction partner of LTA with both the *Escherichia coli* two-hybrid system and a phage display method. After confirming the interaction of LTA protein and galectin-2 *in vitro* and *in vivo*, we also explored the association between genetic variants in *LGALS2* and risk of MI. We found one SNP (rs7291467; 3279C > T) in *LGALS2*; this variant decreases the level of galectin-2 (encoded by *LGALS2*) mRNA expression and showed a statistically significant association for MI [27, 28, 31]. Other researchers properly replicated the finding for rs7291467 SNP with MI by a meta-analysis [19]. This genetic variant affected the mRNA level of *LGALS2* and resulted in altered cellular secretion of LTA, and then which affected the inflammation status. We also identified that galectin-2 interacts with tubulins, important components of the microtubule complex, suggesting a role in intracellular trafficking [27, 28, 31]. LTA seems to be another protein that utilizes the microtubule cytoskeleton network for translocation, and galectin-2 mediates LTA trafficking through binding to microtubules, although the detailed role of galectin-2 in this trafficking machinery complex has yet to be elucidated.

Interaction of LTA and its cell surface receptor strongly activates nuclear factor kB (NFkB) by proteasomal degradation of its inhibitory partner, I kappa B (IkB) protein; therefore, we have hypothesized that the variation(s) in the genes encoding proteasomal proteins could confer MI susceptibility. Therefore, we have performed

comprehensive association analysis for genes encoding proteasome subunits using selected tagging SNP by linkage disequilibrium structure and identified a significant association for an SNP (rs1048990) in *PSMA6*, encoding a proteasome subunit, alpha type 6 with MI [29, 31]. Another large Chinese study robustly replicated our association with approximately the same number as our study and a meta-analysis [21]. The associated SNP, existing within 5'UTR of exon 1 in *PSMA6*, facilitated the mRNA level of *PSMA6*. Furthermore, the reduction of the mRNA expression level of *PSMA6* with short interfering RNA in cultured human cells, including coronary vascular endothelial cells, inhibited the activity of NFkB, a central mediator of inflammation regulating I kappa B stabilization [15]. Therefore, expression levels of PSMA6 protein affect the degree of inflammation reaction, suggesting that the functional variant of *PSMA6* is a genetic factor with an increased risk for MI in Japanese and Asian populations.

We further systematically explored molecular pathways associated with an increased risk of MI using a modified tandem affinity purification method [27] and identified BRAP, as a binding partner of galectin-2. We also explored genetic associations for tag SNPs in *BRAP* with MI and found the tight association for two SNPs, rs3782886 and rs11066001, in *BRAP* with increased MI risk ($P < 10^{-20}$, OR = ~1.5). Both other Japanese and Taiwanese cohorts precisely replicated the associations. Allele frequencies of these variants were hardly detected in Centre d'Etude du Polymorphism Humain individuals and Yoruba individuals, indicating that these SNPs are likely to be specific only to Asian populations. Conventional risk factors such as age, gender, diabetes, lipidemia, smoking, and blood pressure were not associated with the variants, suggesting that the variants in *BRAP* are of independent increased genetic risk for MI [30, 31].

BRCA1 associated protein (BRAP) is also known to be an E3 ubiquitin ligase that interacts with Ras and is associated with MAP kinase signaling through regulation of the scaffolding activity of kinase suppressor of ras (KSR). The MAP kinase pathway has an important physiological function associated with cell growth, cell survival regulation, cell differentiation, cell transformation, and pro-inflammatory factor production. Experiment for *BRAP* knock-down revealed suppression of NFkB activation in human coronary artery endothelial cells, suggesting that altered expression of *BRAP* might affect the expression of NFkB-dependent inflammatory molecules. We also identified that several molecules related to inflammation and cell proliferation, such as major components of IKK signalosome interacting with BRAP molecules (Fig. 2.1) [20]. Together, the findings showed that the degree of inflammation through activation of NFkB-IKK signalosome might be enhanced by up-regulated *BRAP* expression from risk alleles, thereby implying an important role in MI pathogenesis. Figure 2.1 shows the hypothetical implication of the BRAP cascade/pathway and immune/inflammation proteins in MI pathogenesis. Additional exploration of BRAP and immune/inflammatory molecules may provide useful information for exploring a novel therapeutic strategy with pharmaceutical/biological approaches. To date, we have constructed an ELISA system to screen possible molecules to intervene between BRAP and IKK signalosome (NFKBIB). We have

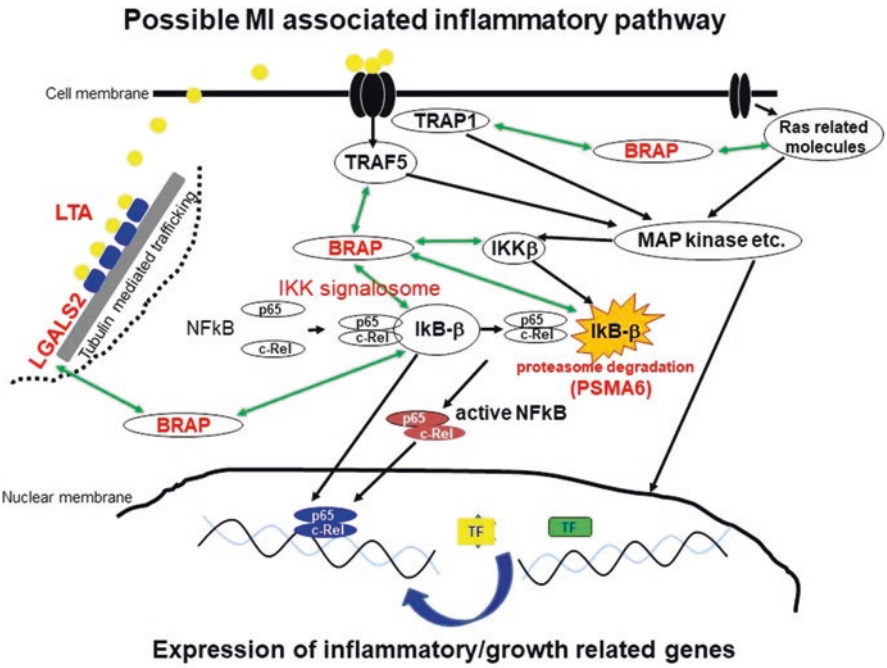

Fig. 2.1 Possible inflammatory cascade for pathogenesis of myocardial infarction. Green arrows indicate direct interaction for BRAP. *TF* transcription factor

also identified several molecules that possibly intervene between BRAP and NFKBIB interaction (unpublished).

2.3 Large Scale GWASs Reveal 98 CAD Loci

To date, 98 loci have been identified for CAD with genome-wide significance ($p < 5 \times 10^{-8}$) by comprehensive GWASs from around the world, mainly Western countries (Table 2.1). The estimated effect size from odds ratios for each variant are not so high, and still there is "missing heritability," similar to other common disorders [23]. These CAD loci are associated with conventional risk factors and only 28 loci are observed (Table 2.1; 18 associated with lipid traits, and 11 with blood pressure), indicating that there are large uncertain mechanisms with fundamental roles for CAD pathogenesis that remain to be elucidated [13, 31, 45].

In 2007, some GWASs with several thousands of Caucasian samples and several millions of SNP variants detected the association between variants on chromosome 9p21.3 and CAD [44], which, excluding African ancestry, have been robustly replicated in other races [37, 38]. At an early age of CAD onset, the risk ratio of this genetic factor increases with a small effect, but may be independent of other

Table 2.1 Coronary artery disease (CAD) genetic loci from large-scale genome-wide association study (P < 5 x 10–8)

SNP ID	Chromosome	Gene	Possible CAD-related function	Risk/non-risk allele	RAF	OR	Year
rs11206510[a]	1p32.3	*PCSK9*	LDL metabolism	T/C	0.81	1.15	2009
rs17114036	1p32.2	*PPAP2B*	Lipid synthesis	A/G	0.91	1.17	2011
rs599839[a]	1p13.3	*SORT1*	LDL metabolism	A/G	0.77	1.29	2007
rs4845625	1q21	*IL6R*	Inflammation	T/C	0.47	1.06	2013
rs11810571	1q21.3	*TDRKH*	–	G/C	0.79	1.07	2017
rs1892094	1q24.2	*ATP1B1*	–	C/T	0.5	1.04	2017
rs6700559	1q32.1	*DDX59–CAMSAP2*	–	C/T	0.53	1.04	2017
rs2820315	1q32.1	*LMOD1*	Smooth muscle cell activation	T/C	0.3	1.05	2017
rs17465637	1q41	*MIA3*	Inhibition of inflammatory cell proliferation	C/A	0.71	1.2	2007
rs515135[a]	2p24-p23	*APOB*	Cholesterol metabolism	G/A	0.83	1.07	2013
rs6544713[a]	2p21	*ABCG5-ABCG8*	Cholesterol metabolism	T/C	0.3	1.06	2013
rs1561198	2p11.2	*VAMP5-VAMP8-GGCX*	–	A/G	0.45	1.06	2013
rs2252641	2q22.3	*ZEB2-AC074093.1*	–	G/A	0.46	1.06	2013
rs6725887	2q33.1	*WDR12*	–	C/T	0.14	1.17	2009
rs2571445	2q35	*TNS1*	Smooth muscle cell–extracellular matrix associations	A/G	0.39	1.04	2017
rs1250229[b]	2q35	*FN1*	Smooth muscle cell–extracellular matrix associations	T/C	0.26	1.07	2017
rs2972146	2q36.3	*LOC646736*	–	T/G	0.65	1.06	2017
rs7623687	3p21.31	*RHOA-AMT-TCTA-CDHR4-KLHDC8B*	–	A/C	0.86	1.07	2017
rs4618210	3p24.3	*PLCL2*	Inflammation	G/A	0.42	1.1	2014
rs142695226	3q21.2	*UMPS-ITGB5*	Cell proliferation, adhesion	G/T	0.14	1.08	2017
rs9818870	3q22.3	*MRAS*	Cell proliferation, adhesion	T/C	0.15	1.15	2009
rs433903	3q25.2	*ARHGEF26-DHX36*	–	G/A	0.86	1.08	2017

rs10857147[b]	4p21.21	*PRDM8-FGF5*	Cell growth, invasion	T/A	0.27	1.06	2017
rs17087335	4q12	*REST-NOA1*	–	T/G	0.21	1.11	2015
rs11723436	4q27	*MAD2L1-PDE5A*	–	G/A	0.31	1.05	2017
rs273909[b]	4q31.1-q31.2	*GUCYIA3*	Cell differentiation, chemotaxis	G/A	0.81	1.08	2013
rs1878406	4q31.22	*EDNRA*	Vasoconstriction, inflammation	T/C	0.15	1.1	2013
rs35879803	4q31.21	*ZNF827*	–	C/A	0.7	1.05	2017
rs11748327	5p15.3	*IRX1*	–	C/T	0.76	1.25	2011
rs1800449	5q23.1	*LOX*	Smooth muscle cell–extracellular matrix associations	T/C	0.17	1.07	2017
rs273909	5q31.1	*SLC22A4-SLC22A5*	–	C/T	0.14	1.07	2013
rs246600	5q31.3	*ARHGAP26*	Smooth muscle cell–extracellular matrix associations	T/C	0.48	1.05	2017
rs12526453	6p24	*PHACTR1*	–	C/G	0.65	1.12	2009
rs35541991	6p22.3	*HDGFL1*	–	C/CA	0.31	1.05	2017
rs3798220[a]	6q25.3	*LPA*	Lipid metabolism	C/T	0.02	1.92	2009
rs17609940[a]	6p21.31	*ANKS1A*	–	G/C	0.75	1.07	2011
rs12190287	6q23.2	*TCF21*	–	C/G	0.62	1.08	2011
rs6903956	6p24.1	*C6orf105*	–	A/G	0.07	1.65	2011
rs6929846	6p22.1	*BTN2A1*	–	T/C	0.06	1.51	2011
rs10947789	6p21	*KCNK5*	–	T/C	0.76	1.07	2013
rs4252120	6q26	*PLG*	Inflammation	T/C	0.73	1.07	2013
rs10953541	7q22.3	*BCAP29*	–	C/T	0.8	1.07	2011
rs11556924	7q32.2	*ZC3HC1*	–	C/T	0.62	1.09	2011
rs2023938	7p21.1	*HDAC9*	Hematopoiesis	G/A	0.1	1.08	2013
rs10237377	7q34	*PARP12*	–	G/T	0.65	1.05	2017
rs3918226	7q36.1	*NOS3*	Production of nitric oxide	T/C	0.06	1.26	2015
rs264[a]	8p22	*LPL*	Lipid synthesis	G/A	0.86	1.11	2013

(continued)

Table 2.1 (continued)

SNP ID	Chromosome	Gene	Possible CAD-related function	Risk/non-risk allele	RAF	OR	Year
rs2954029[a]	8q24.13	*TRIB1*	Lipid metabolism	A/T	0.55	1.06	2013
rs1333049	9p21.3	*CDKN2A, B/ANRIL/IFNW1/IFNA21*	Cell proliferation, inflammation	C/G	0.47	1.47	2007
rs579459[a]	9q34.2	*ABO*	Thrombogenesis	C/T	0.21	1.33	2011
rs1412444	10q23.2-q23.3	*LIPA*	Lipid-related	T/C	0.42	1.09	2011
rs501120	10q11.1	*CXCL12*	Inflammation, lipid metabolism	T/C	0.87	1.17	2009
rs2505083	10p11.23	*KIAA1462*	Endothelial cell function	C/T	0.38	1.07	2011
rs12413409[b]	10q24.32	*CYP17A1, CNNM2, NT5C2*	Lipid synthesis	G/A	0.89	1.12	2011
rs10940293	11p15.4	*SWAP70*	Cell migration and adhesion	A/G	0.55	1.05	2015
rs1351525[b]	11p15.2	*ARNTL*	Lipogenesis	T/A	0.67	1.05	2017
rs12801636[a]	11q13.1	*PCNX3*	Lipid metabolism	G/A	0.77	1.05	2017
rs590121	11q13.5	*SERPINH1*	Plaque rupture (serine protease inhibitor derived from smooth muscle cells)	T/G	0.3	1.05	2017
rs974819	11q22.3	*PDGFD*	Inflammation, lipid synthesis	T/C	0.32	1.07	2011
rs964184[a]	11q23.3	*ZNF259, APOA5-A4-C3-A1*	LDL metabolism	G/C	0.13	1.13	2011
rs10841443	12p12.2	*RP11-664H17.1*	–	G/C	0.67	1.05	2017
rs11170820	12q13.13	*HOXC4*	–	G/C	0.08	1.1	2017
rs3184504[a,b]	12q24	*SH2B3*	–	T/C	0.38	1.13	2009
rs671	12q24	*BRAP-ALDH2*	Inflammation	A/G	0.28	1.43	2012
rs11830157	12q24.2	*KSR2*	Inflammation, cell proliferation	G/T	0.36	1.12	2015
rs2258287[a], rs2244608[a]	12q24.31	*C12orf43–HNF1A*	Lipid metabolism	A/C	0.34	1.05	2017
rs11057830[a]	12q24.31	*SCARB1*	HDL receptor	A/G	0.16	1.07	2017
rs11057401[a]	12q24.31	*CCDC92*	–	T/A	0.69	1.06	2017
rs9319428	13q12	*FLT1*	Angiogenesis, inflammation	A/G	0.32	1.06	2013
rs4773144	13q34	*COL4A1, COL4A2*	Plaque destabilization	G/A	0.44	1.07	2011

rs3832966	14q24.3	*TMED10, ZC2HC1C, RPS6KL1, NEK9, EIF2B2, ACYP1*	–	Insertion/ deletion	0.46	1.05	2017
rs2895811	14q32.2	*HHIPL1*	–	C/T	0.43	1.07	2011
rs56062135	15q22.3	*SMAD3*	Cell proliferation	C/T	0.79	1.17	2015
rs6494488	15q22.31	*OAZ2, RBPMS2*	–	A/G	0.82	1.05	2017
rs3825807	15q25.1	*ADAMTS7*	Smooth muscle cell activation	A/G	0.57	1.19	2011
rs17514846[b]	15q26.1	*FURIN-FES*	Protease convertase	A/C	0.44	1.07	2013
rs8042271	15q26.1	*MFGE8-ABHD2*	Anti-inflammatory (*MFGE8*), cell adhesion, migration (*ABHD2*)	G/A	0.9	1.1	2015
rs1050362[a]	16q22.2	*DHX38*	Cell growth	C/A	0.38	1.04	2017
rs33928862[b]	16q23.1	*BCAR1*	Cell migration, survival, transformation, invasion	Deletion/ insertion	0.51	1.05	2017
rs3851738	16q23.1	*CFDP1*	–	C/G	0.6	1.05	2017
rs7500448[b]	16q23.3	*CDH13*	Cell adhesion	A/G	0.77	1.07	2017
rs216172	17p13.3	*SMG6, SRR*	–	C/G	0.37	1.07	2011
rs12936587	17p11.2	*RASD1, SMCR3, PEMT*	–	G/A	0.56	1.07	2011
rs46522	17q21.32	*UBE2Z, GIP, ATP5G1, SNF8*	Insulin resistance (GIP)	T/C	0.53	1.06	2011
rs17608766[b]	17q21.32	*GOSR2*	–	C/T	0.14	1.07	2017
rs7212798	17q23.2	*BCAS3*	Control cell polarity and motility in endothelial cells	C/T	0.15	1.08	2015
rs1867624	17q23.3	*PECAM1*	Cell–cell adhesion	T/C	0.61	1.04	2017
rs663129	18q21.3	*PMAIP1-MC4R*	Generation of reactive oxygen species	A/G	0.26	1.06	2015
rs1122608[a]	19p13	*LDLR*	LDL metabolism	G/T	0.75	1.15	2009
rs3803915	19p13.3	*AP3D1-DOT1L-SF3A2*	–	C/A	0.19	1.12	2014
rs2075650[a]	19p13.32	*APOE-APOC1*	LDL metabolism	G/A	0.14	1.14	2011

(continued)

Table 2.1 (continued)

SNP ID	Chromosome	Gene	Possible CAD-related function	Risk/non-risk allele	RAF	OR	Year
rs12976411	19q13.1	*ZNF507-LOC400684*	–	T/A	0.09	1.49	2015
rs138120077	19q13.2	*HNRNPUL1, TGFB1, CCDC97*	–	Deletion/ insertion	0.14	1.07	2017
rs8108632	19q13.2	*TGFB1, B9D2*	–	T/A	0.48	1.05	2017
rs1964272	19q13.32	*SNRPD2*	–	G/A	0.51	1.05	2017
rs867186[b]	20q11.22	*PROCR*	Endothelial cell function	A/G	0.89	1.08	2017
rs9982601	21q22	*SLC5A3-MRPS6-KCNE2*	–	T/C	0.13	1.2	2009
rs180803	22q11.2	*POM121L9P-ADORA2A*	Anti-inflammatory (*ADORA2A*)	G/T	0.97	1.2	2015

ID identifier, *RAF* risk allele frequency, *OR* odds ratio; –, unknown function for CAD

[a]Associated with lipid traits

[b]Associated with blood pressure

conventional risk factors such as lipidemia. Moreover, the association between the 9p21.3 locus and the increased risk of other diseases including abdominal aortic and intracranial aneurysms, type 2 diabetes, Alzheimer's disease, subclinical phenotype for CAD, and cancers, has been observed, but in different variants from CAD, indicating pleiotropic effects of the associations for the 9p21 locus and many disorders. The 9p21.3 variants associated with CAD exist in nearby *CDKN2B-AS1*, a long noncoding RNA (lncRNA), close to the genes *CDKN2A* and *B*, encoding cyclin-dependent kinase inhibitor proteins. An association of the higher expression of mRNA for *CDKN2B-AS1* with the CAD risk allele of 9p21.3 was found in the functional analysis; however, an inverse association was observed in the expression of *CDKN2A/B* mRNA. In adipose tissue, a statistical association between *CDKN2B* expression and the 9p21.3 SNP was revealed in an eQTL analysis. Identification of a putative enhancer for the 9p21.3 CAD locus and subsequent chromatin conformation capture to detect long-range chromosome interaction by Harismendy et al. revealed that the interval of the enhancer interacts physically with the chromosome loci, *CDKN2A/B*, *MTAP*, and further chromosomes downstream of *IFNA21*, encoding interferon alpha 21, in vascular endothelial cells. On the contrary, other studies follow up the above findings with several cells, including aortic smooth muscle and endothelial cells did not support interferon-related inflammatory cascade for 9p21.3 variant suggesting that there might be unidentified uncertain mechanisms for the 9p21.3 risk variant [22].

Reilly et al. performed a GWAS in CAD patients with MI and those without MI and identified an association with a protective role on several SNPs tagging the O allele in the ABO blood group at chromosome 9p34.2 with MI [35]. A Japanese population with MI replicated this association [12], but no association was found with CAD [41]. *ABO* contributes to the blood group system. The gene encodes proteins (transferase A, alpha 1–3-N-acetylgalactosaminyltransferase; transferase B, alpha 1–3-galactosyltransferase), which transfer carbohydrate to von Willebrand Factor (vWF). By a deletion of guanine-258 near the N-terminus of the protein, the O allele encodes a protein without any enzymatic activity and thus cannot modify the vWF molecule, which is assumed to enhance the proteolysis of vWF and results in circulating vWF and Factor VIII in lower concentrations. Associations of ABO blood group with LDL-C, type 2 diabetes and inflammatory adhesion molecules, and ACE activity are also observed. These findings suggest that ABO proteins might have multiple functions implicating thrombosis and/or plaque rupture that are associated with the risk of MI. In the future, the clarification of the detailed mechanism associated with MI with ABO and clinical studies are required, people with blood group A, B or AB may receive therapies such as antiplatelet agent treatment [3, 38].

As a druggable CAD-associated gene in GWAS hits, we can suggest a gene named *PCSK9,* which encodes a calcium-dependent serine endoprotease and belonging to the proprotein convertase subtilisin/kexin (PCSK) family, an enzyme that cleaves latent precursor proteins to biologically/physiologically active molecules. In an initial study, PCSK9 was identified as a protein encoded by a gene with gain-of-function mutations for two families with hypercholesterolemia [1], was a druggable molecule that dramatically reduced LDL-C, and its clinical use was

investigated [39]. PCSK9 binds with LDL cholesterol receptor in liver and resolves the receptor to inactivate it. A full human monoclonal antibody for PCSK9 to inhibit the LDL receptor interaction certainly decreased circulating levels of LDL cholesterol in humans, which seemed to have no side effects in a phase III clinical study [34]. GWAS also identified another molecule, *FURIN,* which also encodes a PCSK family member, and mainly expressed atherosclerotic plaques in humans, indicating that the molecule might be a druggable molecule druggab to atherosclerotic diseases. We need further investigation into atherosclerotic diseases and FURIN function. These findings implicate that the genetic diversities for disease risk are associated with the pathogenesis of certain disorders and significantly contribute to discovery of the druggable targets for atherosclerotic diseases, implying the significant power of comprehensive genetic analysis, including GWAS, that contributes to exploring novel unanticipated insights/knowledge for precision medicine.

On the other hand, several GWASs identified six CAD loci in an East Asian population, among which we have identified four loci with genome-wide significance for CAD in a Japanese population that are close to the genes *IRX1*, *BRAP-ALDH2*, *PLCL2*, and *AP3D1-DOT1L-SF3A2* on chromosome 5p15, 3q24, 12q24, and 19p13 respectively [2, 12]. In another Japanese study, Yamada et al. identified an SNP with functionality into *BTN2A1* on chromosome 6 [47] and Takeuchi et al. also showed a *BRAP-ALDH2* locus [41] for CAD risk with GWAS significance. In the GWAS with Han Chinese, Wang et al. reported a 6p24 locus for the certain risk of CAD [46], whereas, these Asian-associated loci for CAD failed in the Caucasian GWASs. This racial difference may be explained partly by the variance among race in allelic frequencies and the study power owing to the small number of samples and the difference in relative risk ratio. The ethnic diversity in the architecture of genetic structure such as the presence of undiscovered hidden variations and accurate differences in linkage disequilibrium structure in Europeans may be an influence. We have left to answer the questions related in ethnic diversity for these loci associated with several traits.

The hypothetic functions of the molecules near the CAD loci are shown in Table 2.1. We could divide these functions roughly into five groups associated with lipid metabolism: inflammation, cell adhesion, cell migration, cell proliferation, and unknown function. A mega GWAS by the CARDIoGRAMplusC4D Consortium identified 15 novel CAD loci in 2013. They also conducted pathway analysis *in silico* and identified four common pathways: liver/retinoid X receptor activation, atherosclerosis signaling, acute phase response signaling, and retinoid X receptor activation, molecules related to lipid metabolism and inflammation. Furthermore, a large meta-analysis GWAS included a further 30 risk loci for CAD mainly in Caucasians and implicated in CAD with arterial-wall-specific and blood vessel morphogenesis, cell adhesion, cell migration, angiogenesis, insulin pathway, signaling of nitric oxide, and inflammation/immune pathway [13, 16, 25, 45]. The accompanying role related to thrombosis/atherosclerosis and rupture of plaque for these molecules and pathways remains to be elucidated. This evidence may enhance the discovery of therapeutic/diagnostic targets for further biological, physiological, and pharmacological examination.

2.4 Genetic Variants with High Odds Ratios

Renovation of technologies for nucleic acid sequencing and informatics permit us the comprehensive discovery of rare variants (mutations) with pathogenicity in the whole genome and whole exome with large populations. Exome sequencing with several thousand European and African individuals with CAD and investigation into the association between the rare variants and plasma triglyceride levels were performed by the Exome Sequencing Project [42]. They found druggab loss-of-function mutations in *APOC3*, encoding apolipoprotein C. In triglyceride levels, the carriers with the variants were 39% lower than those of carriers without the variants and were 40% lower in the CAD risk than those of non-carriers. The rare coding variants in two genes, *APOA5* and *LDLR*, were respectively associated increased MI risk with exome-wide significance in another exome sequencing project with ~ 5000 individuals with early onset MI and controls [9]. The increased risk ratios for MI individuals were 4.2-fold for *LDLR* variants and 2.2-fold for *APOA5* variants. The dysfunctional mutations in nitric oxide signaling genes, *GUCY1A3* and *CCT7*, were associated with increased MI risk by yet another exome study with a large MI family. *In vitro* and *in vivo* functional experiments showed enhanced thrombosis formation by reduced nitric oxide signaling via downregulated expression and enzyme activity of the mutated proteins. Dewey et al. in the DiscovEHR human genetic study [7, 8] reports sequences with tens of thousands of people for whole exons of the angiopoietin-like 3 and 4 genes (*ANGPTL3* and *ANGPTL4*). They identified loss-of-function variants and examined the association with the variants. They also identified that the variants were associated with lower lipid levels (LDL, HDL, triglycerides, and total cholesterol for the variants of *ANGPTL3*, and triglycerides for the variants of *ANGPTL4*) and reduced CAD risk. They additionally showed that the reduction of the lipid levels and odds of atherosclerotic CAD by pharmacological inactivation targeted these genes. These findings provide novel insights for early diagnosis of asymptomatic patients and biological, physiological, and pharmaceutical investigation for precision medicine of atherosclerotic disorders, even if these variants have only a small impact on CAD heritability (< ~ 1%).

2.5 Summary

To our knowledge, the world's first GWAS with no hypothesis in Japanese ultimately elucidated a candidate pathology of MI by implicating an inflammatory/immune cascade including the IKK signalosome and BRAP encoded by the gene was certainly associated with increased MI/CAD risk in the Asian population. The inflammatory cascade, including NFkB signaling, plays a pivotal role in the pathogenesis of CAD/atherosclerosis as also suggested in an analysis by CARDIoGRAMplusC4D consortium and others. It would be a common final pathway that emerges as inflammation that is present in CAD/atherosclerosis. These

findings resulted from hypothesis-free, genome-wide, and comprehensive studies, which indicates the potent power of the GWAS to identify a pathogenetic pathway for this common but serious disease. The genetic knowledge could also apply to facilitating future biological and pharmaceutical investigations to develop innovative diagnoses and therapies to adapt to individuals who suffer from certain disease, which is called precision medicine.

Many GWASs identified large numbers of unanticipated genetic risk loci for CAD/atherosclerosis, and highlighted new clues to future preventive medicine for the development of genetic diagnosis and new druggable CAD targets. Each genetic factor contributes to CAD pathogenesis with modest effects and gathering of all variants cannot explain well the previously estimated heritability by epidemiological findings. Although several exome sequencing studies discovered pivotal CAD risk variants and contributed valuable knowledge to the medical field, the exploration of missing heritability remains to be clarified. There seem to be many common variants with small effects for CAD susceptibility that remain to be clarified; however, the additional rare variants that have a relatively large effect on transcriptional regulatory elements, including histone modification regions, Dnase hypersensitivity, and methylation sites, may contribute to the investigation of missing heritability. Whole-genome sequencing for these regulatory elements with appropriately large individuals and suitable informatics techniques will elucidate this common issue for common diseases in the future. Only a small portion of genetic variants identified through GWAS is dependent on conventional risk factors and hardly explains the molecular function that mediates CAD susceptibility; thus, comprehending the molecular mechanisms for CAD/atherosclerosis risks affected by these genetic factors will be focused on in the next era.

Coronary artery disease is attributable to arterial defects, is a common but serious disorder, and is a leading cause of death worldwide. Elucidating the genetic architecture contributing to CAD pathogenesis would lead to the discovery of innovative diagnoses, preventive measures, and therapy that can be adapted to each individual suffering from the disorder, the so-called precision medicine.

References

1. Abifadel M, Varret M, Rabeè JD, Allard D, Ouguerram K, Devillers M et al (2003) Mutations in PCSK9 cause autosomal dominant hypercholesterolemia. Nat Genet 34:154–156
2. Aoki A, Ozaki K, Sakata Y, Takahashi A, Kubo M et al (2011) SNPs on chromosome 5p15.3 associated with myocardial infarction in Japanese population. J Hum Genet 56:47–51
3. Bampali K, Mouzarou A, Lamnisou K, Babalis D (2013) Genetics and coronary artery disease: present and future. Hell J Cardiol 55(2):156–163
4. Braunwald E (1997) Shattuck lecture–cardiovascular medicine at the turn of the millennium: triumphs, concerns and opportunities. N Engl J Med 337:1360–1369
5. Breslow JW (1997) Cardiovascular disease burden increases, NIH funding decreases. Nat Med 3:600–601
6. Collins FS, Guyer MS, Charkravarti A (1997) Variations on a theme: cataloging human DNA sequence variation. Science 278:1580–1581

7. Dewey, FE. Gusarova, V. O'Dushlaine, C. Gottesman, O. Trejos, J. et al. Inactivating variants in *ANGPTL4* and risk of coronary artery disease. N Engl J Med 374, 1123–1133 (2016)
8. Dewey, FE. Gusarova, V. Dunbar, RL. O'Dushlaine, C. Schurmann, C. et al. Genetic and pharmacologic inactivation of *ANGPTL3* and cardiovascular disease. N Engl J Med (2017) 377(3), 211–221
9. Do R, Stitziel NO, Won HH, Jørgensen AB, Duga S, Angelica Merlini P (2015) Exome sequencing identifies rare *LDLR* and *APOA5* alleles conferring risk for myocardial infarction. Nature 518(7537):102–106
10. Ebana Y, Ozaki K, Inoue K, Sato H, Iida A, Lwin H et al (2007) A functional SNP in ITIH3 is associated with susceptibility to myocardial infarction. J Hum Genet 52:220–229
11. Haga H, Yamada R, Ohnishi Y, Nakamura Y, Tanaka T (2002) Gene-based SNP discovery as part of the Japanese Millennium Genome project: identification of 190,562 genetic variations in the human genome. J Hum Genet 47:605–610
12. Hirokawa M, Morita H, Tajima T, Takahashi A, Ashikawa K, Miya F et al (2014) A genome-wide association study identifies PLCL2 and AP3D1-DOT1L-SF3A2 as new susceptibility loci for myocardial infarction in Japanese. Eur J Hum Genet 23:374–380
13. Howson JMM, Zhao W, Barnes D, Ho W-K, Young R et al (2017) Fifteen new risk loci for coronary artery disease highlight arterial-wall-specific mechanisms. Nat Genet 49:1113–1119
14. Ishii N, Ozaki K, Sato H, Mizuno H, Saito S, Takahashi A et al (2006) Identification of a novel non-coding RNA, MIAT, that confers risk of myocardial infarction. J Hum Genet 51:1087–1099
15. Karin M, Delhase M (2000) The I kappa B kinase (IKK) and NF-kappa B: key elements of proinflammatory signalling. Semin Immunol 12:85–98
16. Klarin D, Zhu QM, Emdin CA, Chaffin M, Horner S, MacMillan BJ et al (2017) Genetic analysis in UK Biobank links insulin resistance and transendothelial migration pathways to coronary artery disease. Nat Genet 49:1392–1397
17. Konta A, Ozaki K, Sakata Y, Takahashi A, Morizono T et al (2016) A functional SNP in *FLT1* increases risk of coronary artery disease in a Japanese population. J Hum Genet 61:435–441
18. Lander ES (1996) The new genomics: global views of biology. Science 274:536–539
19. Lian J, Fang P, Dai D, Ba Y, Yang X, Huang X et al (2014) Association between *LGALS2* 3279C>T and coronary artery disease: a case-control study and a meta-analysis. Biomed Rep 2:879–885
20. Liao YC, Wang YS, Guo YC, Ozaki K, Tanaka T, Lin HF et al (2011) BRAP activates the inflammatory cascades and increases the risk for carotid atherosclerosis. Mol Med 17:1065–1074
21. Liu X, Wang X, Shen Y, Wu L, Ruan X, Lindpaintner K et al (2009) The functional variant rs1048990 in *PSMA6* is associated with susceptibility to myocardial infarction in a Chinese population. Atherosclerosis 206(1):199–203
22. McPherson R (2013) Chromosome 9p21.3 locus for coronary artery disease: how little we know. J Am Coll Cardiol 62(15):1382–1383
23. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ (2009) Finding the missing heritability of complex diseases. Nature 461:747–753
24. Ohnishi Y, Tanaka T, Ozaki K, Yamada R, Suzuki H, Nakamura Y (2001) A high-throughput SNP typing system for genomewide association studies. J Hum Genet 46:471–477
25. Nelson CP, Goel A, Butterworth AS, Kanoni S, Webb TR, Marouli E et al (2017) Association analyses based on false discovery rate implicate new loci for coronary artery disease. Nat Genet 49:1385–1391
26. Ozaki K, Ohnishi Y, Iida A, Sekine A, Yamada R, Tsunoda T et al (2002) Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction. Nat Genet 32:650–654
27. Ozaki K, Inoue K, Sato H et al (2004) Functional variation in LGALS2 confers risk of myocardial infarction and regulates lymphotoxin-alpha secretion in vitro. Nature 429:72–75

28. Ozaki K, Tanaka T (2005) Genome-wide association study to identify SNPs conferring risk of myocardial infarction and their functional analyses. Cell Mol Life Sci 62:1804–1813
29. Ozaki K, Sato H, Iida A, Iida A, Ohnishi Y, Sekine A et al (2006) A functional SNP in PSMA6 confers risk of myocardial infarction in the Japanese population. Nat Genet 38:921–225
30. Ozaki K, Sato H, Inoue K, Tsunoda T, Sakata Y, Mizuno H et al (2009) SNPs in BRAP associated with risk of myocardial infarction in Asian populations. Nat Genet 41:329–333
31. Ozaki K, Tanaka T (2016) Molecular genetics of coronary artery disease. J Hum Genet 61:71–77
32. Peden JF, Farrall M (2011) Thirty-five common variants for coronary artery disease: the fruits of much collaborative labour. Hum Mol Genet 20:R198–R205
33. PROCARDIS Consortium (2004) A trio family study showing association of the lymphotoxin-alpha N26 (804A) allele with coronary artery disease. Eur J Hum Genet 12:770–774
34. Raal FJ, Stein EA, Dufour R, Turner T, Civeira F, Burgess L et al (2015) PCSK9 inhibition with evolocumab (AMG 145) in heterozygous familial hypercholesterolaemia (RUTHERFORD-2): a randomised, double-blind, placebo-controlled trial. Lancet 385:331–340
35. Reilly MP, Li M, He J, Ferguson JF, Stylianou IM, Mehta NN et al (2011) Identification of ADAMT7 as a novel locus for coronary atherosclerosis and association of ABO with myocardial infarction in the presence of coronary atherosclerosis: two genome wide association studies. Lancet 377:382–392
36. Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. Science 273:1516–1517
37. Roberts R, Stewart AF (2012) Genes and coronary artery disease. J Am Coll Cardiol 60(18):1715–1721
38. Roberts R (2014) Genetics of coronary artery disease: an update. Methodist Debakey Cardiovasc J 10(1):7–12
39. Stein EA, Mellis S, Yancopoulos GD, Stahl N, Logan D, Smith WB et al (2012) Effect of a monoclonal antibody to PCSK9 on LDL cholesterol. N Engl J Med 366(12):1108–1118
40. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A et al (2015) An integrated map of structural variation in 2,504 human genomes. Nature 526:75–81
41. Takeuchi F, Yokota M, Yamamoto K, Nakashima E, Katsuya T, Asano H et al (2012) Genome-wide association study of coronary artery disease in the Japanese. Eur J Hum Genet 20:333–340
42. TG and HDL Working Group of the Exome Sequencing Project (2014) National Heart, Lung, and Blood Institute. Loss-of-function mutations in APOC3, triglycerides, and coronary disease. N Engl J Med 371(1):22–31
43. The 1000 Genomes Project Consortium (2015) A global reference for human genetic variation. Nature 526:68–74
44. The Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447:661–678
45. Verweij N, Eppinga RN, Hagemeijer Y, van der Harst P (2017) Identification of 15 novel risk loci for coronary artery disease and genetic risk of recurrent events, atrial fibrillation and heart failure. Sci Rep 7:2761
46. Wang F, Xu CQ, He Q, Cai JP, Li XC, Wang D et al (2011) Genome-wide association identifies a susceptibility locus for coronary artery disease in the Chinese Han population. Nat Genet 43:345–349
47. Yamada Y, Nishida T, Ichihara S, Sawabe M, Fuku N, Nishigaki Y et al (2011) Association of a polymorphism of BTN2A1 with myocardial infarction in East Asian populations. Atherosclerosis 215:145–152

Chapter 3
Genetic and Functional Genetics of Autoimmune Diseases

Kazuhiko Yamamoto, Kazuyoshi Ishigaki, Akari Suzuki, and Yuta Kochi

Abstract The majority of autoimmune diseases are multi-factorial diseases that develop through the interaction of several factors, such as genetic and environmental factors. A limited number of disease susceptibility genes, including those of the major histocompatibility complex have been known to exist for several decades. After these eras, genome-wide association studies have been used for more than 10 years to identify susceptibility genes for certain autoimmune diseases. These findings have contributed to our understanding of the pathogenesis of these diseases. As the analysis of susceptibility genes has progressed, it has become apparent that many disease susceptibility gene variants are involved at the expression level of genes. Furthermore, expression of genes related to disease pathogenesis is cell-specific, with involvement of epigenetic mechanisms. Genetic information exists before the onset of disease, and thus has a causal relationship to the disease. Therefore, the analysis of genomic function in human immunology research is essential, with regard to understanding the pathological mechanisms as well as having applications for drug discovery. In this article, we discuss these issues, with a particular focus on rheumatoid arthritis.

Keywords Autoimmune diseases · Rheumatoid arthritis · GWAS · eQTL · Functional genomics

K. Yamamoto (✉) · A. Suzuki · Y. Kochi
Laboratory for Autoimmune Diseases, Riken Center for Integrative Medical Sciences, Yokohama, Japan
e-mail: kazuhiko.yamamoto@riken.jp; yuta.kochi@riken.jp

K. Ishigaki
Laboratory for Autoimmune Diseases, Riken Center for Integrative Medical Sciences, Yokohama, Japan

Laboratory for Statistical Analysis, Riken Center for Integrative Medical Sciences, Yokohama, Japan

T. Tsunoda et al. (eds.), *Genome-Wide Association Studies*,
https://doi.org/10.1007/978-981-13-8177-5_3

3.1 Genetic Factors in Autoimmune Diseases

Genetic factors have been reported to be involved in the pathogenesis of autoimmune diseases. The concordance rates for autoimmune diseases in monozygotic twins are higher than those for dizygotic twins, and these are higher than the prevalence in the general population. For example, the concordance rate of monozygotic twins with rheumatoid arthritis (RA) is approximately 12–15%, compared to 2–4% for dizygotic twins [1]. On the other hand, the prevalence of RA in the general population has been reported to be 0.5–1%, although there are some differences among ethnic groups. These pieces of evidence strongly indicate genetic factors are involved in the majority of autoimmune diseases, even though factors such as infections, socioeconomic status and the environment are also likely to be involved.

3.2 The Major Histocompatibility Complex Group of Genes

The major histocompatibility complex (MHC), also called the Human Leukocyte Antigen complex (HLA) in human, is important in immune responses. Association between genetic variants in the *HLA* gene and autoimmune diseases has been reported, not only in RA but other diseases such as systemic lupus erythematosus (SLE), ankylosing spondylitis, Behcet's disease, Graves' disease, and type 1 diabetes. Since class I and class II molecules of HLA have the function of presenting antigen to T cells, the association with these immune diseases is understandable. However, as will be describe later, it is not clear whether antigen presentation is the only function of HLA in the pathogenesis of autoimmune diseases.

Historically, serological typing of the HLA class II molecules exhibited HLA-DR1 and HLA-DR4 had strong associations with RA. Furthermore, it was reported that several alleles of *DRB1* ∗01:01, 04:01, 04:04, 04:05 that encode for the β chain of the DR antigen are involved. As the stretch of amino acid residues (70–74) that corresponded to the β chain hypervariable region encoded by these alleles is common, a shared epitope (SE) hypothesis was proposed by Gregersen et al. [2]. However, as recent studies revealed that the amino acid positions 11 and 13 in HLA-DRB1 were also influential, revising the SE hypothesis has been recommended [3]. Nevertheless, over all the concept of the SE hypothesis has not changed, indicating that the HLA class II molecules translated from disease susceptible alleles bind and present the epitopes of RA specific antigens. It has been shown that HLA class II molecules of RA susceptible alleles have high avidities to citrullinated peptides. In fact, anti-citrullinated protein antibodies (ACPA) are auto-antibodies with the highest disease specificity in RA. According to these lines of data, the functional understanding of the RA susceptibility HLA class II genes is in progress [4].

Meanwhile, Okada and colleagues, together with the International Collaborative Research Group, utilized the HLA imputation method to undertake a large-scale fine-mapping analysis of the *HLA* gene sequence [5]. They analyzed samples

collected from Japanese people (6244 RA and 23,731 controls) and Asian (7097) and European (23,147) control populations. They found that in addition to the classical *HLA* genes such as *HLA-DRB1*, *HLA-DPB1*, *HLA-B*, the *HLA-DOA* gene, a non-classical *HLA* gene, was also involved in RA. Unlike the classical *HLA* genes, changes in expression levels of the non-classical *HLA* gene shown to be involved in the onset of the disease [5].

3.3 Analysis of Non-HLA Genetic Factors by Genome-Wide Association Analysis

Along with decoding the whole human genome, it has become possible to analyze disease susceptibility gene polymorphisms of common diseases, including autoimmune diseases. Among several different variants, single nucleotide polymorphisms (SNPs) were shown to be important because of feasibility of their analysis and the functional significance of the variants. Since the beginning of this century, researchers at the RIKEN institute focused on this area of research using a method that could be described as a prototype of the method that is now widely used. Further, the international HapMap project analyzed haplotype blocks and selected tag SNPs [6]. Along with progress in the development of technology to type genetic variants using microarrays, genome wide association studies (GWAS) have become a common approach [7]. Since 2007, GWAS of common diseases has been performed globally, with many reports published on autoimmune diseases [8, 9].

Okada et al. recently undertook a meta-analysis of GWAS studies of RA, analyzing 29,880 RA patients and 73,758 controls, with Asian and European ancestries [9]. They identified that 42 novel loci were associated with RA, increasing the total number of gene loci to 101 that show susceptibility to RA. However, it is important to understand that each locus has potentially multiple genes in a linkage disequilibrium block. Therefore, various databases were integrated to estimate genes and SNPs, most likely to be associated with RA among each locus. Furthermore, RA associated genes were found to be significantly enriched (via the network of protein-protein interactions) in target genes of drugs currently being used to treat RA. These findings not only provided us with important information about the pathogenesis of RA, but also demonstrated a new strategy for drug discovery using GWAS. For example, some drugs for other diseases were also found potentially to targets RA genes.

Although some of the RA risk variants were found to be involved in generating qualitatively different proteins with alterations in amino acid regions, many other RA risk SNPs were involved at the expression levels of genes [9]. This is called expression quantitative trait loci (eQTLs) (Fig. 3.1). It has been estimated that the accumulation of differences in gene function and expression levels due to such genetic variants was indeed associated with the pathogeneses of polygenic diseases. For example, it was reported that 53% of chromosomal regions associated with

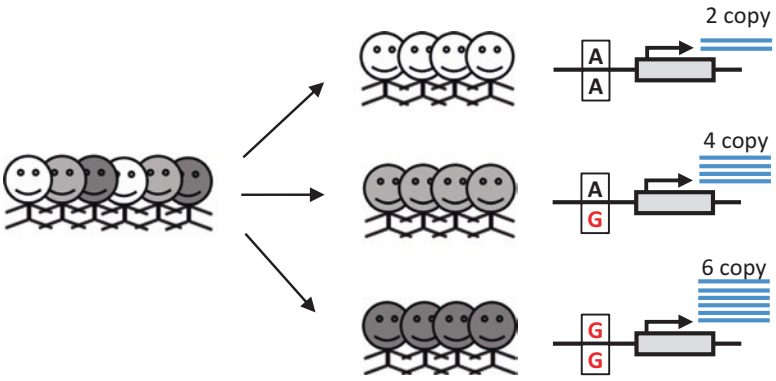

Fig. 3.1 An eQTL (expression quantitative trait loci): a quantitative trait loci with gene expression levels affected by nucleotide sequences differences

Celiac diseases revealed by GWAS are eQTL, being involved at the level of gene expression. Recently, it has been observed that by including variants that regulate genes at a distance (trans), disease risk SNPs became more frequently eQTL than the initial estimation [10].

3.4 Hidden Heritability in GWAS of Autoimmune Diseases

GWAS has been performed under the common disease-common variant hypothesis or "common diseases are caused by common genetic variants". In fact, more than 100 disease susceptibility loci that occur in relatively high frequencies have been elucidated, even in one disease such as RA. However, it has been suggested that we are not able to sufficiently explain all of the genetic contribution to disease, even when combining the GWAS risk variants. Thus, it is possible that more common variants are involved or rare variants, occurring at a lower frequency.

Sequencing of the exon region of the sialic acid acetylesterase (SIAE) gene, previously shown to be involved in immune-tolerance pathogenesis in mice, was performed on patients with autoimmune disease as well as healthy human samples [11]. As a result, rare variants that lacked gene function were found in patients with multiple autoimmune diseases. The frequency of these variants was significantly different between patients (24/923) and healthy individuals (2/648). The odds ratio to inherit the SIAE variants was 8.3 in patients with RA, whilst the odds ratio for the majority of common variants of autoimmune disease susceptibility genes was 1.1–1.5. Thus, the contribution of each of these rare variants to the pathogenesis of individuals is high. Conversely, these types of rare variants are at a low frequency in the general population. Subsequently, many studies have investigated the incidence of rare variants in patients with autoimmune diseases, the majority observing difficulty of identifying causal rare variants.

3.5 Genetic Factors Common to Multiple Autoimmune Diseases

There are several disease susceptibility genes that are common to many types of autoimmune diseases. It has also been reported that multiple autoimmune diseases have been observed in the same family and an individual can suffer from multiple autoimmune diseases. Thus, it is possible that the basic mechanisms for different autoimmune diseases are shared. However, it is important to point out that genetic factors common to multiple diseases do not apply to all autoimmune diseases. For example, *STAT4* is a common risk gene for RA and SLE. *STAT4* is a transcription factor that regulates important cytokines, such as IL-12, IL-23 and type I interferon and is thought to be involved in the differentiation of Th1 type and Th17 type CD4+ T cells. On the other hand, it has been reported that *STAT3* was associated with Crohn's disease and multiple sclerosis. This suggests that different transcription factors are expressed by different helper T cell subsets and thus contribute to different diseases. Furthermore, when we look at the pathway involving disease susceptibility genes for RA, many of the genes involved in NF-κB signaling, such as *CD40*, *TRAF1*, *TNFAIP3*, *PRKCO*, *TNFRSF14* have been found to be involved. This suggests the NF-κB signaling pathway is also involved in the pathogenesis of RA.

Further examples of genes associated with multiple autoimmune diseases are described below:

1. *PTPN22* (protein tyrosine phosphatase nonreceptor-type 22)

PTPN22 is a gene that encodes lymphoid tyrosine phosphatase (LYP), which suppresses a signal from both T cell receptor and B cell receptors [12]. The SNP in the *PTPN22* gene responsible for type 1 diabetes, SLE, RA and Graves' disease in European American people has a specific amino acid substitution, R620W. Interestingly, the R620W variants has not been identified in East Asian populations, including Japanese people. Upon signal transduction, tyrosine kinase phosphorylates tyrosine of various molecules, while tyrosine phosphatase such as *PTPN 22* dephosphorylates them. Therefore, *PTPN22* is considered to act as a negative regulator of antigen receptor signaling.

2. *CTLA4* (cytotoxic T lymphocyte-associated protein 4)

CTLA4 has been identified as a susceptibility gene for several autoimmune diseases such as RA, SLE, type 1 diabetes and Graves' disease [13]. The *CTLA4* gene encodes a molecule that is expressed on the cell membrane of T cells where it transmits an inhibitory signal. It also plays an important role in regulatory T cells. A soluble molecule exists as a splicing variant in *CTLA4*, and the expression of this soluble *CTLA4* decreases in disease susceptible alleles. Thus it is estimated the soluble CTLA4 molecule plays an important role in maintaining immune tolerance.

3. *FCRL3* (Fc receptor-like 3)

Genetic variants in the *FCRL3* gene have been shown to be associated with RA, SLE and Graves' disease [14]. The site of the risk SNP in the promoter region of *FCRL3* strongly binds to the transcription factor NF-κB, increasing gene expression [14]. The molecule *FCRL3* was highly expressed by mature B cells. When it was strongly cross-linked to the B cell receptor, *FCRL3* suppressed the signal, suggesting *FCRL3* is associated with B cell related immune tolerance. Furthermore, a functional relationship between the *FCRL3* SNP and regulatory T cells has also been reported.

4. *IL23R* (interleukin 23 receptor)

It has been identified that the *IL23R* is a disease association gene that encodes a protein with an amino acid change (R381Q) in Crohn's disease, psoriasis and ankylosing spondylitis. The functional effect of this mutation is unclear. However, since IL23 is a cytokine essential for differentiation and maintenance of Th17 lymphocytes, it is estimated that Th17 lymphocytes have an important role in these diseases. Conversely, the association of this polymorphism with RA and SLE is not clear.

5. *CCR 6* (chemokine (C – C motif) receptor 6)

The *CCR6* gene has also been shown to be associated with susceptibility to RA, Graves' disease and Crohn's disease [15]. Gene expression in the disease susceptible allele is relatively higher, most likely due to differences in binding to transcription factors. The protein encoded by *CCR6* is expressed on T cells, B cells and dendritic cells, and highly expressed by Th17 T cell subsets, suggesting a role in the migration of Th17 T cells to inflammatory sites.

3.6 *PADI4* (Polymorphisms Specifically Associated with RA)

We reported that a gene encoding an enzyme called peptidylarginine deiminase type 4 or *PADI4* was associated with susceptibility to RA in the Japanese population [16]. Initially, *PAD14* was classified as an ethnic specific disease susceptibility gene because the result was replicated in Asians (Japanese, Korean and Chinese) by large-scale follow-up analysis, but not in those with European ancestries. However, it has also been shown recently using meta-analyses (with better detection capabilities), that European and American populations have the susceptible allele. There are two main *PADI4* gene haplotypes, and mRNA transcribed from the RA susceptible haplotype is more stable than that from the non-susceptible haplotype. The enzyme encoded by the *PADI* genes, PAD, is involved in post-translational modification, converting an arginine residue into citrulline. As a result of this citrullination reaction, the protein loses a positive charge. Therefore, citrullination may influence the three-dimensional structure of the molecule, potentially altering its antigenicity and

function. Although various autoantibodies have been detected in the sera of RA patients, ACPA have been found to be highly specific (from 89% to 98%). Joint destruction was reported to be more advanced in patients who were positive for ACPA. These findings affirmed an association between *PADI4* and RA, not only based on genetics but on biological consequences. However, it was shown that *PADI4* was also involved in the formation of neutrophil extracellular traps (NETs). Therefore, the involvement of *PAD14* in the pathogenesis of RA needs further investigation [17].

3.7 From GWAS to Functional Genomics

As previously discussed, the majority of SNPs associated with risk for autoimmune diseases have been found to be eQTL. Thus, genomic function of disease susceptible variants could be studied based on this information. Recently, we sampled peripheral blood mononuclear cells (PBMC) donated by 100 healthy individuals and separated using a cell sorter, for the major immune cells such as CD4+ T cells, CD8+ T cells, B cells, NK cells and monocytes. Gene expression by these cells was quantified using a next generation sequencer (RNA-Seq). Furthermore, we analyzed the relationship of gene expression with genetic variants, and created an eQTL catalog (Fig. 3.2). As most of the previous studies analyzed whole blood leukocytes, this study generated a novel database to investigate how genetic variants influence the expression levels of specific genes, in particular immune cells (Published at the National Bioscience Database Center (NBDC)). The expression levels of genes differed in each cell type. It is understood that the epigenome provides another mechanism for cell specificity. Thus, we could determine which cell types express specific genes with the variants associated with disease risk. This information can only be obtained when cell subsets are analyzed separately.

Furthermore, by applying the eQTL catalog, we developed a new method to analyze the pathogenesis of immune diseases, focusing on the direction of abnormal gene expression regulation caused by risk-associated variants. Usually, a single risk variants could not be anticipated to have significant influence on disease onset. However, with this method, the influence of multiple risk variants could be evaluated and the results interpreted based on our understanding of a certain pathway. Specifically, by analyzing the genetic information of RA patients and healthy individuals, we predicted the effects of 176 genes involved in TNF receptor downstream pathways in CD4+ T cells. As a result of our analysis, it was confirmed that the activation of TNF receptor regulated pathways in CD4 + T cells was important for the pathology of RA [17]. Importantly, this result was obtained from healthy individuals, therefore not influenced by environmental factors such as treatment regimes. Therefore, only the genetic contribution to the disease could be evaluated. The TNF inhibitor is an effective treatment option for RA patients and TNF signaling is known to be important in the pathogenesis of RA. However, our study identified that downstream pathways of the TNF receptor in CD4+ T cells were specifically involved in the patogenesis [18].

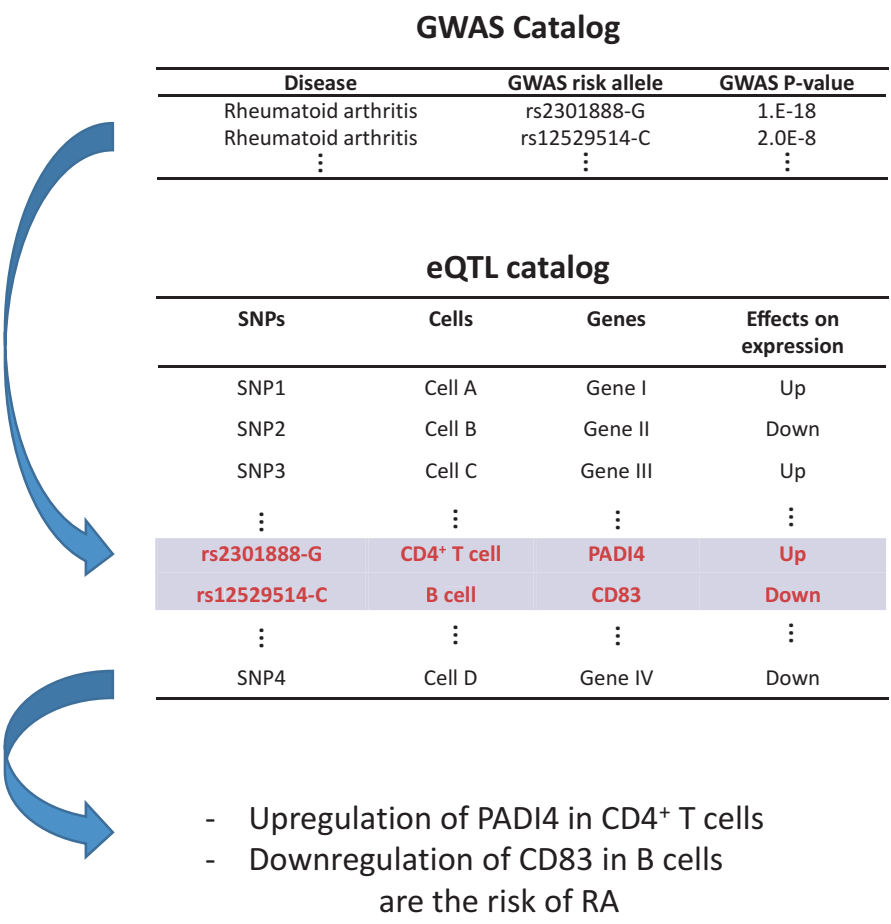

Fig. 3.2 Example of an eQTL catalog and example of a list of DNA variants involved in the pathogenesis of rheumatoid arthritis

3.8 Epigenetic Research

As described above, the concordance rates of monozygotic twins with RA are higher than dizygotic twins. However, since these rates are not 100%, factors other than genetics are also involved in pathogenesis of RA. It has been suggested that environmental factors working through the epigenome have an important role in autoimmune diseases, by controlling gene expression in a cell specific manner. In fact, differences in DNA methylation and gene expression between monozygotic twins displaying the onset of SLE have been observed [19].

Since the early 1990s, the methylation levels in peripheral T cells have been observed to be lower in active SLE than inactive SLE. An analysis of genome-wide DNA methylation in CD4+ T cells indicated changes in methylation levels according

to the progression of SLE [20]. More comprehensively, a study that compared methylation of about 46,000 CpG sites on DNA in CD4+ T cells, CD 19+ B cells and CD 14 + monocytes from both SLE patients and healthy individuals was reported [21]. This study showed methylation was lower in SLE patients compared to healthy controls, especially in the vicinity of the gene where the alleles were associated with risk for SLE. The hypomethylation levels of T cell, B cell and monocytes between SLE patients and controls was significant. With regards to histone modification, histone H3 and H4 hypoacetylation and H3K9 hypomethylation in CD4+ T cells has been reported in SLE patients but not compared to healthy individuals [22].

Similar to those observed in SLE patients, DNA from synovial fibroblast-like cells collected from RA patients was demethylated, most likely inducing aggressive granulation in tissues [23]. Suppression of DNA methyl-transferase 1 (*DNMT1*) activity results in demethylation. For example, abnormal demethylation of the CpG site of the *MMP13* gene increased *MMP13* expression, and subsequently degradation of type II collagen in cartilage by the protease activity of MMP 13. With regards to histone modification, histone deacetylase (*HDAC*) activity differs between various types of cells. For example, *HDAC* expression is enhanced in peripheral blood [24] and synovial fibroblast-like cells, but the total *HDAC* activity has been reported to be low in synovial tissue [23]. In RA synovial tissue, hypomethylation of histone H3K9, and hyperacetylation of histone H3, H4 have also been observed [25].

Abnormal expression of microRNA (miRNA) has been also associated with autoimmune diseases [26], although miRNA regulation slightly differs from the fundamental concept of the epigenetic modifications described above. It has been shown that miR-146a is a regulator of inflammatory cytokines such as TNF-α. Interestingly, miR-146a expression levels decreased in patients with SLE but were elevated in RA patients [26]. It has also been reported that long-chain ncRNA (lncRNA), was involved in autoimmune diseases through tissue-specific transcriptional regulation and present at higher levels than miRNA [27].

3.9 The Integration of Functional Genomics into Human Immunology Research

Our immune system consists of higher-order functions through the interaction of various cell types, molecules and genes. To date, the immune system has mainly been investigated using mouse models, including through inactivation of specific genes (knockout mice). Overall, mouse and human immune systems are similar. However, there are differences, in particular where a treatment shown to be effective for an immunological disease in mice does not work in humans. Therefore, it has been recognized that immunological research in humans is important. In this respect, human studies that examine only the immune responses may not provide information about causal relationships. For example, data on gene expression, protein expression, or epigenetic changes alone cannot indicate whether they are a

cause or consequence. Therefore, similar to gene knockout studies in mice, an investigative methodology that clarifies both cause and consequence should be adopted for human immunology research. In this context, the study of disease-susceptible genetic variants is important. With the exception of antigen receptor genes, a patient's genetic information exists before the disease onset and does not change. These findings provide us with evidence into the causal relationship of the observed phenomenon and its pathogenesis. As discussed earlier, many of the disease susceptible variants identified by GWAS have been found to function as an e-QTL, regulating the expression levels of genes. The SNP regions associated with RA significantly overlap the histone mark of an active promoter and enhancer in T cells from RA patients [28]. Therefore, using global genomic information, qualitative and quantitative analyses of gene expression together with information about disease susceptible variants, cell specific epigenomes and proteins, we will better understand the pathogenic components of immuno-competent cells in various immune-related diseases. This research will make it possible to elucidate causal intermediate phenotypes such as gene expression, epigenome and protein expression patterns in individual diseases. By comprehensively understanding the human immune system, it could be possible to elucidate the immune status of each individual in more detail, making precision medicine a reality [29, 30].

References

1. Silman AJ et al (1993) Twin concordance rates for rheumatoid arthritis: results from a nationwide study. Br J Rheumatol 32:903–907
2. Gregersen PK et al (1987) The shared epitope hypothesis. An approach to understanding the molecular genetics of susceptibility to rheumatoid arthritis. Arthritis Rheum 30(11):1205–1213
3. Raychaudhuri S et al (2012) Five amino acids in three HLA proteins explain most of the association between MHC and seropositive rheumatoid arthritis. Nat Genet 44:291–296
4. Scally SW et al (2013) A molecular basis for the association of the HLA-DRB1 locus, citrullination, and rheumatoid arthritis. J Exp Med 210:2569–2582
5. Okada Y et al (2016) Contribution of a Non-classical HLA Gene, HLA-DOA, to the Risk of Rheumatoid Arthritis. Am J Hum Genet 99:366–374
6. International HapMap Consortium (2005) A haplotype map of the human genome. Nature 437:1299–1320
7. Hirschhorn JN et al (2005) Genome-wide association studies for common diseases and complex traits. Nat Rev Genet 6:95–108
8. Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447:661–678
9. Okada Y et al (2014) Genetics of rheumatoid arthritis contributes to biology and drug discovery. Nature 506:376–381
10. Ram R et al (2017) Effects of type 1 diabetes risk alleles on immune cell gene expression. Genes 8:167
11. Surolia I et al (2010) Functionally defective germline variants of sialic acid acetylesterase in autoimmunity. Nature 466:243–247

12. Gregersen PK et al (2006) Genetics of autoimmune diseases–disorders of immune homeostasis. Nat Rev Genet 7:917–928
13. Ueda H et al (2003) Association of the T-cell regulatory gene CTLA4 with susceptibility to autoimmune disease. Nature 423:506–511
14. Kochi Y et al (2005) A functional variant in FCRL3, encoding Fc receptor-like 3, is associated with rheumatoid arthritis and several autoimmunities. Nat Genet 37:478–485
15. Kochi Y et al (2010) A regulatory variant in CCR6 is associated with rheumatoid arthritis susceptibility. Nat Genet 42:515–519
16. Suzuki A et al (2003) Functional haplotypes of PADI4, encoding citrullinating enzyme peptidylarginine deiminase 4, are associated with rheumatoid arthritis. Nat Genet 34:395–402
17. Seri Y et al (2015) Peptidylarginine deiminase type 4 deficiency reduced arthritis severity in a glucose-6-phosphate isomerase-induced arthritis model. Sci Rep 5:13041
18. Ishigaki K et al (2017) Polygenic burdens on cell-specific pathways underlie the risk of rheumatoid arthritis. Nat Genet 49:1120–1125
19. Javirre BM et al (2010) Changes in the pattern of DNA methylation associate with twin discordance in systemic lupus erythematosus. Genome Res 20:170–179
20. Jeffries MA et al (2011) Genomewide DNA methylation patterns in CD4+ T cells from patients with systemic lupus erythematosus. Epigenetics 6:593–601
21. Absher DM et al (2013) Genome-wide DNA methylation analysis of systemic lupus erythematosus reveals persistent hypomethylation of interferon genes and compositional changes to CD4+ T-cell populations. PLoS Genet 9:e1003678
22. Hu N et al (2008) Abnormal histone modification patterns in lupus CD4+ T cells. J Rheumatol 35:804–810
23. Nakano K et al (2013) DNA methylome signature in rheumatoid arthritis. Ann Rheum Dis 72:110–117
24. Gillespie J et al (2012) Histone deacetylases are dysregulated in rheumatoid arthritis and a novel histone deacetylase 3-selective inhibitor reduces interleukin-6 production by peripheral blood mononuclear cells from rheumatoid arthritis patients. Arthritis Rheum 64:418–442
25. Karouzakis E et al (2009) DNA hypomethylation in rheumatoid arthritis synovial fibroblasts. Arthritis Rheum 60:3613–3622
26. Dai R et al (2011) MicroRNA, a new paradigm for understanding immunoregulation, inflammation, and autoimmune diseases. Transl Res 157:163–179
27. Huang W et al (2015) DDX5 and its associated lncRNA Rmrp modulate TH17 cell effector functions. Nature 528:517–522
28. Trynka G et al (2013) Chromatin marks identify critical cell types for fine mapping complex trait variants. Nat Genet 45:124–130
29. Ye CJ et al (2014) Intersection of population variation and autoimmunity genetics in human T cell activation. Science 345:1254665
30. Farh KK-H et al (2015) Genetic and epigenetic fine mapping of causal autoimmune disease variants. Nature 518:337–343

Chapter 4
Genome-Wide Association Study for Type 2 Diabetes

Minako Imamura, Momoko Horikoshi, and Shiro Maeda

Abstract Genome-wide association studies (GWAS) have facilitated a substantial and rapid rise in the number of confirmed genetic susceptibility variants for type 2 diabetes (T2D) and glycemic traits. Approximately 90 variants for conferring susceptibility to T2D and 80 variants for glycemic traits have been identified until the end of 2016. This success has led to widespread hope that the findings will translate into improved clinical care for the increasing numbers of patients with diabetes. Potential areas or clinical translation include risk prediction and subsequent disease prevention, pharmacogenomics, and the development of novel therapeutics. In contrast, worldwide efforts to identify susceptibility loci to diabetic nephropathy have not been successful so far, and most of heritability for diabetic nephropathy remains to be elucidated. Uncovering the missing heritability is essential to the progress of T2D genetic studies and to the translation of genetic information into clinical practice.

Keywords Type 2 diabetes · Insulin secretion · Insulin resistance · Nephropathy · Chronic kidney diseases

M. Imamura · S. Maeda (✉)
Department of Advanced Genomic and Laboratory Medicine, Graduate School of Medicine, University of the Ryukyus, Nishihara, Japan

Division of Clinical Laboratory and Blood Transfusion, University of the Ryukyus Hospital, Nishihara, Japan

Laboratory for Endocrinology, Metabolism and Kidney Diseases, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan
e-mail: smaeda@med.u-ryukyu.ac.jp

M. Horikoshi
Laboratory for Endocrinology, Metabolism and Kidney Diseases, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan

T. Tsunoda et al. (eds.), *Genome-Wide Association Studies*,
https://doi.org/10.1007/978-981-13-8177-5_4

4.1 GWAS for Type 2 Diabetes

More than 400 million people are affected by diabetes mellitus worldwide, and the number of patents is estimated to rise to more than 600 million by 2040 [1]. Increasing prevalence of diabetes is a serious concern in many countries. Of the total global diabetes rate, 90% are living with type 2 diabetes (T2D), which is characterized by insulin resistance in peripheral tissues and impairments of insulin secretion from pancreatic β-cells. Although the current rise in T2D prevalence is explained mainly by changes in life-style, complex genetic determinants are widely considered to contribute to an inherent susceptibility to this disease [2–5]. A sibling relative risk of T2D was reported to be approximately 2 [4], and its heritability has been estimated at 30–70% [5]. Like other common diseases, the pathogenesis of T2D is considered polygenic, and the effects of individual genetic factors are modest by themselves [6]. Development of high-throughput genotyping technologies and statistical and computational software has allowed remarkable progress over the past decades in the research fields for genome-wide search to discover novel genetic loci for T2D susceptibility [6]. In 2007, five GWAS for T2D performed by European and American Groups identified robust susceptibility loci to European T2D, and until 2016, more than 90 T2D susceptibility loci have been identified through GWAS in different ethnic groups. The empirical threshold for statistical significance used here is $p < 5 \times 10^{-8}$ unless a different study-wise threshold has been applied and noted. It is also important to remember that loci are labeled by the gene(s) nearest to or functionally plausible for the association signal and that they do not necessarily explain the true functional gene responsible for the signal.

Genetics of T2D: Before the GWAS Era

Prior to the GWAS era, the importance of genetic factors in the etiology of T2D had been well established through family and twin studies [2–5], and the linkage analysis and candidate-gene association studies were applied as the primary approaches to identify susceptibility loci for diseases or phenotypic traits [7, 8]. Reynisdottir et al. identified segments in chromosomes 5 and 10 with suggestive linkage to T2D [8], and showed that the chromosome 10 region harbored the *TCF7L2* [9]. Subsequently, the association of *TCF7L2* with T2D was replicated not only in populations of European descent but also in other ethnic groups [10–16], including the Japanese [17, 18]. Candidate-gene association studies showed that *PPARG* [19] and *KCNJ11* [20] were susceptibility genes for T2D. Both genes encode targets of antidiabetes medications (thiazolidinediones and sulfonylureas, respectively) and harbor missense variants associated with T2D: P12A in *PPARG* and E23K in *KCNJ11* [19, 20]. The successful identification of these genes encouraged the genetic study of T2D; however, these classical approaches were not recognized as suitable to identify variants that have a smaller effect on disease susceptibility. Therefore, the

discovery of novel T2D susceptible loci had been challenging, and a more powerful strategy was needed to overcome this difficulty.

The Initial Phase of GWAS Era of T2D Genetics (2007–2008)

In 2007, GWAS for T2D was conducted in a French population composed of 661 cases and 614 controls, covering 392,935 SNP (single nucleotide polymorphism) loci. This study identified novel association signals at *SLC30A8*, *HHEX*, *LOC387761*, and *EXT2* and validated the association at *TCF7L2* previously identified through linkage analysis [21]. Shortly after the French GWAS, the Icelandic study group confirmed the association of *SLC30A8*, *HHEX*, and the newly identified *CDKAL1* with T2D [22]. At the same time, three collaborating groups, the Wellcome Trust Case Control Consortium/United Kingdom Type 2 Diabetes Genetics consortium (WTCCC/UKT2D), the Finland-United States Investigation of NIDDM (FUSION), and the Diabetes Genetics Initiative (DGI), reported the consistent associations of *SCL30A8*, *HHEX*, *CDKAL1*, *IGF2BP2*, and *CDKN2A/B* with T2D in European populations [23–25]. These novel loci and two previously known variants (*PPARG* P12A and *KCNJ11* E23K) were confirmed by multiple replication studies composed of European and non-European populations with the exception of *LOC387761* and *EXT2*. Thus, the first round of European GWAS confirmed eight T2D susceptibility loci across multiple ethnic groups: *TCF7L2*, *SLC30A8*, *HHEX*, *CDKAL1*, *IGF2BP2*, *CDKN2A/B*, *PPARG*, and *KCNJ11* [21–25]. In addition to these eight loci, the WTCCC/UKT2D study identified a strong association between *FTO* variants and T2D, although the effect of *FTO* variants on conferring susceptibility to T2D was mostly mediated through increase in body weight [26].

After the first round of European GWAS, an effort was made to increase sample size so that common variants with smaller effect sizes would be detectable. WTCCC/UKT2D, FUSION, and DGI combined their data to form the Diabetes Genetics Replication and Meta-analysis (DIAGRAM) consortium. Six additional novel loci, *JAZF1*, *CDC123/CAMK1D*, *TSPAN/LGR5*, *THADA*, *ADAMTS9*, and *NOTCH2*, were identified in a genome-wide scan comprising a substantial sample size (4549 cases and 5579 controls) and more than 2.2 million SNPs (either directory genotyped or imputed), followed by replication testing [27].

GWAS in Groups of East Asian Descent (2008–2011)

Over the past decades, many Asian countries have experienced a dramatic increase in the prevalence of T2D. Cumulative evidence suggests that Asians may be more susceptible than populations of European ancestry to insulin resistance and diabetes, which was thought to be due to differences in interethnic genetic inheritance [28]. Many of the association of the T2D loci identified by European GWAS,

especially the first round of GWAS, have been confirmed in Japanese populations [6, 29, 30]. However, there are significant interethnic differences in the risk allele frequency or in effect sizes at several loci, which may affect the power to detect associations in these populations. For example, risk allele frequencies of *TCF7L2* variants showing the strongest effect on T2D in European populations are very few in the Japanese (~5%) compared to populations of European descent (~40%) [17, 18]. Consequently, the association of *TCF7L2* variants and T2D appears statistically less significant in the Japanese [17, 18]. In addition, the effects of some loci identified through European T2D GWAS were not consistent in Japanese populations [6, 29, 30]. Therefore, it is necessary to identify ethnic group-specific T2D susceptibility loci, those have not been captured by the European studies, to explain T2D heritability in populations of Asian descent.

In 2008, two independent Japanese GWAS, conducted by Millennium genome project (MHLW) and BioBankJapan (BBJ), simultaneously identified the *KCNQ1* locus as a strong T2D susceptibility locus in the Japanese [31, 32]; this was the first established T2D susceptibility locus through non-European GWAS. Subsequent replication studies performed in different ethnic groups revealed that single nucleotide variants located at intron 15 of *KCNQ1* had the strongest effects on conferring susceptibility to T2D in several East Asian populations [33–36]. The association of the *KCNQ1* locus with T2D was replicated in European populations, but the minor allele frequencies in Europeans were considerably lower than those in East Asian populations (~7% in Europeans versus ~40% in East Asians). Thus, in contrast to *TCF7L2*, the attributable fraction of *KCNQ1* on T2D susceptibility was relatively small in European populations. Since the *KCNQ1* locus was not captured in the European studies, this finding emphasizes the importance of examining susceptibility loci in different ethnic groups. Although the two Japanese GWAS successfully identified the *KCNQ1* locus, these studies had limited sample sizes at the initial stage of the genome-wide scan: MHLW, 187 T2D cases vs. 752 controls [32]; BBJ, 194 T2D cases vs. 1558 controls [31].

A Japanese GWAS of a larger sample size (4470 T2D vs. 3071 controls) discovered additional two T2D susceptibility loci, *UBE2E2* and *C2CD4A-C2CD4B* in 2010 [37]. Associations between these loci and T2D were confirmed in East Asian replication study [37] and large-scale European GWAS afterward [38], suggesting GWAS for T2D using non-European as well as European populations is useful to facilitate identification of both ethnicity-specific and common-susceptibility loci among different ethnic groups.

An effort was made to increase sample size in East Asian population as well as in European combined their data to form Asian Genetic Epidemiology Network (AGEN) consortium [39]. Eight additional novel loci, *GLIS3*, *PEPD*, *FITM-R3HDML-HNF4A*, *KCNK16*, *MAEA*, *GCC1-PAX4*, *PSMD6*, and *ZFAND3*, were identified in a genome-wide scan comprising a substantial sample size (6952 cases and 11,865 controls) followed by replication testing (Stage 2 in silico replication analysis 5843 cases and 4574 controls de novo replication analysis 12,284 cases and 13,172 controls) [39].

GWAS with Imputation and Large-Scale Meta-Analyses (2012–)

In order to identify common variants of weaker effects, efforts have been made to increase sample size by combining association data from multiple cohorts by meta-analyses. DIAGRAM consortium has constantly developed the scale of collaboration, incremental meta-analyses (DIAGRAM+ and DIAGRAM v3) [38, 40] adding GWA data from further studies from European descent to DIAGRAM v1 data (DIAGRAM+; total of 8130 cases and 38,097 controls [40], DIAGRAM v3; total of 12,171 cases and 56,862 controls [38]) together with extensive replication have identified additional loci (12 and 8 loci, respectively).

In the meantime, four additional loci (*ANK1*, *MIR129-LEP*, *GPSM1*, and *SLC16A11-SLC16A13*) have been identified by Japanese GWAS, with increment of the sample size [41] and number of variants examined by the imputation of genotypes [29, 41]. The latest Japanese GWAS meta-analysis has identified seven additional T2D susceptibility loci (*CCDC85A*, *FAM60A*, *DMRTA1*, *ASB3*, *ATP8B2*, *MIR4686*, and *INAFM2*) in a genome-wide scan comprising the largest sample size in the East Asian population (15,463 cases and 26,183 controls) followed by replication testing (7936 cases and 5539 controls) [30].

Thus, larger GWAS meta-analyses combined multiple cohorts with homogeneous populations have continued to expand the number of T2D loci. In 2014, motivated by a consistency of common variant associations observed across different populations [42, 43], a trans-ethnic GWAS meta-analysis of more than 110,000 individuals, which combined GWAS data in multiple ethnic groups including European, East Asian, South Asian, and Mexican/Mexican American, has been performed [44]. Seven additional new loci for T2D susceptibility were successfully identified by combining GWAS from multiple ancestry groups, which highlighted the benefits of trans-ethnic GWAS [44].

Established susceptibility loci for T2D identified by 2016 are shown in Fig. 4.1.

What Have T2D GWAS Brought About So Far?

Identified Loci for T2D Linked More Frequently to β-Cell Function than to Insulin Sensitivity

The etiology of T2D is a combination of β-cell dysfunction and insulin resistance, which is promoted by either genetic or environmental factors (e.g., obesity, Westernized diet, and lifestyle). Interestingly, majority of the known T2D susceptible variants appear to influence insulin secretion rather than insulin resistance. For example, large meta-analysis from DIAGRAM+ demonstrated that of 31 confirmed T2D susceptibility loci, 10 (*MTNR1B*, *SLC30A8*, *THADA*, *TCF7L2*, *KCNQ1*, *CAMK1D*, *CDKAL1*, *IGF2BP2*, *HNF1B*, and *CENTD2*) were nominally associated with reduced homeostasis model assessment of β-cell function (HOMA-β) which

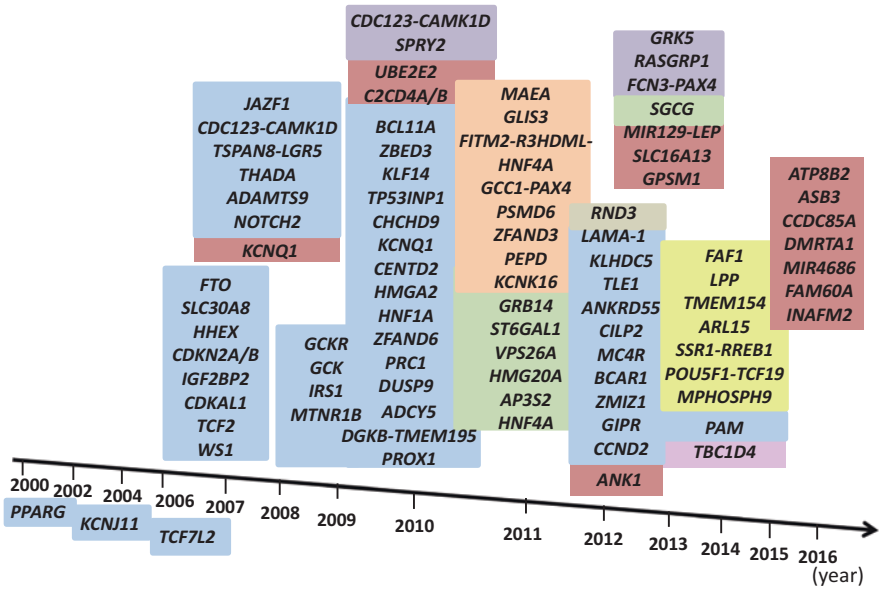

Fig. 4.1 Established T2D susceptible loci. The *x*-axis shows the year of publication. Background color indicates ethnic composition of the samples in the discovery GWAS: European (blue), Japanese (red), Chinese (purple), African American (gray), East Asian (orange), South Asian (green), trans-ethnic (yellow), and Inuit (pink)

estimates steady-state β-cell function and only 3 (*PPARG*, *FTO*, and *KLF14*) were associated with HOMA of insulin resistance (IR)(HOMA-IR), an indicator for insulin resistance [40]. Consistent result was observed in larger study afterward [45]. Moreover, the loci identified in the early phase of Japanese GWAS, namely, *KCNQ1*, *UBE2E2*, *C2CD4A-C2CD4B*, and *ANK1* were shown to be associated with decreased β-cell function in nondiabetic control groups [29, 32, 34, 37]. Prior to the accumulation of GWAS data, a genetic predisposition to insulin resistance had been considered to play dominant roles in development of T2D, especially in populations of European origin [40]. The results obtained from GWAS, however, emphasize the crucial role of the pancreatic β-cell in the onset of T2D, and a genetic predisposition to reduced β-cell function may contribute more to the susceptibility to T2D.

Missing Heritability

GWAS have successfully identified novel T2D susceptibility loci that had not been captured by classical approaches. However, it has been estimated that only ~10% of the known T2D heritability could be explained by those T2D susceptibility loci [38, 46]. Because polygenic analyses in the European ancestry GWAS have suggested many more common variant loci not yet reaching genome-wide significance could contribute to the heritability of T2D susceptibility [38, 46], residual genetic

variance explained by a long tail of common variant signals of lesser effect could be captured in larger-scale analyses of various individual ethnic populations or trans-ethnic meta-analysis. The rationale of GWAS is based on the "common disease-common variant" hypothesis, and studies have focused on finding common variants associated with the disease; therefore, susceptibility variants having a minor allele frequency (MAF) of less than 1% are frequently missed, with limited exceptions [47, 48]. It has been a matter of considerable debate whether low-frequency risk variants, which could be evaluated by next generation sequencing and may have relatively large effects, could explain the missing heritability. To test this hypothesis, the GoT2D and T2D-GENES consortia performed whole-genome sequencing (WGS, $n = 2657$) and whole-exome sequencing (WES, $n = 12{,}940$) with 26.7 million variants, including 4.16 million low frequency (0.5 <MAF <5%) or 6.26 million rare (MAF <0.5%) variants. The results indicated variants associated with T2D after sequencing were overwhelmingly common (MAF >5%); therefore they concluded that this sequencing analysis did not support the idea that lower-frequency variants have a major role in predisposition to T2D [49], although sample sizes for initial WGS/WES were considered too small for rare variants analyses.

Translation of T2D Genetics into Clinical Practice

The Possibility of Disease Prediction and Prevention

One of the most anticipated clinical uses of genetic information is to predict an individual's risk of developing T2D. Indeed, genetic investigations suggested lifestyle intervention arm of the Diabetes Prevention Program (DPP) attenuated genetic risk defined by carrying *TCF7L2* risk allele [10] or GRS constructed with 34 confirmed loci attenuated risk of developing diabetes [50]; these are good examples of the clinical usefulness of genetic testing to allow detection of high-risk individuals with whom physicians should aggressively intervene. Since the discovery of multiple T2D risk genetic variants, genetic risk score (GRS) calculated based on the number of risk alleles in subjects who developed disease has become a common approach to indicate individual's genetic risk. Our study group examined the utility of GRS based on 49 established T2D loci (GRS-49) in the Japanese (Fig. 4.2) [51]. GRS-49 was significantly associated with T2D risk in a Japanese population, and those with a GRS ≥60 (5.7% of the population examined) were 9.81 times as likely to have type 2 diabetes compared with those with a GRS <46 (4.2% of the population examined) (Fig. 4.2b) [51]. The result suggested even though the impact of each T2D susceptibility locus was very small, accumulation of genetic information was useful to detect a high-risk group for the disease in a population. However, the area under the receiver operating characteristic (ROC) curves for GRS was 0.624, and the effect of adding GRS into clinical factor (age, sex, and BMI) was as small as 0.03 even though the incremental effect was statistically significant (Fig. 4.2c) [51]. The performance of genetic prediction models using GRS has been evaluated

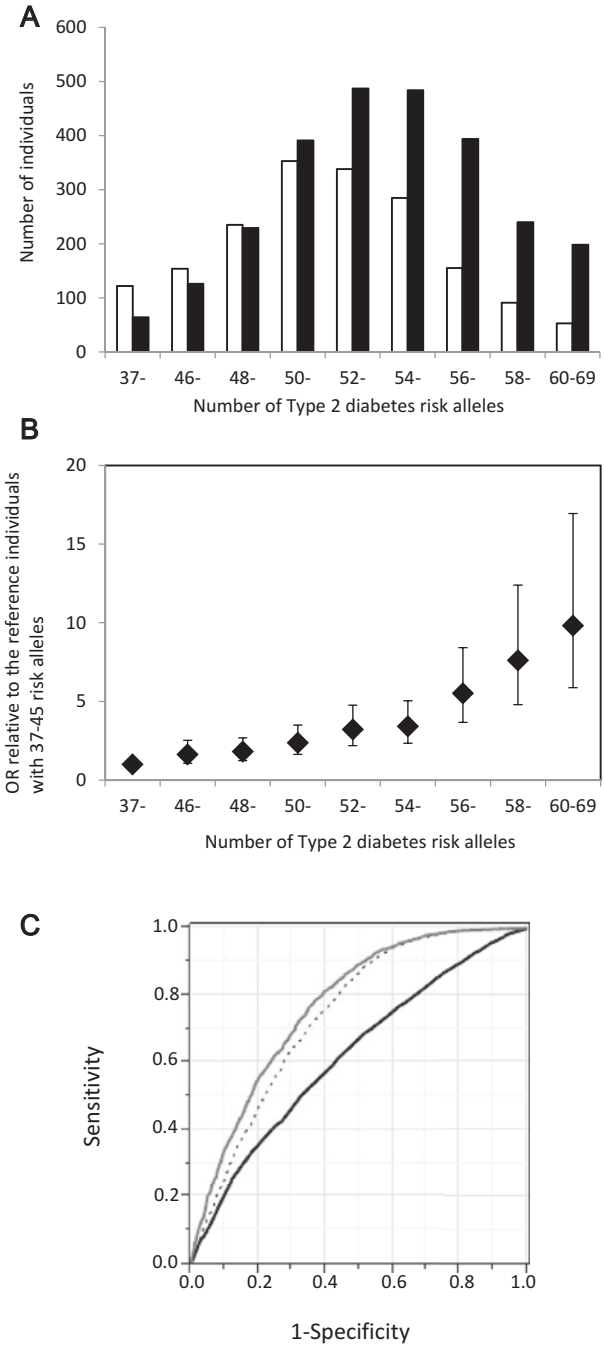

Fig. 4.2 Evaluation of a genetic risk score (GRS) constructed by summing up the number of risk alleles for GWAS-derived 49 single nucleotide variants (SNVs) in 4399 Japanese participants [51].

in over 30 studies including Asian and European with case-control study sets or prospective cohorts [52]. The results were consistent among these studies including ours [51]: AUCs of genetic information alone for T2D were 0.579–0.641 and incremental predictive performance of T2D using established marker is statistically significant but limited [52]. Insufficient information is available to construct a genetic risk score for T2D because of so-called missing heritability, and it is far from translating into clinical practice at present. Identification of causal variants, epigenetic modifications, gene-gene interactions, and gene-environment interactions as well as uncovering residual T2D susceptible genetic variants may improve the clinical utility of genetic information for T2D prediction [52].

The Possibility of Identifying Novel Biological Mechanisms and Therapeutic Targets

Because GWAS is a biology-agnostic method to detect genetic variations that predispose to a disease, the results may contribute to identify novel biological mechanisms, which may lead to discover novel therapeutic targets for T2D. However, uncovering underlying molecular mechanisms by which the loci contribute to susceptibility to type 2 diabetes has been behind, compared with GWAS discovery. A major obstacle is that the causal variants and molecular mechanisms for diabetes risk are unknown in the most of the identified T2D susceptibility loci. Furthermore, most genetic risk variants are found in the intronic or noncoding regions of genes and are more likely to affect regulation of transcription rather than gene function. Thus, it has been challenging to elicit novel biological insight, which may uncover the disease pathogenesis and guide drug discovery from GWAS derived genetic information.

To identify biological candidate for causal genes at established T2D risk loci systematically, our study group utilized an in silico pipeline, originally developed by Okada et al. [53], using various publicly available bioinformatics methods based on (i) functional annotation, (ii) cis-acting expression quantitative trait loci, (iii) pathway analyses, (iv) genetic overlap with monogenic diabetes, (v) knockout mouse phenotypes, and (vi) PubMed text mining [30]. Seven genes (*PPARG*, *KCNJ11*, *ABCC8*, *GCK*, *KIF11*, *GSK3B*, and *JUN*) were identified as potential drug targets for T2D treatment by integrating disease-associated variants with diverse genomic and biological datasets and subsequent drug target search (Fig. 4.3) [30]. Of these, *PPARG*, *KCNJ11*, and *ABCC8* have been well known as targets for the

Fig. 4.2 (continued) (**a**) Distribution of the number of risk alleles in patients with T2D (black bars, $n = 2613$) and controls (white bars, $n = 1786$). (**b**) Odds rates (ORs) for individual groups with different number of T2D risk alleles relative to the reference group having 37–45.5 risk alleles. The vertical bars represent 95% confidence intervals. (**c**) Receiver Operating Characteristic plot for model 1 containing GRS (black line, area under the curve (AUC) = 0.624); model 2 containing sex, age, and body mass index (BMI) (broken line, AUC = 0.743); and model 3 containing GRS, age, sex, and BMI (gray line; AUC = 0.773)

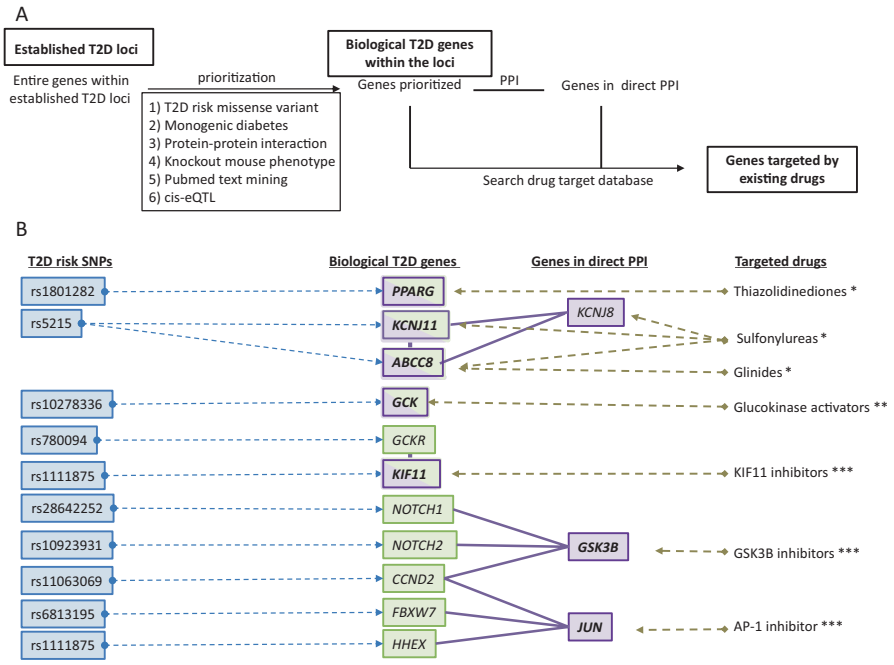

Fig. 4.3 Discovery of potential drug targets for the treatment of T2D [30]. (**a**) Strategy for drug targets search based on the genetic information derived from GWAS. Biological T2D risk genes were selected from among the T2D potential risk genes located in any of the established T2D risk loci, using a scoring system by summing up the number of prioritization criteria satisfied. We selected novel T2D therapeutic targets from the overlapping genes between the drug target genes and the biological T2D risk genes or genes those products are in direct PPI with the biological T2D risk gene products. (**b**) Representative connections between T2D risk SNVs (blue), T2D biological genes (green), drug target genes (purple), and targeted drugs. ∗ Approved compounds for T2D treatment ∗∗Compounds for T2D treatment under clinical trial ∗∗∗ Compounds under clinical trial for treatment against diseases other than T2D

already approved T2D treatment options, and a *GCK* activator is currently undergoing clinical trials for the T2D treatment. Thus, this in silico pipeline was capable to detect drug target of established T2D treatment suggesting the capability for developing novel T2D treatment. Inhibitors for *KIF11*, *GSK3B*, and *JUN* were under clinical trial for the treatments of cancers (*KIF11*, *GSK3B*) or rheumatoid arthritis (*JUN*); these compounds could also be potential treatments for T2D [30]. Thus, systematic approaches for integrating the findings of genetic, biological, and pharmacological studies could be a useful strategy for developing new T2D treatments.

4.2 GWAS of Metabolic Traits

The etiology of T2D is characterized by reduced insulin secretion due to impaired beta-cell dysfunction and the presence of insulin resistance. The heritability of insulin secretion, peripheral insulin action, and nonoxidative glucose metabolism has been investigated in young and old Danish twins and was estimated that 75–84%, 53–55%, and 48–50% were attributed to genetic factor, respectively, showing that there is a strong genetic component in the etiology of these traits [54]. As a result, we would expect to find genetic loci associated with these traits through non-hypothesis-driven GWAS, and see new loci, which we would not have known to be implemented in these traits. As GWAS for type 2 diabetes have been successful in identifying many susceptibility loci (please see the section described above), so has been the case for insulin secretion and action. There are many kinds of metabolic traits such as lipid, adiponectin, and leptin levels that play an important role in the metabolism of type 2 diabetes. Here we will focus on genetic loci reported for fasting glycemic traits, including fasting glucose and insulin, proinsulin, and hemoglobin A1c (HbA1c).

GWAS of Common Variants for Glycemic Traits

European Studies

Before the advent of the GWAS era, a few loci were demonstrated to be influencing fasting glucose level in healthy individuals. Using candidate gene approach, association studies identified variants in three genes, *GCK*, *G6PC2*, and *GCKR* [55–58], unequivocally implemented in fasting glucose level. The first GWAS to report genetic loci for diabetes-related quantitative traits was conducted on HbA1c level. Pare et al. evaluated 337,343 SNPs in 14,618 nondiabetic women of Caucasian ancestry in the Women's Genome Health Study [59]. In addition to confirming the HbA1c association at *GCK* and *G6PC2*, they identified a novel locus at *HK1*. Another locus, *SLC30A8*, which was known for its association with T2D reached border-line genome-wide significance ($p = 9.8 \times 10^{-8}$).

The Meta-Analyses of Glucose and Insulin traits Consortium (MAGIC) investigators undertook a series of GWAS on fasting glycemic traits in nondiabetic individuals and succeeded in identifying several genetic loci (Fig. 4.4). By 2011, their efforts led to the discovery of 16 loci for fasting glucose level (known *G6PC2*, *MTNR1B*, *GCK*, *GCKR*, *SLC30A8*, *TCF7L2*; recently reported *DGKB-TMEM195*; novel *ADCY5*, *MADD*, *ADRA2A*, *CRY2*, *FADS1*, *GLIS3*, *SLC2A2*, *PROX1*, and

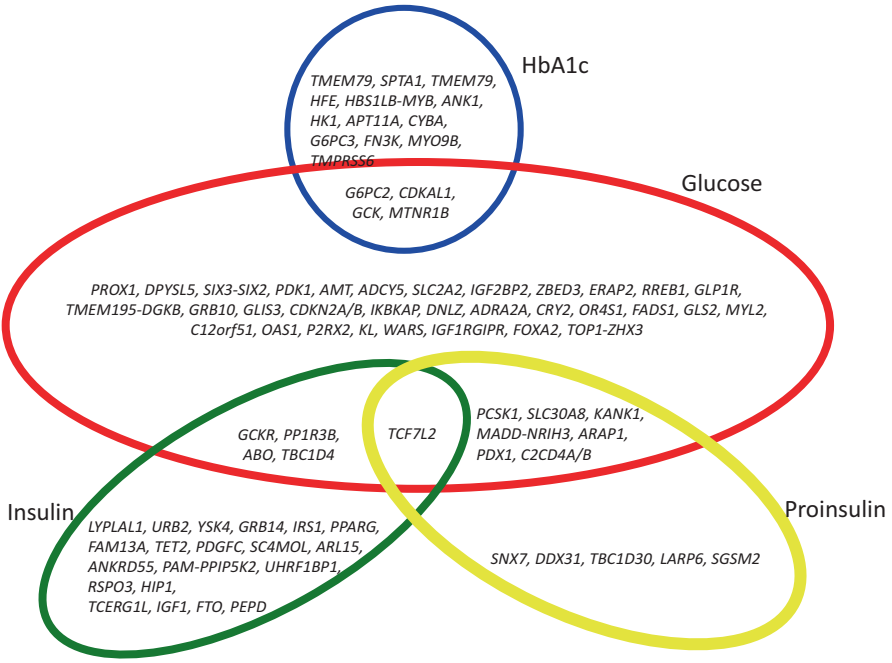

Fig. 4.4 Schematic view of the >80 established loci for fasting glycemic traits, including fasting glucose, insulin, proinsulin, and HbA1c

C2CD4B), 2 for fasting insulin level/HOAM-IR (known *GCKR* and novel *IGF1*), 5 for postoral glucose tolerance test (OGTT) (*GIPR*, *GCKR*, *ADCY5*, *TCF7L2*, *C2CD4A/B*), 10 for proinsulin level (*MADD*, *SLC30A8*, *TCF7L2*, *C2CD4A/B*, *PCSK1*, *ARAP1*, *LARP6*, *SGSM2*; body mass index (BMI) adjusted locus *SNX7*, women-specific locus *DDX31*), and 10 for HbA1c (known *HK1*, *MTNR1B*, *GCK*, *G6PC2*; novel *SPTA1*, *FNK3*, *HFE*, *TMPRSS6*, *ANK1*, *APT11A/TUBGCP3*) [60–63]. This brought the number of loci associated with one or more glycemic traits to 31. These studies highlighted several important biological pathways involved in glucose and insulin metabolism, such as signal transduction, cell proliferation, development, glucose-sense, and circadian regulation. It also demonstrated that on one hand, studying genetics of glycemic trait can help identify T2D risk loci but, on the other hand, that not all loci associated with glycemic traits in healthy population (with glucose level in the "physiological" range) affect the risk of T2D (with glucose level in the "pathological" range), showing that there are un-overlapping mechanisms between fasting glucose elevation and development of T2D.

MAGIC investigators extended their effort by increasing the sample size for discovery GWAS from 46,186 to 133,010 nondiabetic participants and incorporating Illumina CardioMetabochip, a custom iSELECT array of ~200 k SNPs that covers putative association signals for a wide range of cardiometabolic traits and fine-maps established loci [64]. This approach identified 41 novel loci associated with glyce-

mic traits, raising the number of loci associated with fasting glucose level to 36, fasting insulin to 19 and 2 h postprandial glucose (2hGlu) to 9 (Fig. 4.4). The large increase in the number of insulin-associated loci (from 2 to 19) was partly owing to the incorporation of analyses with and without adjustment for BMI [64]. The authors speculated that because BMI explained more of the variance in fasting insulin level than in fasting glucose (R^2 32.6% vs. 8.6%), BMI adjustment provided more opportunity to detect true genetic associations for fasting insulin level by removing the variance in insulin level influenced by BMI. These loci affecting fasting insulin concentration showed association with lipid levels and fat distribution, suggesting impact on insulin resistance. Of the total 53 glycemic loci, 33 were also associated with increased risk of T2D ($q < 0.05$). Although the overlapping loci between glycemic traits and T2D were increased, the overlap was incomplete and many glycemic loci had no discernible effect on T2D (Fig. 4.5) [64].

From a similar point of view with the BMI adjusted analysis undertaken by MAGIC investigators, Manning et al. implemented a joint meta-analysis approach to test associations with fasting insulin and glucose concentration accounting for variant by BMI interaction on a genome-wide scale [65]. Six previously unknown loci associated with fasting insulin at genome-wide significance were identified (*COBLL1-GRB14*, *IRS1*, *PPP1R3B*, *PDGFC*, *LYPLAL1*, and *UHRF1BP1*).

To characterize the known 37 T2D loci and examine the relationship with indices of proinsulin processing, insulin secretion, and insulin sensitivity, MAGIC investigators combined data on both basal and dynamic measures to perform cluster analysis [45]. This analysis highlighted clusters characterized by (i) primary effects on insulin sensitivity (*PPARG*, *KLF14*, *IRS1*, *GCKR*), (ii) reduced insulin secretion and fasting hyperglycemia (*MTNR1B*, *GCK*), (iii) defects in insulin processing (*ARAP1*), (iv) influence on insulin processing and secretion without a detectable change in fasting glucose level (*TCF7L2*, *SLC30A8*, *HHEX/IDE*, *CDKAL1*, *CDKN2A/B*), and (v) unclassified (20 loci).

Studies Conducted in Non-European Population

GWAS on glycemic traits in non-European population was conducted around the world. In 2011, a large-scale GWAS meta-analysis on metabolic traits was conducted in East Asian population, identifying one novel locus for fasting glucose at *SIX2-SIX3* [66]. GWAS in African Americans identified novel loci for insulin and insulin resistance assessed by Homeostasis Model Assessment for Insulin Resistance (HOMA-IR) at *SC4NOL* and *TCERG1L* [67]. More recent GWAS in East Asians detected several novel loci for glycemic traits: *C12orf51*, *PDK1-RAPGEF4*, *KANK1*, *IRS1* for fasting glucose; *MYL2*, *C12orf51*, *OAS1* for 1-2hGlu; *TMEM79*, *HBS1L/MYB*, *MYO98*, *CYBA* for HbA1c [68–70]. Among these novel loci, *IRS1* and *C12orf51* were associated with T2D [38, 71]. GWAS in an isolated Inuit population in Greenland has been successful in identifying a common variant in *TBC1D4* associated with higher 2hGlu, 2 h-insulin, 2 h-C-peptide, and reduced insulin sensitivity index [72]. This variant was common (minor allele frequency (MAF) 17%) in

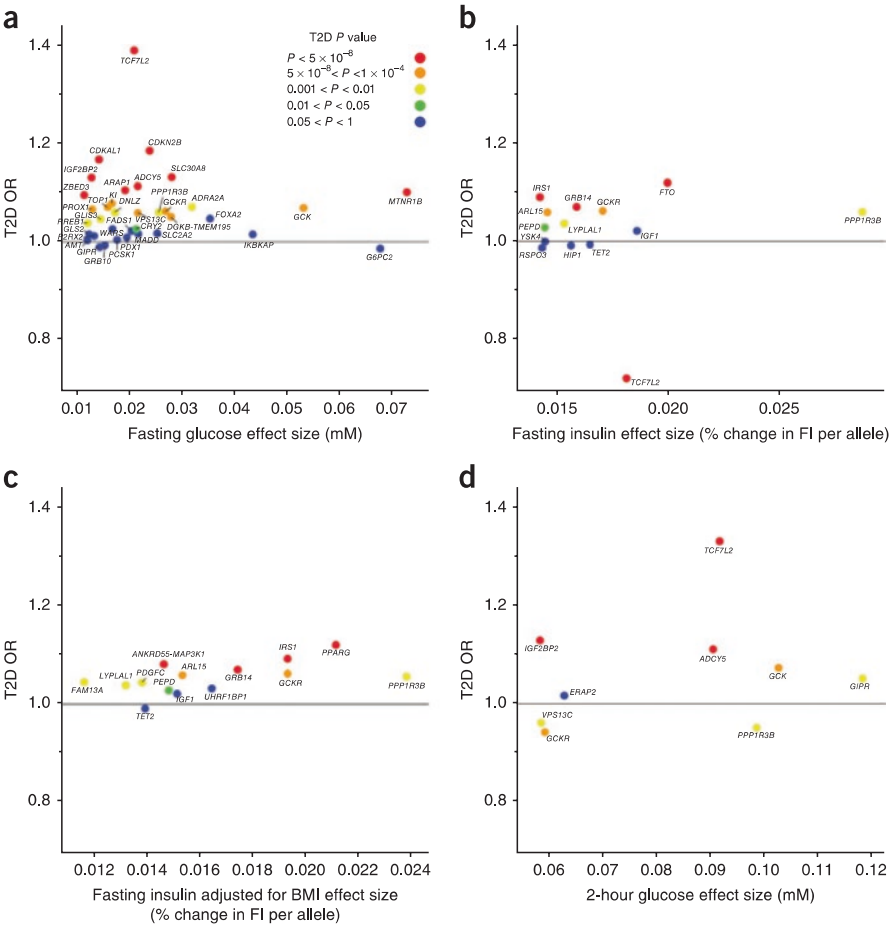

Fig. 4.5 Per-allele β coefficients for glucose and insulin concentrations versus ORs for T2D (reproduced from Scott et al. [64]). (**a**) Fasting glucose concentration versus type 2 diabetes (T2D). (**b**) Fasting insulin (FI) concentration versus T2D. (**c**) Fasting insulin concentration adjusted for body mass index (BMI) versus T2D. (**d**) 2-hour glucose versus T2D. Each locus is color-coded according to the strength of association with T2D as indicated in (**a**)

Greenlandic population, but almost absent in any other population. Homozygous carriers of this *TBC1D4* variant had unprecedentedly high risk of T2D (OR = 10.3).

Exome-Wide Association Analyses for Glycemic Traits

GWAS for fasting glucose and fasting insulin have identified several common variant loci associated with the traits. However, lead SNPs at GWAS loci have relatively modest effect and explain only a small portion of the variance (4.8% and 1.2%,

respectively) [73]. The Illumina HumanExome Beadchip array, a custom array, was designed to facilitate large-scale genotyping of ~250 k mostly rare (MAF <0.5%) and low-frequency (MAF 0.5–5%) protein altering variants selected from sequenced exomes and genomes of ~12,000 individuals. Analyses using this Exomechip have enabled not only to identify novel loci for glycemic traits, but also to clarify the effector transcripts through which the association signals are exerting their effect.

The first report of exome-wide analysis revealed novel loci for low-frequency variants associated with proinsulin level. Low-frequency missense variants in *KANK1* (Arg667His, MAF 2.9%) and *TBC1D30* (Arg279Cys, MAF 2.0%) were associated with proinsulin level. Missense variants in *PAM* (Asp563Gly, MAF 5.3%) and the neighboring *PPIP5K2* (Ser1228Gly, MAF 5.3%) were associated with insulinogenic index [74]. These two missense variants are significantly associated with T2D and are indistinguishable [75, 76]. Exome array analysis identified two low-frequency missense variants in known GWAS signal for fasting proinsulin concentration, which were independent of the known GWAS SNPs. One was Arg766X (MAF 3.7%) in *MADD* and the other was Val996Ile (MAF 1.4%) in *SGSM2*, demonstrating that these two genes were the likely effector transcripts at these loci [74]. Nominal $p < 4.46 \times 10^{-8}$ was used as statistical significance in this analysis, correcting for the number of tests (number of phenotypes multiplied by number of variants tested) conducted [74].

MADD locus was initially identified through GWAS for proinsulin and resides in a region of long-range linkage disequilibrium (LD) that extends >1 Mb in Europeans. Cornes et al. performed targeting deep sequencing of this 11p11.2 locus, encompassing *MADD*, *ACP2*, *NR1H3*, *MYBPC3*, and *SPI1*, and conducted association analysis for fasting glucose and insulin concentration using gene-based test (sequence kernel association test (SKAT)) [77]. SKAT is a useful approach to aggregate low-frequency exonic variants to test against phenotype of interest. Gene-based test at 11p11.2 locus demonstrated that 53 rare variants at NRH13 was jointly associated with fasting insulin, suggesting the existence of >2 independent signals at this locus.

Two other exome-array based analyses for fasting glucose and insulin concentration were reported at the same time. Both studies identified a low-frequency missense variant Ala316Thr (MAF 1.5%) at a novel locus *GLP1R* associated with fasting glucose [73, 78]. The glucose-raising allele of Ala316Thr was associated with lower early insulin secretion, higher 2hGlu concentration and risk of T2D [73]. Multiple low-frequency missense variants at *G6PC2/ABCB11*, a locus known for its strong association with fasting glucose, were reported in both studies. His 177Tyr, Tyr207Ser, Val219Leu (MAF 0.3%, 0.6%, 45.3%, respectively) in *G6PC2* had influence on fasting glucose independently of each other as well as of the known noncoding GWAS common signal [73]. In vitro experiments showed that these missense variants were responsible for the loss of *G6PC2* function through proteosomal degradation, and leads to a reduction in fasting glucose level in human [73]. Gene-based SKAT test demonstrated significant association between *G6PC2* and fasting glucose level [78]. The two studies used study-wide significance based on the number of variants, genes, and tests performed. For example, one of the studies

used $p < 3 \times 10^{-7}$ as significance threshold for single variant analysis and $p < 1.6 \times 10^{-6}$ for gene-based analysis.

Custom Exomechip array contains a certain proportion of noncoding common variants, including known GWAS lead SNPs, in order to facilitate conditional analyses to test evidence for multiple distinct signals at a locus. As a consequence, Exomechip analysis has led to the discovery of several novel loci for glycemic traits with common variants. Exomechip analysis identified additional loci at *GPSM1* and *HNF1A* for Insulinogenic index [74], *ABO* for insulin action (disposition index) and fasting glucose [73, 74], and *URB2* for fasting insulin level [73].

Currently, we are in an exciting time for the discovery of many genetic loci associated with T2D-related quantitative phenotypes. We have summarized >80 loci that have influence on fasting glycemic traits, including fasting glucose, insulin, proinsulin, and HbA1c level (Fig. 4.4). Concurrent approaches using GWAS and Exome array-based analyses have compensatory features to detect these loci. GWAS is widely performed and enables to combine a large number of samples in the meta-analysis. To date, GWAS meta-analysis for fasting glycemic traits are reported on data imputed up to the HapMap reference panel, but ongoing effort to use the latest reference panel for imputation provided by the 1000 Genomes Project will give a better coverage across the low-frequency allele spectrum and is expected to yield many more novel loci for fasting glycemic traits. For exome array-based approach, though it may have limited ability to investigate very rare variants compared to exome sequencing, it is still a cost-effective way and can be more easily performed. Importantly, we have seen proof of principle that exome array genotyping is a powerful way to detect low-frequency variant associations and to enable fine-mapping of the association loci to identify functional variants and effector transcripts through which the association is mediated. The use of these two wheels of analyses is expected to help deciphering the complex picture of the genetics of fasting glycemic traits and its relation with T2D.

4.3 GWAS for Diabetic Nephropathy or Diabetic Kidney Diseases

Diabetic nephropathy is a leading cause of end-stage renal disease (ESRD) in Western countries and Japan. The rising incidence of diabetic nephropathy, especially among patients with type 2 diabetes, is a serious worldwide concern in terms of both poor prognosis and medical costs. Up to now, strict glycemic and/or blood pressure control, protein restriction, or combination of these treatment have been shown to be effective for the prevention of the progression of diabetic nephropathy as well as for reducing cardiovascular events in patients with diabetic nephropathy [79–82]. Furthermore, remission and/or regression of microalbuminuria have also been reported [83–85], and thus the prognosis of subjects with diabetic nephropathy has been significantly improved during the last decade. However, still considerable numbers of subjects were suffered with diabetic nephropathy.

The pathogenesis of diabetic nephropathy appears to be multifactorial, and several environmental and/or genetic factors might be responsible for the development and progression of the disease [86], but precise mechanisms have not been elucidated yet. It has been reported that the cumulative incidence of diabetic retinopathy increased linearly according to the duration of diabetes, whereas the occurrence of nephropathy was almost none after 20–25 years of diabetes duration, and only modest number of individuals with diabetes (~30%) developed diabetic nephropathy [87]. Familial clustering of diabetic nephropathy was also reported both in type 1 [88] and type 2 diabetes [89], From these cumulative evidences, it is suggested that genetic susceptibility plays an important role in the pathogenesis of diabetic nephropathy. Worldwide efforts have been conducted to identify genes conferring susceptibility to diabetic nephropathy, but the efforts by classical approaches, i.e., candidate gene approaches or linkage analyses, have not been successful so far.

GWAS for diabetic nephropathy or diabetic kidney diseases have been performed in European, African American, and Japanese populations. However, the results were not consistent each other, and only a few loci satisfied genome-wide significant level.

GWAS for Diabetic Nephropathy (Diabetic Kidney Disease) in Populations of European Descent

In patients with type 1 diabetes, GWAS for diabetic nephropathy was first conducted by Genetics of Kidneys in Diabetes (GoKinD) study group using 820 cases (284 with proteinuria and 536 with end-stage renal disease) and 885 controls for ~360,000 SNPs, followed by a validation analysis using 1304 participants of the Diabetes Control and Complications Trial (DCCT)/Epidemiology of Diabetes Interventions and Complications (EDIC) study, a long-term, prospective investigation of the development of diabetes- associated complications [90]. Four SNP loci were reported to show suggestive associations through the GWAS, rs10868025 near *FRMD3* (9q21.32), rs39059 within *CHN2* (7p14.3), rs451041 within *CARS* (11p15.4), and rs1411766/rs1742858 near *MYO16/IRS2* (13q33.3). Among the four loci, association of two loci, *FRMD3* and *CARS*, were validated in the DCCT/EDIC study, although the association did not attain genome-wide significant level. The association of the four loci were further evaluated in 66 extended families of European ancestry, the Joslin Study of Genetics of Nephropathy in Type 2 Diabetes Family Collection, the results indicated that *FRMD3* locus showed evidence of association with diabetic nephropathy (advanced diabetic nephropathy or advanced diabetic nephropathy plus high microalbuminuria) or with albuminuria (log transformed albumin to creatinine ratio) [91]. The association of *FRMD3* locus with diabetic end-stage renal disease was observed in African American patients with type 2 diabetes lacking two *MYH9* E1 risk haplotypes, which was well-known strong risk for nondiabetic kidney diseases in African Americans [92]. In Japanese

patients with type 2 diabetes, rs1411766 at ch. 13q33.3 was associated with diabetic nephropathy, and the association attained a genome-wide significant level after integration of two data, Japanese type 2 diabetes and Caucasian type 1 diabetes, by a meta-analysis [93].

In 2012, a meta-analysis of diabetic nephropathy for patients with type 1 diabetes in populations of European origin was performed by the Genetics of Nephropathy-an International Effort (GENIE) consortium [94]. The analysis using advanced diabetic nephropathy (4409 overt proteinuria or end-stage renal disease) and 6691 controls identified that rs7588550 within *ERBB4* showed suggestive evidence of associated with diabetic nephropathy. In a subsequent sub-analysis for end-stage renal disease, 1786 cases and 8718 controls including patients with overt proteinuria, two loci, rs7583877 in the *AFF3* (2q11.2) and rs12437854 in *RGMA*/*MCTP2* locus (15q26.1), were associated with ESRD with a genome-wide significant level. However, these associations were not validated in independent case-control studies [95]. Genotype imputation using directly genotyped data and linkage disequilibrium data in 1000 genomes database for patients with type 1 diabetes was performed in the Finnish Diabetic Nephropathy (FinnDiane) study. The analysis for 11,133,962 tested SNPs and subsequent first and second stage analyses, comprising of 2142 cases and 2494 controls, identified rs1326934 within the *SORBS1* as top signal for susceptibility to diabetic nephropathy, but the association did not reach a genome-wide significant level [96]. Sex stratified GWAS for diabetic nephropathy in European patients with type 1 diabetes identified rs4972593 on chromosome 2q31.1 as susceptibility to ESRD only in women, but not in men, and the results were replicated in independent replication studies [97].

In a GWAS meta-analysis for quantitative traits analysis regarding kidney functions in 54,450 individuals, variants within *CUBN* showed genome-wide significant association with urinary albumin-to-creatinine ratio (UACR), and it was also shown that an effect size on logarithmic UACR values was fourfold larger among 5825 individuals with diabetes (0.19 log[mg/g], $p = 2.0 \times 10^{-5}$) compared with 46,061 individuals without diabetes (0.045 log[mg/g], $p = 8.7 \times 10^{-6}$; $p = 8.2 \times 10^{-4}$ for difference) [98]. In this analysis, rs649529 at *RAB38*/*CTSC* locus on chromosome 11q14 and rs13427836 in *HS6ST1* on chromosome 2q21 were associated with UACR only in patients with diabetes.

GWAS for Diabetic Nephropathy in the Japanese Population

In order to identify genes conferring susceptibility to diabetic nephropathy, we have performed a GWAS for diabetic nephropathy using 188 Japanese patients with type 2 diabetes [99, 100]. We commenced an association study using SNPs from a Japanese SNP database [101, 102] established prior to the HapMap database. We screened approximately 100,000 gene-based SNP loci, and the genotype and allele frequencies of 94 nephropathy cases, defined as patients with overt proteinuria or ESRD were compared with those of 94 controls defined as patients with

normoalbuminuria and diabetic retinopathy. Approximately 80,000 SNP loci were successfully genotyped, and 1615 SNP loci with $p < 0.01$ were selected, and forwarded to the validation study. These 1615 SNP loci were analyzed further in a greater number of subjects to clarify their statistical significance. As a result, several SNP loci, including the *SLC12A3* locus [103], *ELMO1* locus [104], and *NCALD* locus [105] were found to be associated with diabetic nephropathy.

Solute Carrier Family 12, Member 3 (*SLC12A3*)

The *SLC12A3*, at chromosome 16q13, encodes a thiazide-sensitive Na + -Cl- cotransporter that mediates reabsorption of Na^+ and Cl^- at the renal distal convoluted tubule; this molecule is the target of thiazide diuretics. Mutations in *SLC12A3* are responsible for Gitelman syndrome [106], which is inherited as an autosomal recessive trait characterized by hypokalemia, metabolic alkalosis, hypomagnesemia, hypocalciuria, and volume depletion. A coding SNP in exon 23 of the *SLC12A3* (rs11643718, +78 G to A: Arg913Gln) was shown to be associated with diabetic nephropathy ($p = 0.00002$, odds ratio 2.53 [95% CI 1.64–3.90]). The results implicated that substitution of Arg913 to Gln in the *SLC12A3* might reduce the risk to develop diabetic nephropathy. The association of rs11643718 with diabetic nephropathy was replicated in independent case-control studies, including Japanese [107], South Asian [108], and Malaysian subjects [109] with type 2 diabetes. Rs11643718 was associated with end-stage renal disease in Korean patients with type 2 diabetes, but direction of effect was opposite to that in the original report [110]. Rs11643718 did not show significant effect in Caucasian patients with type 2 diabetes (Table 4.1) [111].

Engulfment and Cell Motility 1 (*ELMO1*)

We identified that the *ELMO1* was a likely candidate for conferring susceptibility to diabetic nephropathy (rs741301, intron 18 + 9170, GG vs. GA+AA, $\chi^2 = 19.9$, $p = 0.000008$, odds ratio: 2.67, 95%CI 1.71–4.16) [104]. The association of *ELMO1* locus with diabetic nephropathy was observed also in African American patients

Table 4.1 Effect of non-synonymous SNP (rs11643718, Arg913Gln) within the *SLC12A3 with diabetic nephropathy*

Ethnicity	n Case: control	Odds ratio	95% CI	*p* value	Allele frequency
Japanese	716: 543	0.443	0.309–0.636	0.00002	0.076
Japanese	71: 193	0.09	0.01–0.92	0.043	0.143
Malaysian	124: 259	0.547	0.308–0.973	0.038	0.112
South Indian	583: 601	0.658	0.459–0.943	0.020	0.101
Korean	175: 183	2.295	1.573–3.239	0.003	0.055
European	277: 164	1.213	0.775–1.897	0.397	0.098

with type 2 diabetes [112], Caucasian patients with type 1 diabetes [113], South Indian patients with type 2 diabetes [108], Chinese patients with type 2 diabetes [114], and American Indian patients with type 2 diabetes [115], although associated SNPs or direction of the effects varied among the individual studies. The *ELMO1* gene, on chromosome 7p14, is a known mammalian homologue of the *C. elegans* gene, ced-12, which is required for engulfment of dying cells and for cell migration [116]. *ELMO1* has also been reported to cooperate with CrkII and Dock180, which are homologues of *C. elegans* ced-2, ced-5, respectively, to promote phagocytosis and cell shape changes [116, 117]. However, until then no evidence had been reported to suggest a role for this gene in the pathogenesis of diabetic nephropathy. By in situ hybridization using the kidney of normal and diabetic mice, we found that *ELMO1* expression was weakly detectable mainly in tubular and glomerular epithelial cells in normal mouse kidney, and was clearly elevated in the kidney of diabetic mice. Subsequent in vitro analysis revealed that *ELMO1* expression was elevated in cells cultured under high glucose conditions (25 mM) compared to cells cultured under normal glucose conditions (5.5 mM). Furthermore, we identified that the expression of extracellular matrix protein genes, such as Type 1 collagen and fibronectin, were increased in cells that over-expressing *ELMO1*, whereas the expression of MMPs (matrix metalloproteinase) was decreased [104, 118]. Therefore, it is suggested that persistent excess of *ELMO1* in subjects with disease susceptibility allele leads to the overaccumulation of extracellular matrix proteins and to the development and progression of diabetic glomerulosclerosis. It has been also reported that excess of *Elmo1* accelerated the progression of renal injury in mouse model of diabetes, whereas *Elmo1* depletion protected the renal injury in these mice [119]. In contrast, experiments using zebrafish suggested that *elmo1* had a protective role in the progression of renal injury under diabetic conditions [120].

The association of *NCALD* locus with diabetic nephropathy was not replicated in an independent population, and the association of above mentioned loci identified in Japanese GWAS for diabetic nephropathy did not attain a genome-wide significant level.

Acetyl-Coenzyme a Carboxylase Beta Gene (*ACACB*)

We extended the previous GWAS for diabetic nephropathy to the SNPs with *p* values between 0.01 and 0.05 and provide evidence that a SNP, rs2268388, within the acetyl-coenzyme A carboxylase beta gene (*ACACB*; MIM: 601557) contributes to an increased prevalence of proteinuria in patients with type 2 diabetes across different ethnic populations [121].

The frequency of the T allele of rs2268388 was consistently higher among patients with type 2 diabetes with proteinuria (combined meta-analysis gave a *p* value of 5.35×10^{-8} in the Japanese, 2.3×10^{-9} for all populations). The association of rs2268388 was replicated in patients with type 2 diabetes in different ethnic groups, including Han Chinese [122] and Indians [123].

Expression of *ACACB* was observed in adipose tissue, heart, and skeletal muscle, and, to a lesser extent, in the kidney. The results of in situ hybridization with normal mouse kidney revealed that *Acacb* was localized to glomerular epithelial cells and tubular epithelial cells. We also observed the expression of *ACACB* in cultured human renal proximal tubular epithelial cells (hRPTECs). In cultured hRPTECs, a 29-bp DNA fragments containing the SNP region had significant enhancer activity, and fragments corresponding to the disease susceptibility allele had stronger enhancer activity than those for the major allele [121].

The quantitative real-time PCR (polymerase chain reaction) using glomeruli isolated from these mice revealed that the expression of *Acacb* was increased in the glomeruli of diabetic db/db mice compared to those of control mice [124]. Furthermore, overexpression of ACACB in hRPTECs resulted in remarkable increase of the expressions of genes encoding pro-inflammatory cytokines, including IL-6, CXCL1, CXCL2, CXCL5, and CXCL6.

Combining these results with the finding in the genetic study, it is suggested that ACACB contributes to conferring susceptibility to diabetic nephropathy at least in part, via the effects of the pro-inflammatory cytokines, and the ACACB-IL-6 or ACACB-CXCLs systems may be considered as new pathways for the development and progression of diabetic nephropathy.

GWAS for Diabetic Nephropathy in Other Ethnic Groups

An African American GWAS for diabetic nephropathy evaluated 965 ESRD patients with type 2 diabetes and control individuals without type 2 diabetes or kidney disease for 832,357 SNP loci, and in addition to *MYH9-APOL1* locus, which is already known susceptibility to nondiabetic kidney diseases, several loci, *RPS12*, *LIMK2*, *SFI1*, were associated with ESRD in patients with type 2 diabetes, although any association did not attain a genome-wide significant level [125].

Results of multiethnic GWAS meta-analysis, including African American, American Indian, European, and Mexican, identified significant association of rs955333 at 6q25.2 with diabetic nephropathy [126].

Susceptibility loci for diabetic nephropathy or diabetic kidney disease with genome-wide significant association are listed in Table 4.2.

4.4 Future Perspective

After the human genome (sequencing) project was completed [127, 128], a large body of information on the human genome has been accumulated [129]. Simultaneously, high-throughput genotyping technologies as well as statistical methods and/or tools for handling innumerable datasets have been developed. Then, genome-wide association studies for investigating genes associated with disease

Table 4.2 Genetic loci associated with diabetic nephropathy

Ethnicity	Nearest gene	Chromosome	Phenotype	Type of diabetes	Method	Replication
Japanese	ACACB	12q24.11	Overt proteinuria	Type 2	GWAS	Yes
European	AFF3 RGMA-MCTP2	2q11.2	End-stage renal disease	Type 1	GWAS	No
European	rs4972593	2q31.1	End-stage renal disease (women only)	Type 1	GWAS	Yes
European	GLRA3	4q34.1	Urinary albumin excretion rate	Type 1	GWAS	No
European	EPO	7q22.1	End-stage renal disease + proliferative retinopathy	Type 1	Candidate gene approach	No
Multiethnic	rs955333	6q25.2	Overt proteinuria + end-stage renal disease	Type 2	GWAS	No
European	CUBN	10q13	Urinary albumin excretion rate	Type 2	GWAS	Yes
European	SLC19A3	2q36.3	Advanced retinopathy + end-stage renal disease	Type 1	Candidate gene approach	No

susceptibility across the entire human genome have been facilitated, and more than 2000 loci susceptible to various diseases or traits have been discovered [130].

Although this is excellent progress, it has also been recognized that the information obtained from GWAS is still insufficient for clinical application. The focus of ongoing research efforts includes detailed functional characterization of the identified T2D susceptibility variants and the search for missing heritability.

Certain modifications of the GWAS study design will be necessary to uncover the missing heritability. Much larger intra- or trans-ethnic sample sizes will be required to increase the power to detect true signals, which may be conducted in meta-analyses. Examining populations of non-European descent is likely to identify additional T2D loci, and this should be performed more vigorously. Association analyses of low frequency variants for T2D are an additional option. Additionally, it has been shown that the study using small and historically isolated populations may have advantages to identify novel susceptibility to the disease [72]. In this report, GWAS for glycemic traits using a relatively small number of Greenlandic inuits (~2500) identified the nonsense variants in the *TBC1D4*, which had a striking effect on susceptibility to T2D (OR = ~10). Since similar success was reported to identify novel missense variants within *CREBRF* associated with obesity in the Samoan population [131], unique variants with a large effect size are conserved in geneti-

cally homogeneous populations, and GWAS in these populations, even if its sample size is not so large, are useful to identify novel susceptibility to T2D.

Characterizing disease biology is another relevant goal of genetic studies for T2D, which has been behind compared with GWAS discovery. Recent biological and clinical studies have suggested possible means to increase the translational use of genetic findings through convergence on common resources and workflows, regarding comprehensive gene expression data, epigenomics, PPI networks, and information of cellular and animal models [30, 53, 132]. In order to exploit these trends to advance biological understanding of T2D, it is urgent needs of establishment and effective utilization of publicly available databases including genetic data with large-scale sample size with rich phenotype information, epigenomic and transcriptomic data for diverse tissue types, and comprehensive biological data resource from cellular and animal models.

References

1. IDF diabetes atlas ver 7. (2015) http://www.diabetesatlas.org/
2. Poulsen P, Kyvik KO, Vaag A, Beck-Nielsen H (1999) Heritability of type II (non-insulin-dependent) diabetes mellitus and abnormal glucose tolerance – a population-based twin study. Diabetologia 42:139–145
3. Groop L, Forsblom C, Lehtovirta M, Tuomi T, Karanko S, Nissén M, Ehrnström BO, Forsén B, Isomaa B, Snickars B, Taskinen MR (1996) Metabolic consequences of a family history of NIDDM (the Botnia study): evidence for sex-specific parental effects. Diabetes 45:1585–1593
4. Hemminki K, Li X, Sundquist K, Sundquist J (2010) Familial risks for type 2 diabetes in Sweden. Diabetes Care 33:293–297
5. Almgren P, Lehtovirta M, Isomaa B, Sarelin L, Taskinen MR, Lyssenko V, Tuomi T, Groop L, Botnia Study Group (2011) Heritability and familiality of type 2 diabetes and related quantitative traits in the Botnia study. Diabetologia 54:2811–2819
6. Imamura M, Maeda S (2011) Genetics of type 2 diabetes: the GWAS era and future perspectives. Endocr J 58:723–739
7. Horikawa Y, Oda N, Cox NJ, Li X, Orho-Melander M, Hara M, Hinokio Y, Lindner TH, Mashima H, Schwarz PE, del Bosque-Plata L, Horikawa Y, Oda Y, Yoshiuchi I, Colilla S, Polonsky KS, Wei S, Concannon P, Iwasaki N, Schulze J, Baier LJ, Bogardus C, Groop L, Boerwinkle E, Hanis CL, Bell GI (2000) Genetic variation in the gene encoding calpain-10 is associated with type 2 diabetes mellitus. Nat Genet 26:163–175
8. Reynisdottir I, Thorleifsson G, Benediktsson R, Sigurdsson G, Emilsson V, Einarsdottir AS, Hjorleifsdottir EE, Orlygsdottir GT, Bjornsdottir GT, Saemundsdottir J, Halldorsson S, Hrafnkelsdottir S, Sigurjonsdottir SB, Steinsdottir S, Martin M, Kochan JP, Rhees BK, Grant SF, Frigge ML, Kong A, Gudnason V, Stefansson K, Gulcher JR (2003) Localization of a susceptibility gene for type 2 diabetes to chromosome 5q34-q35.2. Am J Hum Genet 73:323–325
9. Grant SF, Thorleifsson G, Reynisdottir I, Benediktsson R, Manolescu A, Sainz J, Helgason A, Stefansson H, Emilsson V, Helgadottir A, Styrkarsdottir U, Magnusson KP, Walters GB, Palsdottir E, Jonsdottir T, Gudmundsdottir T, Gylfason A, Saemundsdottir J, Wilensky RL, Reilly MP, Rader DJ, Bagger Y, Christiansen C, Gudnason V, Sigurdsson G, Thorsteinsdottir U, Gulcher JR, Kong A, Stefansson K (2006) Variant of transcription factor 7-like 2 (*TCF7L2*) gene confers risk of type 2 diabetes. Nat Genet 38:320–323

10. Florez JC, Jablonski KA, Bayley N, Pollin TI, de Bakker PI, Shuldiner AR, Knowler WC, Nathan DM, Altshuler D, Diabetes Prevention Program Research Group (2006) TCF7L2 polymorphisms and progression to diabetes in the diabetes prevention program. N Engl J Med 355:241–250
11. Groves CJ, Zeggini E, Minton J, Frayling TM, Weedon MN, Rayner NW, Hitman GA, Walker M, Wiltshire S, Hattersley AT, McCarthy MI (2006) Association analysis of 6,736 U.K. subjects provides replication and confirms *TCF7L2* as a type 2 diabetes susceptibility gene with a substantial effect on individual risk. Diabetes 55:2640–2644
12. Zhang C, Qi L, Hunter DJ, Meigs JB, Manson JE, van Dam RM, Hu FB (2006) Variant of transcription factor 7-like 2 (*TCF7L2*) gene and the risk of type 2 diabetes in large cohorts of U.S. women and men. Diabetes 55:2645–2648
13. Scott LJ, Bonnycastle LL, Willer CJ, Sprau AG, Jackson AU, Narisu N, Duren WL, Chines PS, Stringham HM, Erdos MR, Valle TT, Tuomilehto J, Bergman RN, Mohlke KL, Collins FS, Boehnke M (2006) Association of transcription factor 7-like 2 (*TCF7L2*) variants with type 2 diabetes in a Finnish sample. Diabetes 55:2649–2653
14. Damcott CM, Pollin TI, Reinhart LJ, Ott SH, Shen H, Silver KD, Mitchell BD, Shuldiner AR (2006) Polymorphisms in the transcription factor 7-like 2 (*TCF7L2*) gene are associated with type 2 diabetes in the Amish: replication and evidence for a role in both insulin secretion and insulin resistance. Diabetes 55:2654–2659
15. Saxena R, Gianniny L, Burtt NP, Lyssenko V, Giuducci C, Sjögren M, Florez JC, Almgren P, Isomaa B, Orho-Melander M, Lindblad U, Daly MJ, Tuomi T, Hirschhorn JN, Ardlie KG, Groop LC, Altshuler D (2006) Common single nucleotide polymorphisms in *TCF7L2* are reproducibly associated with type 2 diabetes and reduce the insulin response to glucose in nondiabetic individuals. Diabetes 55:2890–2895
16. Cauchi S, Meyre D, Dina C, Choquet H, Samson C, Gallina S, Balkau B, Charpentier G, Pattou F, Stetsyuk V, Scharfmann R, Staels B, Frühbeck G, Froguel P (2006) Transcription factor *TCF7L2* genetic study in the French population: expression in human beta-cells and adipose tissue and strong association with type 2 diabetes. Diabetes 55:2903–2908
17. Hayashi T, Iwamoto Y, Kaku K, Hirose H, Maeda S (2007) Replication study for the association of *TCF7L2* with susceptibility to type 2 diabetes in a Japanese population. Diabetologia 50:980–984
18. Horikoshi M, Hara K, Ito C, Nagai R, Froguel P, Kadowaki T (2007) A genetic variation of the transcription factor 7-like 2 gene is associated with risk of type 2 diabetes in the Japanese population. Diabetologia 50:747–751
19. Altshuler D, Hirschhorn JN, Klannemark M, Lindgren CM, Vohl MC, Nemesh J, Lane CR, Schaffner SF, Bolk S, Brewer C, Tuomi T, Gaudet D, Hudson TJ, Daly M, Groop L, Lander ES (2000) The common PPARgamma Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. Nat Genet 26:76–80
20. Gloyn AL, Weedon MN, Owen KR, Turner MJ, Knight BA, Hitman G, Walker M, Levy JC, Sampson M, Halford S, McCarthy MI, Hattersley AT, Frayling TM (2003) Large-scale association studies of variants in genes encoding the pancreatic beta-cell KATP channel subunits Kir6.2 (*KCNJ11*) and SUR1 (*ABCC8*) confirm that the *KCNJ11* E23K variant is associated with type 2 diabetes. Diabetes 52:568–572
21. Sladek R, Rocheleau G, Rung J, Dina C, Shen L, Serre D, Boutin P, Vincent D, Belisle A, Hadjadj S, Balkau B, Heude B, Charpentier G, Hudson TJ, Montpetit A, Pshezhetsky AV, Prentki M, Posner BI, Balding DJ, Meyre D, Polychronakos C, Froguel P (2007) A genome-wide association study identifies novel risk loci for type 2 diabetes. Nature 445:881–885
22. Steinthorsdottir V, Thorleifsson G, Reynisdottir I, Benediktsson R, Jonsdottir T, Walters GB, Styrkarsdottir U, Gretarsdottir S, Emilsson V, Ghosh S, Baker A, Snorradottir S, Bjarnason H, Ng MC, Hansen T, Bagger Y, Wilensky RL, Reilly MP, Adeyemo A, Chen Y, Zhou J, Gudnason V, Chen G, Huang H, Lashley K, Doumatey A, So WY, Ma RC, Andersen G, Borch-Johnsen K, Jorgensen T, van Vliet-Ostaptchouk JV, Hofker MH, Wijmenga C, Christiansen C, Rader DJ, Rotimi C, Gurney M, Chan JC, Pedersen O, Sigurdsson G, Gulcher JR, Thorsteinsdottir

U, Kong A, Stefansson K (2007) A variant in *CDKAL1* influences insulin response and risk of type 2 diabetes. Nat Genet 39:770–775

23. Saxena R, Voight BF, Lyssenko V, Burtt NP, de Bakker PI, Chen H, Roix JJ, Kathiresan S, Hirschhorn JN, Daly MJ, Hughes TE, Groop L, Altshuler D, Almgren P, Florez JC, Meyer J, Ardlie K, Bengtsson Boström K, Isomaa B, Lettre G, Lindblad U, Lyon HN, Melander O, Newton-Cheh C, Nilsson P, Orho-Melander M, Råstam L, Speliotes EK, Taskinen MR, Tuomi T, Guiducci C, Berglund A, Carlson J, Gianniny L, Hackett R, Hall L, Holmkvist J, Laurila E, Sjögren M, Sterner M, Surti A, Svensson M, Svensson M, Tewhey R, Blumenstiel B, Parkin M, Defelice M, Barry R, Brodeur W, Camarata J, Chia N, Fava M, Gibbons J, Handsaker B, Healy C, Nguyen K, Gates C, Sougnez C, Gage D, Nizzari M, Gabriel SB, Chirn GW, Ma Q, Parikh H, Richardson D, Ricke D, Purcell S (2007) Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. Science 316:1331–1336
24. Zeggini E, Weedon MN, Lindgren CM, Frayling TM, Elliott KS, Lango H, Timpson NJ, Perry JR, Rayner NW, Freathy RM, Barrett JC, Shields B, Morris AP, Ellard S, Groves CJ, Harries LW, Marchini JL, Owen KR, Knight B, Cardon LR, Walker M, Hitman GA, Morris AD, Doney AS, Wellcome Trust Case Control Consortium (WTCCC), McCarthy MI, Hattersley AT (2007) Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. Science 316:1336–1341
25. Scott LJ, Mohlke KL, Bonnycastle LL, Willer CJ, Li Y, Duren WL, Erdos MR, Stringham HM, Chines PS, Jackson AU, Prokunina-Olsson L, Ding CJ, Swift AJ, Narisu N, Hu T, Pruim R, Xiao R, Li XY, Conneely KN, Riebow NL, Sprau AG, Tong M, White PP, Hetrick KN, Barnhart MW, Bark CW, Goldstein JL, Watkins L, Xiang F, Saramies J, Buchanan TA, Watanabe RM, Valle TT, Kinnunen L, Abecasis GR, Pugh EW, Doheny KF, Bergman RN, Tuomilehto J, Collins FS, Boehnke M (2007) A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. Science 316:1341–1345
26. Frayling TM, Timpson NJ, Weedon MN, Zeggini E, Freathy RM, Lindgren CM, Perry JR, Elliott KS, Lango H, Rayner NW, Shields B, Harries LW, Barrett JC, Ellard S, Groves CJ, Knight B, Patch AM, Ness AR, Ebrahim S, Lawlor DA, Ring SM, Ben-Shlomo Y, Jarvelin MR, Sovio U, Bennett AJ, Melzer D, Ferrucci L, Loos RJ, Barroso I, Wareham NJ, Karpe F, Owen KR, Cardon LR, Walker M, Hitman GA, Palmer CN, Doney AS, Morris AD, Smith GD, Hattersley AT, McCarthy MI (2007) A common variant in the *FTO* gene is associated with body mass index and predisposes to childhood and adult obesity. Science 316:889–894
27. Zeggini E, Scott LJ, Saxena R, Voight BF, Marchini JL, Hu T, de Bakker PI, Abecasis GR, Almgren P, Andersen G, Ardlie K, Boström KB, Bergman RN, Bonnycastle LL, Borch-Johnsen K, Burtt NP, Chen H, Chines PS, Daly MJ, Deodhar P, Ding CJ, Doney AS, Duren WL, Elliott KS, Erdos MR, Frayling TM, Freathy RM, Gianniny L, Grallert H, Grarup N, Groves CJ, Guiducci C, Hansen T, Herder C, Hitman GA, Hughes TE, Isomaa B, Jackson AU, Jørgensen T, Kong A, Kubalanza K, Kuruvilla FG, Kuusisto J, Langenberg C, Lango H, Lauritzen T, Li Y, Lindgren CM, Lyssenko V, Marvelle AF, Meisinger C, Midthjell K, Mohlke KL, Morken MA, Morris AD, Narisu N, Nilsson P, Owen KR, Palmer CN, Payne F, Perry JR, Pettersen E, Platou C, Prokopenko I, Qi L, Qin L, Rayner NW, Rees M, Roix JJ, Sandbaek A, Shields B, Sjögren M, Steinthorsdottir V, Stringham HM, Swift AJ, Thorleifsson G, Thorsteinsdottir U, Timpson NJ, Tuomi T, Tuomilehto J, Walker M, Watanabe RM, Weedon MN, Willer CJ, Wellcome Trust Case Control Consortium, Illig T, Hveem K, Hu FB, Laakso M, Stefansson K, Pedersen O, Wareham NJ, Barroso I, Hattersley AT, Collins FS, Groop L, McCarthy MI, Boehnke M, Altshuler D (2008) Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. Nat Genet 40:638–645
28. Chan JC, Malik V, Jia W, Kadowaki T, Yajnik CS, Yoon KH, Hu FB (2009) Diabetes in Asia: epidemiology, risk factors, and pathophysiology. JAMA 301:2129–2140
29. Imamura M, Maeda S, Yamauchi T, Hara K, Yasuda K, Morizono T, Takahashi A, Horikoshi M, Nakamura M, Fujita H, Tsunoda T, Kubo M, Watada H, Maegawa H, Okada-Iwabu M, Iwabu M, Shojima N, Ohshige T, Omori S, Iwata M, Hirose H, Kaku K, Ito C, Tanaka Y,

Tobe K, Kashiwagi A, Kawamori R, Kasuga M, Kamatani N, Diabetes Genetics Replication and Meta-analysis (DIAGRAM) Consortium, Nakamura Y, Kadowaki T (2012) A single-nucleotide polymorphism in ANK1 is associated with susceptibility to type 2 diabetes in Japanese populations. Hum Mol Genet 21:3042–3049

30. Imamura M, Takahashi A, Yamauchi T, Hara K, Yasuda K, Grarup N, Zhao W, Wang X, Huerta-Chagoya A, Hu C, Moon S, Long J, Kwak SH, Rasheed A, Saxena R, Ma RC, Okada Y, Iwata M, Hosoe J, Shojima N, Iwasaki M, Fujita H, Suzuki K, Danesh J, Jørgensen T, Jørgensen ME, Witte DR, Brandslund I, Christensen C, Hansen T, Mercader JM, Flannick J, Moreno-Macías H, Burtt NP, Zhang R, Kim YJ, Zheng W, Singh JR, Tam CH, Hirose H, Maegawa H, Ito C, Kaku K, Watada H, Tanaka Y, Tobe K, Kawamori R, Kubo M, Cho YS, Chan JC, Sanghera D, Frossard P, Park KS, Shu XO, Kim BJ, Florez JC, Tusié-Luna T, Jia W, Tai ES, Pedersen O, Saleheen D, Maeda S, Kadowaki T (2016) Genome-wide association studies in the Japanese population identify seven novel loci for type 2 diabetes. Nat Commun 7:10531
31. Unoki H, Takahashi A, Kawaguchi T, Hara K, Horikoshi M, Andersen G, Ng DP, Holmkvist J, Borch-Johnsen K, Jørgensen T, Sandbaek A, Lauritzen T, Hansen T, Nurbaya S, Tsunoda T, Kubo M, Babazono T, Hirose H, Hayashi M, Iwamoto Y, Kashiwagi A, Kaku K, Kawamori R, Tai ES, Pedersen O, Kamatani N, Kadowaki T, Kikkawa R, Nakamura Y, Maeda S (2008) SNPs in *KCNQ1* are associated with susceptibility to type 2 diabetes in East Asian and European populations. Nat Genet 40:1098–1102
32. Yasuda K, Miyake K, Horikawa Y, Hara K, Osawa H, Furuta H, Hirota Y, Mori H, Jonsson A, Sato Y, Yamagata K, Hinokio Y, Wang HY, Tanahashi T, Nakamura N, Oka Y, Iwasaki N, Iwamoto Y, Yamada Y, Seino Y, Maegawa H, Kashiwagi A, Takeda J, Maeda E, Shin HD, Cho YM, Park KS, Lee HK, Ng MC, Ma RC, So WY, Chan JC, Lyssenko V, Tuomi T, Nilsson P, Groop L, Kamatani N, Sekine A, Nakamura Y, Yamamoto K, Yoshida T, Tokunaga K, Itakura M, Makino H, Nanjo K, Kadowaki T, Kasuga M (2008) Variants in *KCNQ1* are associated with susceptibility to type 2 diabetes mellitus. Nat Genet 40:1092–1097
33. Lee YH, Kang ES, Kim SH, Han SJ, Kim CH, Kim HJ, Ahn CW, Cha BS, Nam M, Nam CM, Lee HC (2008) Association between polymorphisms in *SLC30A8*, *HHEX*, *CDKN2A/B*, *IGF2BP2*, *FTO*, *WFS1*, *CDKAL1*, *KCNQ1* and type 2 diabetes in the Korean population. J Hum Genet 53:991–998
34. Tan JT, Nurbaya S, Gardner D, Ye S, Tai ES, Ng DP (2009) Genetic variation in *KCNQ1* associates with fasting glucose and beta-cell function: a study of 3,734 subjects comprising three ethnicities living in Singapore. Diabetes 58:1445–1449
35. Hu C, Wang C, Zhang R, Ma X, Wang J, Lu J, Qin W, Bao Y, Xiang K, Jia W (2009) Variations in *KCNQ1* are associated with type 2 diabetes and beta cell function in a Chinese population. Diabetologia 52:1322–1325
36. Liu Y, Zhou DZ, Zhang D, Chen Z, Zhao T, Zhang Z, Ning M, Hu X, Yang YF, Zhang ZF, Yu L, He L, Xu H (2009) Variants in *KCNQ1* are associated with susceptibility to type 2 diabetes in the population of mainland China. Diabetologia 52:1315–1321
37. Yamauchi T, Hara K, Maeda S, Yasuda K, Takahashi A, Horikoshi M, Nakamura M, Fujita H, Grarup N, Cauchi S, Ng DP, Ma RC, Tsunoda T, Kubo M, Watada H, Maegawa H, Okada-Iwabu M, Iwabu M, Shojima N, Shin HD, Andersen G, Witte DR, Jørgensen T, Lauritzen T, Sandbæk A, Hansen T, Ohshige T, Omori S, Saito I, Kaku K, Hirose H, So WY, Beury D, Chan JC, Park KS, Tai ES, Ito C, Tanaka Y, Kashiwagi A, Kawamori R, Kasuga M, Froguel P, Pedersen O, Kamatani N, Nakamura Y, Kadowaki T (2010) A genome-wide association study in the Japanese population identifies susceptibility loci for type 2 diabetes at *UBE2E2* and *C2CD4A-C2CD4B*. Nat Genet 42:864–868
38. Morris AP, Voight BF, Teslovich TM, Ferreira T, Segrè AV, Steinthorsdottir V, Strawbridge RJ, Khan H, Grallert H, Mahajan A, Prokopenko I, Kang HM, Dina C, Esko T, Fraser RM, Kanoni S, Kumar A, Lagou V, Langenberg C, Luan J, Lindgren CM, Müller-Nurasyid M, Pechlivanis S, Rayner NW, Scott LJ, Wiltshire S, Yengo L, Kinnunen L, Rossin EJ, Raychaudhuri S, Johnson AD, Dimas AS, Loos RJ, Vedantam S, Chen H, Florez JC, Fox

C, Liu CT, Rybin D, Couper DJ, Kao WH, Li M, Cornelis MC, Kraft P, Sun Q, van Dam RM, Stringham HM, Chines PS, Fischer K, Fontanillas P, Holmen OL, Hunt SE, Jackson AU, Kong A, Lawrence R, Meyer J, Perry JR, Platou CG, Potter S, Rehnberg E, Robertson N, Sivapalaratnam S, Stančáková A, Stirrups K, Thorleifsson G, Tikkanen E, Wood AR, Almgren P, Atalay M, Benediktsson R, Bonnycastle LL, Burtt N, Carey J, Charpentier G, Crenshaw AT, Doney AS, Dorkhan M, Edkins S, Emilsson V, Eury E, Forsen T, Gertow K, Gigante B, Grant GB, Groves CJ, Guiducci C, Herder C, Hreidarsson AB, Hui J, James A, Jonsson A, Rathmann W, Klopp N, Kravic J, Krjutškov K, Langford C, Leander K, Lindholm E, Lobbens S, Männistö S, Mirza G, Mühleisen TW, Musk B, Parkin M, Rallidis L, Saramies J, Sennblad B, Shah S, Sigurðsson G, Silveira A, Steinbach G, Thorand B, Trakalo J, Veglia F, Wennauer R, Winckler W, Zabaneh D, Campbell H, van Duijn C, Uitterlinden AG, Hofman A, Sijbrands E, Abecasis GR, Owen KR, Zeggini E, Trip MD, Forouhi NG, Syvänen AC, Eriksson JG, Peltonen L, Nöthen MM, Balkau B, Palmer CN, Lyssenko V, Tuomi T, Isomaa B, Hunter DJ, Qi L, Wellcome Trust Case Control Consortium, Meta-Analyses of Glucose and Insulin-related traits Consortium (MAGIC) Investigators, Genetic Investigation of ANthropometric Traits (GIANT) Consortium, Asian Genetic Epidemiology Network–Type 2 Diabetes (AGEN-T2D) Consortium, South Asian Type 2 Diabetes (SAT2D) Consortium, Shuldiner AR, Roden M, Barroso I, Wilsgaard T, Beilby J, Hovingh K, Price JF, Wilson JF, Rauramaa R, Lakka TA, Lind L, Dedoussis G, Njølstad I, Pedersen NL, Khaw KT, Wareham NJ, Keinanen-Kiukaanniemi SM, Saaristo TE, Korpi-Hyövälti E, Saltevo J, Laakso M, Kuusisto J, Metspalu A, Collins FS, Mohlke KL, Bergman RN, Tuomilehto J, Boehm BO, Gieger C, Hveem K, Cauchi S, Froguel P, Baldassarre D, Tremoli E, Humphries SE, Saleheen D, Danesh J, Ingelsson E, Ripatti S, Salomaa V, Erbel R, Jöckel KH, Moebus S, Peters A, Illig T, de Faire U, Hamsten A, Morris AD, Donnelly PJ, Frayling TM, Hattersley AT, Boerwinkle E, Melander O, Kathiresan S, Nilsson PM, Deloukas P, Thorsteinsdottir U, Groop LC, Stefansson K, Hu F, Pankow JS, Dupuis J, Meigs JB, Altshuler D, Boehnke M, McCarthy MI, DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium (2012) Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. Nat Genet 44:981–990

39. Cho YS, Chen CH, Hu C, Long J, Ong RT, Sim X, Takeuchi F, Wu Y, Go MJ, Yamauchi T, Chang YC, Kwak SH, Ma RC, Yamamoto K, Adair LS, Aung T, Cai Q, Chang LC, Chen YT, Gao Y, Hu FB, Kim HL, Kim S, Kim YJ, Lee JJ, Lee NR, Li Y, Liu JJ, Lu W, Nakamura J, Nakashima E, Ng DP, Tay WT, Tsai FJ, Wong TY, Yokota M, Zheng W, Zhang R, Wang C, So WY, Ohnaka K, Ikegami H, Hara K, Cho YM, Cho NH, Chang TJ, Bao Y, Hedman ÅK, Morris AP, McCarthy MI, DIAGRAM Consortium, MuTHER Consortium, Takayanagi R, Park KS, Jia W, Chuang LM, Chan JC, Maeda S, Kadowaki T, Lee JY, Wu JY, Teo YY, Tai ES, Shu XO, Mohlke KL, Kato N, Han BG, Seielstad M (2011) Meta-analysis of genome-wide association studies identifies eight new loci for type 2 diabetes in east Asians. Nat Genet 44:67–72
40. Voight BF, Scott LJ, Steinthorsdottir V, Morris AP, Dina C, Welch RP, Zeggini E, Huth C, Aulchenko YS, Thorleifsson G, McCulloch LJ, Ferreira T, Grallert H, Amin N, Wu G, Willer CJ, Raychaudhuri S, McCarroll SA, Langenberg C, Hofmann OM, Dupuis J, Qi L, Segrè AV, van Hoek M, Navarro P, Ardlie K, Balkau B, Benediktsson R, Bennett AJ, Blagieva R, Boerwinkle E, Bonnycastle LL, Bengtsson Boström K, Bravenboer B, Bumpstead S, Burtt NP, Charpentier G, Chines PS, Cornelis M, Couper DJ, Crawford G, Doney AS, Elliott KS, Elliott AL, Erdos MR, Fox CS, Franklin CS, Ganser M, Gieger C, Grarup N, Green T, Griffin S, Groves CJ, Guiducci C, Hadjadj S, Hassanali N, Herder C, Isomaa B, Jackson AU, Johnson PR, Jørgensen T, Kao WH, Klopp N, Kong A, Kraft P, Kuusisto J, Lauritzen T, Li M, Lieverse A, Lindgren CM, Lyssenko V, Marre M, Meitinger T, Midthjell K, Morken MA, Narisu N, Nilsson P, Owen KR, Payne F, Perry JR, Petersen AK, Platou C, Proença C, Prokopenko I, Rathmann W, Rayner NW, Robertson NR, Rocheleau G, Roden M, Sampson MJ, Saxena R, Shields BM, Shrader P, Sigurdsson G, Sparsø T, Strassburger K, Stringham HM, Sun Q, Swift AJ, Thorand B, Tichet J, Tuomi T, van Dam RM, van Haeften TW, van Herpt T, van

Vliet-Ostaptchouk JV, Walters GB, Weedon MN, Wijmenga C, Witteman J, Bergman RN, Cauchi S, Collins FS, Gloyn AL, Gyllensten U, Hansen T, Hide WA, Hitman GA, Hofman A, Hunter DJ, Hveem K, Laakso M, Mohlke KL, Morris AD, Palmer CN, Pramstaller PP, Rudan I, Sijbrands E, Stein LD, Tuomilehto J, Uitterlinden A, Walker M, Wareham NJ, Watanabe RM, Abecasis GR, Boehm BO, Campbell H, Daly MJ, Hattersley AT, Hu FB, Meigs JB, Pankow JS, Pedersen O, Wichmann HE, Barroso I, Florez JC, Frayling TM, Groop L, Sladek R, Thorsteinsdottir U, Wilson JF, Illig T, Froguel P, van Duijn CM, Stefansson K, Altshuler D, Boehnke M, McCarthy MI, MAGIC investigators, GIANT Consortium (2010) Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. Nat Genet 42:579–589

41. Hara K, Fujita H, Johnson TA, Yamauchi T, Yasuda K, Horikoshi M, Peng C, Hu C, Ma RC, Imamura M, Iwata M, Tsunoda T, Morizono T, Shojima N, So WY, Leung TF, Kwan P, Zhang R, Wang J, Yu W, Maegawa H, Hirose H, DIAGRAM consortium, Kaku K, Ito C, Watada H, Tanaka Y, Tobe K, Kashiwagi A, Kawamori R, Jia W, Chan JC, Teo YY, Shyong TE, Kamatani N, Kubo M, Maeda S, Kadowaki T (2014) Genome-wide association study identifies three novel loci for type 2 diabetes. Hum Mol Genet 23:239–246
42. Waters KM, Stram DO, Hassanein MT, Le Marchand L, Wilkens LR, Maskarinec G, Monroe KR, Kolonel LN, Altshuler D, Henderson BE, Haiman CA (2010) Consistent association of type 2 diabetes risk variants found in europeans in diverse racial and ethnic groups. PLoS Genet 6:e1001078
43. Saxena R, Elbers CC, Guo Y, Peter I, Gaunt TR, Mega JL, Lanktree MB, Tare A, Castillo BA, Li YR, Johnson T, Bruinenberg M, Gilbert-Diamond D, Rajagopalan R, Voight BF, Balasubramanyam A, Barnard J, Bauer F, Baumert J, Bhangale T, Böhm BO, Braund PS, Burton PR, Chandrupatla HR, Clarke R, Cooper-DeHoff RM, Crook ED, Davey-Smith G, Day IN, de Boer A, de Groot MC, Drenos F, Ferguson J, Fox CS, Furlong CE, Gibson Q, Gieger C, Gilhuijs-Pederson LA, Glessner JT, Goel A, Gong Y, Grant SF, Grobbee DE, Hastie C, Humphries SE, Kim CE, Kivimaki M, Kleber M, Meisinger C, Kumari M, Langaee TY, Lawlor DA, Li M, Lobmeyer MT, Maitland-van der Zee AH, Meijs MF, Molony CM, Morrow DA, Murugesan G, Musani SK, Nelson CP, Newhouse SJ, O'Connell JR, Padmanabhan S, Palmen J, Patel SR, Pepine CJ, Pettinger M, Price TS, Rafelt S, Ranchalis J, Rasheed A, Rosenthal E, Ruczinski I, Shah S, Shen H, Silbernagel G, Smith EN, Spijkerman AW, Stanton A, Steffes MW, Thorand B, Trip M, van der Harst P, van der A DL, van Iperen EP, van Setten J, van Vliet-Ostaptchouk JV, Verweij N, Wolffenbuttel BH, Young T, Zafarmand MH, Zmuda JM, Look AHEAD Research Group, DIAGRAM consortium, Boehnke M, Altshuler D, McCarthy M, Kao WH, Pankow JS, Cappola TP, Sever P, Poulter N, Caulfield M, Dominiczak A, Shields DC, Bhatt DL, Zhang L, Curtis SP, Danesh J, Casas JP, van der Schouw YT, Onland-Moret NC, Doevendans PA, Dorn GW 2nd, Farrall M, GA FG, Hamsten A, Hegele R, Hingorani AD, Hofker MH, Huggins GS, Illig T, Jarvik GP, Johnson JA, Klungel OH, Knowler WC, Koenig W, März W, Meigs JB, Melander O, Munroe PB, Mitchell BD, Bielinski SJ, Rader DJ, Reilly MP, Rich SS, Rotter JI, Saleheen D, Samani NJ, Schadt EE, Shuldiner AR, Silverstein R, Kottke-Marchant K, Talmud PJ, Watkins H, Asselbergs FW, de Bakker PI, McCaffery J, Wijmenga C, Sabatine MS, Wilson JG, Reiner A, Bowden DW, Hakonarson H, Siscovick DS, Keating BJ (2012) Large-scale gene-centric meta-analysis across 39 studies identifies type 2 diabetes loci. Am J Hum Genet 90:410–425
44. DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium, Asian Genetic Epidemiology Network Type 2 Diabetes (AGEN-T2D) Consortium, South Asian Type 2 Diabetes (SAT2D) Consortium, Mexican American Type 2 Diabetes (MAT2D) Consortium, Type 2 Diabetes Genetic Exploration by Nex-generation sequencing in muylti-Ethnic Samples (T2D-GENES) Consortium, Mahajan A, Go MJ, Zhang W, Below JE, Gaulton KJ, Ferreira T, Horikoshi M, Johnson AD, Ng MC, Prokopenko I, Saleheen D, Wang X, Zeggini E, Abecasis GR, Adair LS, Almgren P, Atalay M, Aung T, Baldassarre D, Balkau B, Bao Y, Barnett AH, Barroso I, Basit A, Been LF, Beilby J, Bell GI, Benediktsson R, Bergman RN, Boehm BO, Boerwinkle E, Bonnycastle LL, Burtt N, Cai Q, Campbell H, Carey J, Cauchi

S, Caulfield M, Chan JC, Chang LC, Chang TJ, Chang YC, Charpentier G, Chen CH, Chen H, Chen YT, Chia KS, Chidambaram M, Chines PS, Cho NH, Cho YM, Chuang LM, Collins FS, Cornelis MC, Couper DJ, Crenshaw AT, van Dam RM, Danesh J, Das D, de Faire U, Dedoussis G, Deloukas P, Dimas AS, Dina C, Doney AS, Donnelly PJ, Dorkhan M, van Duijn C, Dupuis J, Edkins S, Elliott P, Emilsson V, Erbel R, Eriksson JG, Escobedo J, Esko T, Eury E, Florez JC, Fontanillas P, Forouhi NG, Forsen T, Fox C, Fraser RM, Frayling TM, Froguel P, Frossard P, Gao Y, Gertow K, Gieger C, Gigante B, Grallert H, Grant GB, Grrop LC, Groves CJ, Grundberg E, Guiducci C, Hamsten A, Han BG, Hara K, Hassanali N, Hattersley AT, Hayward C, Hedman AK, Herder C, Hofman A, Holmen OL, Hovingh K, Hreidarsson AB, Hu C, Hu FB, Hui J, Humphries SE, Hunt SE, Hunter DJ, Hveem K, Hydrie ZI, Ikegami H, Illig T, Ingelsson E, Islam M, Isomaa B, Jackson AU, Jafar T, James A, Jia W, Jöckel KH, Jonsson A, Jowett JB, Kadowaki T, Kang HM, Kanoni S, Kao WH, Kathiresan S, Kato N, Katulanda P, Keinanen-Kiukaanniemi KM, Kelly AM, Khan H, Khaw KT, Khor CC, Kim HL, Kim S, Kim YJ, Kinnunen L, Klopp N, Kong A, Korpi-Hyövälti E, Kowlessur S, Kraft P, Kravic J, Kristensen MM, Krithika S, Kumar A, Kumate J, Kuusisto J, Kwak SH, Laakso M, Lagou V, Lakka TA, Langenberg C, Langford C, Lawrence R, Leander K, Lee JM, Lee NR, Li M, Li X, Li Y, Liang J, Liju S, Lim WY, Lind L, Lindgren CM, Lindholm E, Liu CT, Liu JJ, Lobbens S, Long J, Loos RJ, Lu W, Luan J, Lyssenko V, Ma RC, Maeda S, Mägi R, Männisto S, Matthews DR, Meigs JB, Melander O, Metspalu A, Meyer J, Mirza G, Mihailov E, Moebus S, Mohan V, Mohlke KL, Morris AD, Mühleisen TW, Müller-Nurasyid M, Musk B, Nakamura J, Nakashima E, Navarro P, Ng PK, Nica AC, Nilsson PM, Njølstad I, Nöthen MM, Ohnaka K, Ong TH, Owen KR, Palmer CN, Pankow JS, Park KS, Parkin M, Pechlivanis S, Pedersen NL, Peltonen L, Perry JR, Peters A, Pinidiyapathirage JM, Platou CG, Potter S, Price JF, Qi L, Radha V, Rallidis L, Rasheed A, Rathman W, Rauramaa R, Raychaudhuri S, Rayner NW, Rees SD, Rehnberg E, Ripatti S, Robertson N, Roden M, Rossin EJ, Rudan I, Rybin D, Saaristo TE, Salomaa V, Saltevo J, Samuel M, Sanghera DK, Saramies J, Scott J, Scott LJ, Scott RA, Segrè AV, Sehmi J, Sennblad B, Shah N, Shah S, Shera AS, Shu XO, Shuldiner AR, Sigurđsson G, Sijbrands E, Silveira A, Sim X, Sivapalaratnam S, Small KS, So WY, Stančáková A, Stefansson K, Steinbach G, Steinthorsdottir V, Stirrups K, Strawbridge RJ, Stringham HM, Sun Q, Suo C, Syvänen AC, Takayanagi R, Takeuchi F, Tay WT, Teslovich TM, Thorand B, Thorleifsson G, Thorsteinsdottir U, Tikkanen E, Trakalo J, Tremoli E, Trip MD, Tsai FJ, Tuomi T, Tuomilehto J, Uitterlinden AG, Valladares-Salgado A, Vedantam S, Veglia F, Voight BF, Wang C, Wareham NJ, Wennauer R, Wickremasinghe AR, Wilsgaard T, Wilson JF, Wiltshire S, Winckler W, Wong TY, Wood AR, Wu JY, Wu Y, Yamamoto K, Yamauchi T, Yang M, Yengo L, Yokota M, Young R, Zabaneh D, Zhang F, Zhang R, Zheng W, Zimmet PZ, Altshuler D, Bowden DW, Cho YS, Cox NJ, Cruz M, Hanis CL, Kooner J, Lee JY, Seielstad M, Teo YY, Boehnke M, Parra EJ, Chambers JC, Tai ES, McCarthy MI, Morris AP (2014) Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. Nat Genet 46:234–244

45. Dimas AS, Lagou V, Barker A, Knowles JW, Mägi R, Hivert MF, Benazzo A, Rybin D, Jackson AU, Stringham HM, Song C, Fischer-Rosinsky A, Boesgaard TW, Grarup N, Abbasi FA, Assimes TL, Hao K, Yang X, Lecoeur C, Barroso I, Bonnycastle LL, Böttcher Y, Bumpstead S, Chines PS, Erdos MR, Graessler J, Kovacs P, Morken MA, Narisu N, Payne F, Stancakova A, Swift AJ, Tönjes A, Bornstein SR, Cauchi S, Froguel P, Meyre D, Schwarz PE, Häring HU, Smith U, Boehnke M, Bergman RN, Collins FS, Mohlke KL, Tuomilehto J, Quertemous T, Lind L, Hansen T, Pedersen O, Walker M, Pfeiffer AF, Spranger J, Stumvoll M, Meigs JB, Wareham NJ, Kuusisto J, Laakso M, Langenberg C, Dupuis J, Watanabe RM, Florez JC, Ingelsson E, McCarthy MI, Prokopenko I, MAGIC Investigators (2014) Impact of type 2 diabetes susceptibility variants on quantitative glycemic traits reveals mechanistic heterogeneity. Diabetes 63:2158–2171
46. Stahl EA, Wegmann D, Trynka G, Gutierrez-Achury J, Do R, Voight BF, Kraft P, Chen R, Kallberg HJ, Kurreeman FA, Diabetes Genetics Replication and Meta-analysis Consortium, Myocardial Infarction Genetics Consortium, Kathiresan S, Wijmenga C, Gregersen PK,

Alfredsson L, Siminovitch KA, Worthington J, de Bakker PI, Raychaudhuri S, Plenge RM (2012) Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. Nat Genet 44:483–489

47. Steinthorsdottir V, Thorleifsson G, Sulem P, Helgason H, Grarup N, Sigurdsson A, Helgadottir HT, Johannsdottir H, Magnusson OT, Gudjonsson SA, Justesen JM, Harder MN, Jørgensen ME, Christensen C, Brandslund I, Sandbæk A, Lauritzen T, Vestergaard H, Linneberg A, Jørgensen T, Hansen T, Daneshpour MS, Fallah MS, Hreidarsson AB, Sigurdsson G, Azizi F, Benediktsson R, Masson G, Helgason A, Kong A, Gudbjartsson DF, Pedersen O, Thorsteinsdottir U, Stefansson K (2014) Identification of low-frequency and rare sequence variants associated with elevated or reduced risk of type 2 diabetes. Nat Genet 46:294–298
48. Huyghe JR, Jackson AU, Fogarty MP, Buchkovich ML, Stančáková A, Stringham HM, Sim X, Yang L, Fuchsberger C, Cederberg H, Chines PS, Teslovich TM, Romm JM, Ling H, McMullen I, Ingersoll R, Pugh EW, Doheny KF, Neale BM, Daly MJ, Kuusisto J, Scott LJ, Kang HM, Collins FS, Abecasis GR, Watanabe RM, Boehnke M, Laakso M, Mohlke KL (2013) Exome array analysis identifies new loci and low-frequency variants influencing insulin processing and secretion. Nat Genet 45:197–201
49. Fuchsberger C, Flannick J, Teslovich TM, Mahajan A, Agarwala V, Gaulton KJ, Ma C, Fontanillas P, Moutsianas L, McCarthy DJ, Rivas MA, Perry JR, Sim X, Blackwell TW, Robertson NR, Rayner NW, Cingolani P, Locke AE, Fernandez Tajes J, Highland HM, Dupuis J, Chines PS, Lindgren CM, Hartl C, Jackson AU, Chen H, Huyghe JR, van de Bunt M, Pearson RD, Kumar A, Müller-Nurasyid M, Grarup N, Stringham HM, Gamazon ER, Lee J, Chen Y, Scott RA, Below JE, Chen P, Huang J, Go MJ, Stitzel ML, Pasko D, Parker SC, Varga TV, Green T, Beer NL, Day-Williams AG, Ferreira T, Fingerlin T, Horikoshi M, Hu C, Huh I, Ikram MK, Kim BJ, Kim Y, Kim YJ, Kwon MS, Lee J, Lee S, Lin KH, Maxwell TJ, Nagai Y, Wang X, Welch RP, Yoon J, Zhang W, Barzilai N, Voight BF, Han BG, Jenkinson CP, Kuulasmaa T, Kuusisto J, Manning A, Ng MC, Palmer ND, Balkau B, Stancáková A, Abboud HE, Boeing H, Giedraitis V, Prabhakaran D, Gottesman O, Scott J, Carey J, Kwan P, Grant G, Smith JD, Neale BM, Purcell S, Butterworth AS, Howson JM, Lee HM, Lu Y, Kwak SH, Zhao W, Danesh J, Lam VK, Park KS, Saleheen D, So WY, Tam CH, Afzal U, Aguilar D, Arya R, Aung T, Chan E, Navarro C, Cheng CY, Palli D, Correa A, Curran JE, Rybin D, Farook VS, Fowler SP, Freedman BI, Griswold M, Hale DE, Hicks PJ, Khor CC, Kumar S, Lehne B, Thuillier D, Lim WY, Liu J, van der Schouw YT, Loh M, Musani SK, Puppala S, Scott WR, Yengo L, Tan ST, Taylor HA Jr, Thameem F, Wilson G Sr, Wong TY, Njølstad PR, Levy JC, Mangino M, Bonnycastle LL, Schwarzmayr T, Fadista J, Surdulescu GL, Herder C, Groves CJ, Wieland T, Bork-Jensen J, Brandslund I, Christensen C, Koistinen HA, Doney AS, Kinnunen L, Esko T, Farmer AJ, Hakaste L, Hodgkiss D, Kravic J, Lyssenko V, Hollensted M, Jørgensen ME, Jørgensen T, Ladenvall C, Justesen JM, Käräjämäki A, Kriebel J, Rathmann W, Lannfelt L, Lauritzen T, Narisu N, Linneberg A, Melander O, Milani L, Neville M, Orho-Melander M, Qi L, Qi Q, Roden M, Rolandsson O, Swift A, Rosengren AH, Stirrups K, Wood AR, Mihailov E, Blancher C, Carneiro MO, Maguire J, Poplin R, Shakir K, Fennell T, DePristo M, Hrabé de Angelis M, Deloukas P, Gjesing AP, Jun G, Nilsson P, Murphy J, Onofrio R, Thorand B, Hansen T, Meisinger C, Hu FB, Isomaa B, Karpe F, Liang L, Peters A, Huth C, O'Rahilly SP, Palmer CN, Pedersen O, Rauramaa R, Tuomilehto J, Salomaa V, Watanabe RM, Syvänen AC, Bergman RN, Bharadwaj D, Bottinger EP, Cho YS, Chandak GR, Chan JC, Chia KS, Daly MJ, Ebrahim SB, Langenberg C, Elliott P, Jablonski KA, Lehman DM, Jia W, Ma RC, Pollin TI, Sandhu M, Tandon N, Froguel P, Barroso I, Teo YY, Zeggini E, Loos RJ, Small KS, Ried JS, DeFronzo RA, Grallert H, Glaser B, Metspalu A, Wareham NJ, Walker M, Banks E, Gieger C, Ingelsson E, Im HK, Illig T, Franks PW, Buck G, Trakalo J, Buck D, Prokopenko I, Mägi R, Lind L, Farjoun Y, Owen KR, Gloyn AL, Strauch K, Tuomi T, Kooner JS, Lee JY, Park T, Donnelly P, Morris AD, Hattersley AT, Bowden DW, Collins FS, Atzmon G, Chambers JC, Spector TD, Laakso M, Strom TM, Bell GI, Blangero J, Duggirala R, Tai ES, McVean G, Hanis CL, Wilson JG, Seielstad M, Frayling TM, Meigs JB, Cox NJ, Sladek R, Lander ES, Gabriel S, Burtt NP, Mohlke KL, Meitinger T,

Groop L, Abecasis G, Florez JC, Scott LJ, Morris AP, Kang HM, Boehnke M, Altshuler D, McCarthy MI (2016) The genetic architecture of type 2 diabetes. Nature 536(7614):41–47
50. Hivert MF, Jablonski KA, Perreault L, Saxena R, McAteer JB, Franks PW, Hamman RF, Kahn SE, Haffner S, DIAGRAM Consortium, Meigs JB, Altshuler D, Knowler WC, Florez JC, Diabetes Prevention Program Research Group (2011) Updated genetic score based on 34 confirmed type 2 diabetes Loci is associated with diabetes incidence and regression to normoglycemia in the diabetes prevention program. Diabetes 60:1340–1348
51. Imamura M, Shigemizu D, Tsunoda T, Iwata M, Maegawa H, Watada H, Hirose H, Tanaka Y, Tobe K, Kaku K, Kashiwagi A, Kawamori R, Maeda S (2013) Assessing the clinical utility of a genetic risk score constructed using 49 susceptibility alleles for type 2 diabetes in a Japanese population. J Clin Endocrinol Metab 98:E1667–E1673
52. Wang X, Strizich G, Hu Y, Wang T, Kaplan RC, Qi Q (2016) Genetic markers of type 2 diabetes: Progress in genome-wide association studies and clinical application for risk prediction. J Diabetes 8:24–35
53. Okada Y, Wu D, Trynka G, Raj T, Terao C, Ikari K, Kochi Y, Ohmura K, Suzuki A, Yoshida S, Graham RR, Manoharan A, Ortmann W, Bhangale T, Denny JC, Carroll RJ, Eyler AE, Greenberg JD, Kremer JM, Pappas DA, Jiang L, Yin J, Ye L, Su DF, Yang J, Xie G, Keystone E, Westra HJ, Esko T, Metspalu A, Zhou X, Gupta N, Mirel D, Stahl EA, Diogo D, Cui J, Liao K, Guo MH, Myouzen K, Kawaguchi T, Coenen MJ, van Riel PL, van de Laar MA, Guchelaar HJ, Huizinga TW, Dieudé P, Mariette X, Bridges SL Jr, Zhernakova A, Toes RE, Tak PP, Miceli-Richard C, Bang SY, Lee HS, Martin J, Gonzalez-Gay MA, Rodriguez-Rodriguez L, Rantapää-Dahlqvist S, Arlestig L, Choi HK, Kamatani Y, Galan P, Lathrop M, RACI consortium, GARNET consortium, Eyre S, Bowes J, Barton A, de Vries N, Moreland LW, Criswell LA, Karlson EW, Taniguchi A, Yamada R, Kubo M, Liu JS, Bae SC, Worthington J, Padyukov L, Klareskog L, Gregersen PK, Raychaudhuri S, Stranger BE, De Jager PL, Franke L, Visscher PM, Brown MA, Yamanaka H, Mimori T, Takahashi A, Xu H, Behrens TW, Siminovitch KA, Momohara S, Matsuda F, Yamamoto K, Plenge RM (2014) Genetics of rheumatoid arthritis contributes to biology and drug discovery. Nature 506:376–381
54. Poulsen P, Levin K, Petersen I, Christensen K, Beck-Nielsen H, Vaag A (2005) Heritability of insulin secretion, peripheral and hepatic insulin action, and intracellular glucose partitioning in young and old Danish twins. Diabetes 54:275–283
55. Weedon MN, Clark VJ, Qian Y, Ben-Shlomo Y, Timpson N et al (2006) A common haplotype of the glucokinase gene alters fasting glucose and birth weight: association in six studies and population-genetics analyses. Am J Hum Genet 79:991–1001
56. Bouatia-Naji N, Rocheleau G, Van Lommel L, Lemaire K, Schuit F et al (2008) A polymorphism within the G6PC2 gene is associated with fasting plasma glucose levels. Science 320:1085–1088
57. Sparso T, Andersen G, Nielsen T, Burgdorf KS, Gjesing AP et al (2008) The GCKR rs780094 polymorphism is associated with elevated fasting serum triacylglycerol, reduced fasting and OGTT-related insulinaemia, and reduced risk of type 2 diabetes. Diabetologia 51:70–75
58. Orho-Melander M, Melander O, Guiducci C, Perez-Martinez P, Corella D et al (2008) A common missense variant in the glucokinase regulatory protein gene (GCKR) is associated with increased plasma triglyceride and C-reactive protein but lower fasting glucose concentrations. Diabetes 57:3112–3121
59. Pare G, Chasman DI, Parker AN, Nathan DM, Miletich JP, Zee RY, Ridker PM (2008) Novel association of HK1 with glycated hemoglobin in a non-diabetic population: a genome-wide evaluation of 14, 618 participants in the Women's Genome Health Study. PLoS Genet 4:e1000312
60. Dupuis J, Langenberg C, Prokopenko I, Saxena R, Soranzo N et al (2010) New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. Nat Genet 42:105–116
61. Prokopenko I, Langenberg C, Florez JC, Saxena R, Soranzo N et al (2009) Variants in MTNR1B influence fasting glucose levels. Nat Genet 41:77–81

62. Saxena R, Hivert MF, Langenberg C, Tanaka T, Pankow JS et al (2010) Genetic variation in GIPR influences the glucose and insulin responses to an oral glucose challenge. Nat Genet 42:142–148
63. Strawbridge RJ, Dupuis J, Prokopenko I, Barker A, Ahlqvist E et al (2011) Genome-wide association identifies nine common variants associated with fasting proinsulin levels and provides new insights into the pathophysiology of type 2 diabetes. Diabetes 60:2624–2634
64. Scott RA, Lagou V, Welch RP, Wheeler E, Montasser ME et al (2012) Large-scale association analyses identify new loci influencing glycemic traits and provide insight into the underlying biological pathways. Nat Genet 44:991–1005
65. Manning AK, Hivert MF, Scott RA, Grimsby JL, Bouatia-Naji N et al (2012) A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. Nat Genet 44:659–669
66. Kim YJ, Go MJ, Hu C, Hong CB, Kim et al (2011) Largescale genome-wide association studies in East Asians identify new genetic loci influencing metabolic traits. Nat. Genet 43:990–995
67. Chen G, Bentley A, Adeyemo A, Shriner D, Zhou J et al (2012) Genome-wide association study identifies novel loci association with fasting insulin and insulin resistance in African Americans. Hum Mol Genet 21:4530–4536
68. Go MJ, Hwang JY, Kim YJ, Hee Oh J, Kim YJ et al (2013) New susceptibility loci in MYL2, C12orf51 and OAS1 associated with 1-h plasma glucose as predisposing risk factors for type 2 diabetes in the Korean population. J Hum Genet 58:362–365
69. Hwang JY, Sim X, Wu Y, Liang J, Tabara Y et al (2015) Genome-wide association meta-analysis identifies novel variants associated with fasting plasma glucose in East Asians. Diabetes 64:291–298
70. Chen P, Takeuchi F, Lee JY, Li H, Wu JY et al (2014) Multiple nonglycemic genomic loci are newly associated with blood level of glycated hemoglobin in East Asians. Diabetes 63:2551–2562
71. Go MJ, Hwang JY, Park TJ, Kim YJ, Oh JH et al (2014) Genome-wide association study identifies two novel Loci with sex-specific effects for type 2 diabetes mellitus and glycemic traits in a korean population. Diabetes Metab J 38:375–387
72. Moltke I, Grarup N, Jørgensen ME, Bjerregaard P, Treebak JT, Fumagalli M, Korneliussen TS, Andersen MA, Nielsen TS, Krarup NT, Gjesing AP, Zierath JR, Linneberg A, Wu X, Sun G, Jin X, Al-Aama J, Wang J, Borch-Johnsen K, Pedersen O, Nielsen R, Albrechtsen A, Hansen T (2014) A common Greenlandic *TBC1D4* variant confers muscle insulin resistance and type 2 diabetes. Nature 512:190–193
73. Mahajan A, Sim X, Ng HJ, Manning A, Rivas MA et al (2015) Identification and functional characterization of *G6PC2* coding variants influencing glycemic traits define an effector transcript at the *G6PC2-ABCB11* locus. PLoS Genet 11:e1004876
74. Huyghe JR, Jackson AU, Fogarty MP, Buchkovich ML, Stancakova A et al (2013) Exome array analysis identifies new loci and low-frequency variants influencing insulin processing and secretion. Nat Genet 45:197–201
75. Steinthorsdottir V, Thorleifsson G, Sulem P, Helgason H, Grarup N et al (2014) Identification of low-frequency and rare sequence variants associated with elevated or reduced risk of type 2 diabetes. Nat Genet 46:294–298
76. Fuchsberger C, Flannick J, Teslovich TM, Mahajan A, Agarwala V et al (2016) The genetic architecture of type 2 diabetes. Nature 536:41–47
77. Cornes BK, Brody JA, Nikpoor N, Morrison AC, Dang HC et al (2014) Association of levels of fasting glucose and insulin with rare variants at the chromosome 11p11.2-MADD locus: Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium Targeted Sequencing Study. Circ Cardiovasc Genet 7:374–382
78. Wessel J, Chu AY, Willems SM, Wang S, Yaghootkar H et al (2015) Low-frequency and rare exome chip variants associate with fasting glucose and type 2 diabetes susceptibility. Nat Commun 6:5897

79. The Diabetes Control and Complications Trial Research Group (1993) The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus. N Engl J Med 329:977–986
80. UK Prospective Diabetes Study (UKPDS) Group (1998) Intensive blood-glucose control with sulphonylureas or insulin compared with conventional treatment and risk of complications in patients with type 2 diabetes (UKPDS 33). Lancet 352:837–853
81. Ohkubo Y, Kishikawa H, Araki E, Miyata T, Isami S, Motoyoshi S, Kojima Y, Furuyoshi N, Shichiri M (1995) Intensive insulin therapy prevents the progression of diabetic microvascular complications in Japanese patients with non-insulin-dependent diabetes mellitus: a randomized prospective 6-year study. Diabetes Res Clin Pract 28:103–117
82. Gaede P, Vedel P, Larsen N, Jensen GV, Parving HH, Pedersen O (2003) Multifactorial intervention and cardiovascular disease in patients with type 2 diabetes. N Engl J Med 348:383–393
83. Perkins BA, Ficociello LH, Silva KH, Finkelstein DM, Warram JH, Krolewski AS (2003) Regression of microalbuminuria in type 1 diabetes. N Engl J Med 348:2285–2293
84. Gaede P, Tarnow L, Vedel P, Parving HH, Pedersen O (2004) Remission to normoalbuminuria during multifactorial treatment preserves kidney function in patients with type 2 diabetes and microalbuminuria. Nephrol Dial Transplant 19:2784–2788
85. Araki S, Haneda M, Sugimoto T, Isono M, Isshiki K, Kashiwagi A, Koya D (2005) Factors associated with frequent remission of microalbuminuria in patients with type 2 diabetes. Diabetes 54:2983–2987
86. Brownlee M (2005) The pathobiology of diabetic complications: a unifying mechanism. Diabetes 54:1615–1625
87. Krolewski AS, Warram JH, Rand LI, Kahn CR (1987) Epidemiologic approach to the etiology of type 1 diabetes mellitus and its complications. N Engl J Med 317:1390–1398
88. Quinn M, Angelico MC, Warram JH, Krolewski AS (1996) Familial factors determine the development of diabetic nephropathy in patients with IDDM. Diabetologia 39:940–945
89. Fava S, Azzopardi J, Hattersley AT, Watkins PJ (2000) Increased prevalence of proteinuria in diabetic sibs of proteinuric type 2 diabetic subjects. Am J Kidney Dis 35:708–712
90. Pezzolesi MG, Poznik GD, Mychaleckyj JC, Paterson AD, Barati MT, Klein JB, Ng DP, Placha G, Canani LH, Bochenski J, Waggott D, Merchant ML, Krolewski B, Mirea L, Wanic K, Katavetin P, Kure M, Wolkow P, Dunn JS, Smiles A, Walker WH, Boright AP, Bull SB, DCCT/EDIC Research Group, Doria A, Rogus JJ, Rich SS, Warram JH, Krolewski AS (2009) Genome-wide association scan for diabetic nephropathy susceptibility genes in type 1 diabetes. Diabetes 58:1403–1410
91. Pezzolesi MG, Jeong J, Smiles AM, Skupien J, Mychaleckyj JC, Rich SS, Warram JH, Krolewski AS (2013) Family-based association analysis confirms the role of the chromosome 9q21.32 locus in the susceptibility of diabetic nephropathy. PLoS One 8(3):e60301
92. Freedman BI, Langefeld CD, Lu L, Divers J, Comeau ME, Kopp JB, Winkler CA, Nelson GW, Johnson RC, Palmer ND, Hicks PJ, Bostrom MA, Cooke JN, McDonough CW, Bowden DW (2011) Differential effects of MYH9 and APOL1 risk variants on FRMD3 association with diabetic ESRD in African Americans. PLoS Genet 7:e1002150
93. Maeda S, Araki S, Babazono T, Toyoda M, Umezono T, Kawai K, Imanishi M, Uzu T, Watada H, Suzuki D, Kashiwagi A, Iwamoto Y, Kaku K, Kawamori R, Nakamura Y (2010) Replication study for the association between four Loci identified by a genome-wide association study on European American subjects with type 1 diabetes and susceptibility to diabetic nephropathy in Japanese subjects with type 2 diabetes. Diabetes 59:2075–2079
94. Sandholm N, Salem RM, McKnight AJ, Brennan EP, Forsblom C, Isakova T, McKay GJ, Williams WW, Sadlier DM, Mäkinen VP, Swan EJ, Palmer C, Boright AP, Ahlqvist E, Deshmukh HA, Keller BJ, Huang H, Ahola AJ, Fagerholm E, Gordin D, Harjutsalo V, He B, Heikkilä O, Hietala K, Kytö J, Lahermo P, Lehto M, Lithovius R, Osterholm AM, Parkkonen M, Pitkäniemi J, Rosengård-Bärlund M, Saraheimo M, Sarti C, Söderlund J, Soro-Paavonen A, Syreeni A, Thorn LM, Tikkanen H, Tolonen N, Tryggvason K, Tuomilehto J, Wadén J, Gill

GV, Prior S, Guiducci C, Mirel DB, Taylor A, Hosseini SM, DCCT/EDIC Research Group, Parving HH, Rossing P, Tarnow L, Ladenvall C, Alhenc-Gelas F, Lefebvre P, Rigalleau V, Roussel R, Tregouet DA, Maestroni A, Maestroni S, Falhammar H, Gu T, Möllsten A, Cimponeriu D, Ioana M, Mota M, Mota E, Serafinceanu C, Stavarachi M, Hanson RL, Nelson RG, Kretzler M, Colhoun HM, Panduru NM, Gu HF, Brismar K, Zerbini G, Hadjadj S, Marre M, Groop L, Lajer M, Bull SB, Waggott D, Paterson AD, Savage DA, Bain SC, Martin F, Hirschhorn JN, Godson C, Florez JC, Groop PH, Maxwell AP (2012) New susceptibility loci associated with kidney disease in type 1 diabetes. PLoS Genet 8:e1002921

95. Maeda S, Imamura M, Kurashige M, Araki S, Suzuki D, Babazono T, Uzu T, Umezono T, Toyoda M, Kawai K, Imanishi M, Hanaoka K, Maegawa H, Uchigata Y, Hosoya T (2013) Replication study for the association of 3 SNP loci identified in a genome-wide association study for diabetic nephropathy in European type 1 diabetes with diabetic nephropathy in Japanese patients with type 2 diabetes. Clin Exp Nephrol 17:866–871
96. Germain M, Pezzolesi MG, Sandholm N, McKnight AJ, Susztak K, Lajer M, Forsblom C, Marre M, Parving HH, Rossing P, Toppila I, Skupien J, Roussel R, Ko YA, Ledo N, Folkersen L, Civelek M, Maxwell AP, Tregouet DA, Groop PH, Tarnow L, Hadjadj S (2015) SORBS1 gene, a new candidate for diabetic nephropathy: results from a multi-stage genome-wide association study in patients with type 1 diabetes. Diabetologia 58:543–548
97. Sandholm N, McKnight AJ, Salem RM, Brennan EP, Forsblom C, Harjutsalo V, Mäkinen VP, McKay GJ, Sadlier DM, Williams WW, Martin F, Panduru NM, Tarnow L, Tuomilehto J, Tryggvason K, Zerbini G, Comeau ME, Langefeld CD, FIND Consortium, Godson C, Hirschhorn JN, Maxwell AP, Florez JC, Groop PH, FinnDiane Study Group and the GENIE Consortium (2013) Chromosome 2q31.1 associates with ESRD in women with type 1 diabetes. J Am Soc Nephrol 24:1537–1543
98. Teumer A, Tin A, Sorice R, Gorski M, Yeo NC, Chu AY, Li M, Li Y, Mijatovic V, Ko YA, Taliun D, Luciani A, Chen MH, Yang Q, Foster MC, Olden M, Hiraki LT, Tayo BO, Fuchsberger C, Dieffenbach AK, Shuldiner AR, Smith AV, Zappa AM, Lupo A, Kollerits B, Ponte B, Stengel B, Krämer BK, Paulweber B, Mitchell BD, Hayward C, Helmer C, Meisinger C, Gieger C, Shaffer CM, Müller C, Langenberg C, Ackermann D, Siscovick D, DCCT/EDIC, Boerwinkle E, Kronenberg F, Ehret GB, Homuth G, Waeber G, Navis G, Gambaro G, Malerba G, Eiriksdottir G, Li G, Wichmann HE, Grallert H, Wallaschofski H, Völzke H, Brenner H, Kramer H, Mateo Leach I, Rudan I, Hillege HL, Beckmann JS, Lambert JC, Luan J, Zhao JH, Chalmers J, Coresh J, Denny JC, Butterbach K, Launer LJ, Ferrucci L, Kedenko L, Haun M, Metzger M, Woodward M, Hoffman MJ, Nauck M, Waldenberger M, Pruijm M, Bochud M, Rheinberger M, Verweij N, Wareham NJ, Endlich N, Soranzo N, Polasek O, van der Harst P, Pramstaller PP, Vollenweider P, Wild PS, Gansevoort RT, Rettig R, Biffar R, Carroll RJ, Katz R, Loos RJ, Hwang SJ, Coassin S, Bergmann S, Rosas SE, Stracke S, Harris TB, Corre T, Zeller T, Illig T, Aspelund T, Tanaka T, Lendeckel U, Völker U, Gudnason V, Chouraki V, Koenig W, Kutalik Z, O'Connell JR, Parsa A, Heid IM, Paterson AD, de Boer IH, Devuyst O, Lazar J, Endlich K, Susztak K, Tremblay J, Hamet P, Jacob HJ, Böger CA, Fox CS, Pattaro C, Köttgen A (2016) Genome-wide association studies identify genetic loci associated with albuminuria in diabetes. Diabetes 65:803–817
99. Maeda S (2004) Genome-wide search for susceptibility gene to diabetic nephropathy by gene-based SNP. Diabetes Res Clin Pract 66S:S45–S47
100. Maeda S, Osawa N, Hayashi T, Tsukada S, Kobayashi M, Kikkawa R (2007) Genetic variations associated with diabetic nephropathy and type 2 diabetes in a Japanese population. Kidney Int 72:S43–S48
101. Haga H, Yamada R, Ohnishi Y, Nakamura Y, Tanaka T (2002) Gene-based SNP discovery as part of the Japanese millennium genome project: identification of 190,562 genetic variation in the human genome. J Hum Genet 47:605–610
102. Hirakawa M, Tanaka T, Hashimoto Y, Kuroda M, Takagi T, Nakamura Y (2002) JSNP a database of common gene variations in the Japanese population. Nucleic Acids Res 30:158–162

103. Tanaka N, Babazono T, Saito S, Sekine A, Tsunoda T, Haneda M, Tanaka Y, Fujioka T, Kaku K, Kawamori R, Kikkawa R, Iwamoto Y, Nakamura Y, Maeda S (2003) Association of solute carrier family 12 (sodium/chloride) member 3 with diabetic nephropathy, identified by genome-wide analyses of single nucleotide polymorphisms. Diabetes 52:2848–2853
104. Shimazaki A, Kawamura Y, Kanazawa A, Sekine A, Saito S, Tsunoda T, Koya D, Babazono T, Tanaka Y, Matsuda M, Kawai K, Iiizumi T, Imanishi M, Shinosaki T, Yanagimoto T, Ikeda M, Omachi S, Kashiwagi A, Kaku K, Iwamoto Y, Kawamori R, Kikkawa R, Nakajima M, Nakamura Y, Maeda S (2005) Genetic variations in the gene encoding ELMO1 are associated with susceptibility to diabetic nephropathy. Diabetes 54:1171–1178
105. Kamiyama M, Kobayashi M, Araki S, Iida A, Tsunoda T, Kawai K, Imanishi M, Nomura M, Babazono T, Iwamoto Y, Kashiwagi A, Kaku K, Kawamori R, Ng DP, Hansen T, Gaede P, Pedersen O, Nakamura Y, Maeda S (2007) Polymorphisms in the 3′ UTR in the neurocalcin delta gene affect mRNA stability, and confer susceptibility to diabetic nephropathy. Hum Genet 122(3–4):397–407
106. Simon DB, Nelson-Williams C, Bia MJ, Ellison D, Karet FE, Molina AM, Vaara I, Iwata F, Cushner HM, Koolen M, Gainza FJ, Gitleman HJ, Lifton RP (1996) Gitelman's variant of Bartter's syndrome, inherited hypokalaemic alkalosis, is caused by mutations in the thiazide-sensitive Na-Cl cotransporter. Nat Genet 12:24–30
107. Nishiyama K, Tanaka Y, Nakajima K, Mokubo A, Atsumi Y, Matsuoka K, Watada H, Hirose T, Nomiyama T, Maeda S, Kawamori R (2005) Polymorphism of the solute carrier family 12 (sodium/chloride transporters) member 3, *SLC12A3*, gene at exon 23 (+78G/A: Arg913Gln) is associated with elevation of urinary albumin excretion in Japanese patients with type 2 diabetes: a 10-year longitudinal study. Diabetologia 48:1335–1338
108. Bodhini D, Chidambaram M, Liju S, Revathi B, Laasya D, Sathish N, Kanthimathi S, Ghosh S, Anjana RM, Mohan V, Radha V (2016) Association of rs11643718 *SLC12A3* and rs741301 *ELMO1* variants with diabetic nephropathy in South Indian population. Ann Hum Genet 80:336–341
109. Abu Seman N, He B, Ojala JR, Wan Mohamud WN, Östenson CG, Brismar K, Gu HF (2014) Genetic and biological effects of sodium-chloride cotransporter (*SLC12A3*) in diabetic nephropathy. Am J Nephrol 40:408–416
110. Kim JH, Shin HD, Park BL, Moon MK, Cho YM, Hwang YH, Oh KW, Kim SY, Lee HK, Ahn C, Park KS (2006) *SLC12A3* (solute carrier family 12 member [sodium/chloride] 3) polymorphisms are associated with end-stage renal disease in diabetic nephropathy. Diabetes 55:843–848
111. Ng DP, Nurbaya S, Choo S, Koh D, Chia KS, Krolewski AS (2008) Genetic variation at the SLC12A3 locus is unlikely to explain risk for advanced diabetic nephropathy in Caucasians with type 2 diabetes. Nephrol Dial Transplant 23:2260–2264
112. Leak TS, Perlegas PS, Smith SG, Keene KL, Hicks PJ, Langefeld CD, Mychaleckyj JC, Rich SS, Kirk JK, Freedman BI, Bowden DW, Sale MM (2009) Variants in intron 13 of the *ELMO1* gene are associated with diabetic nephropathy in African Americans. Ann Hum Genet 73:152–159
113. Pezzolesi MG, Katavetin P, Kure M, Poznik GD, Skupien J, Mychaleckyj JC, Rich SS, Warram JH, Krolewski AS (2009) Confirmation of genetic associations at ELMO1 in the GoKinD collection supports its role as a susceptibility gene in diabetic nephropathy. Diabetes 58:2698–2702
114. Wu HY, Wang Y, Chen M, Zhang X, Wang D, Pan Y, Li L, Liu D, Dai XM (2013) Association of ELMO1 gene polymorphisms with diabetic nephropathy in Chinese population. J Endocrinol Investig 36:298–302
115. Hanson RL, Millis MP, Young NJ, Kobes S, Nelson RG, Knowler WC, DiStefano JK (2010) ELMO1 variants and susceptibility to diabetic nephropathy in American Indians. Mol Genet Metab 101:383–390
116. Gumienny TL, Brugnera E, Tosello-Trampont AC, Kinchen JM, Haney LB, Nishiwaki K, Walk SF, Nemergut ME, Macara IG, Francis R, Schedl T, Qin Y, Van Aelst L, Hengartner

MO, Ravichandran KS (2001) CED-12/ELMO, a novel member of the CrkII/Dock180/Rac pathway, is required for phagocytosis and cell migration. Cell 107:27–41
117. Brugnera E, Haney L, Grimsley C, Lu M, Walk SF, Tosello-Trampont AC, Macara IG, Madhani H, Fink GR, Ravichandran KS (2002) Unconventional Rac-GEF activity is mediated through the Dock180-ELMO complex. Nat Cell Biol 4:574–582
118. Shimazaki A, Tanaka Y, Shinosaki T, Ikeda M, Watada H, Hirose T, Kawamori R, Maeda S (2006) ELMO1 increases expression of extracellular matrix proteins and inhibits cell adhesion to ECMs. Kidney Int 70:1769–1776
119. Hathaway CK, Chang AS, Grant R, Kim HS, Madden VJ, Bagnell CR Jr, Jennette JC, Smithies O, Kakoki M (2016) High Elmo1 expression aggravates and low Elmo1 expression prevents diabetic nephropathy. Proc Natl Acad Sci U S A 113:2218–2222
120. Sharma KR, Heckler K, Stoll SJ, Hillebrands JL, Kynast K, Herpel E, Porubsky S, Elger M, Hadaschik B, Bieback K, Hammes HP, Nawroth PP, Kroll J (2016) ELMO1 protects renal structure and ultrafiltration in kidney development and under diabetic conditions. Sci Rep 6:37172
121. Maeda S, Kobayashi MA, Araki S, Babazono T, Freedman BI, Bostrom MA, Cooke JN, Toyoda M, Umezono T, Tarnow L, Hansen T, Gaede P, Jorsal A, Ng DP, Ikeda M, Yanagimoto T, Tsunoda T, Unoki H, Kawai K, Imanishi M, Suzuki D, Shin HD, Park KS, Kashiwagi A, Iwamoto Y, Kaku K, Kawamori R, Parving HH, Bowden DW, Pedersen O, Nakamura Y (2010) A single nucleotide polymorphism within the acetyl-coenzyme A carboxylase beta gene is associated with proteinuria in patients with type 2 diabetes. PLoS Genet 6:e1000842
122. Tang SC, Leung VT, Chan LY, Wong SS, Chu DW, Leung JC, Ho YW, Lai KN, Ma L, Elbein SC, Bowden DW, Hicks PJ, Comeau ME, Langefeld CD, Freedman BI (2010) The acetyl-coenzyme A carboxylase beta (ACACB) gene is associated with nephropathy in Chinese patients with type 2 diabetes. Nephrol Dial Transplant 25:3931–3934
123. Shah VN, Cheema BS, Sharma R, Khullar M, Kohli HS, Ahluwalia TS, Mohan V, Bhansali A (2013) ACACβ gene (rs2268388) and AGTR1 gene (rs5186) polymorphism and the risk of nephropathy in Asian Indian patients with type 2 diabetes. Mol Cell Biochem 372:191–198
124. Kobayashi MA, Watada H, Kawamori R, Maeda S (2010) Overexpression of acetyl-coenzyme A carboxylase beta increases proinflammatory cytokines in cultured human renal proximal tubular epithelial cells. Clin Exp Nephrol 14:315–324
125. McDonough CW, Palmer ND, Hicks PJ, Roh BH, An SS, Cooke JN, Hester JM, Wing MR, Bostrom MA, Rudock ME, Lewis JP, Talbert ME, Blevins RA, Lu L, Ng MC, Sale MM, Divers J, Langefeld CD, Freedman BI, Bowden DW (2011) A genome-wide association study for diabetic nephropathy genes in African Americans. Kidney Int 79:563–572
126. Iyengar SK, Sedor JR, Freedman BI, Kao WH, Kretzler M, Keller BJ, Abboud HE, Adler SG, Best LG, Bowden DW, Burlock A, Chen YD, Cole SA, Comeau ME, Curtis JM, Divers J, Drechsler C, Duggirala R, Elston RC, Guo X, Huang H, Hoffmann MM, Howard BV, Ipp E, Kimmel PL, Klag MJ, Knowler WC, Kohn OF, Leak TS, Leehey DJ, Li M, Malhotra A, März W, Nair V, Nelson RG, Nicholas SB, O'Brien SJ, Pahl MV, Parekh RS, Pezzolesi MG, Rasooly RS, Rotimi CN, Rotter JI, Schelling JR, Seldin MF, Shah VO, Smiles AM, Smith MW, Taylor KD, Thameem F, Thornley-Brown DP, Truitt BJ, Wanner C, Weil EJ, Winkler CA, Zager PG, Igo RP Jr, Hanson RL, Langefeld CD, Family Investigation of Nephropathy and Diabetes (FIND) (2015) Genome-wide association and trans-ethnic meta-analysis for advanced diabetic kidney disease: Family Investigation of Nephropathy and Diabetes (FIND). PLoS Genet 11:e1005352
127. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M,

Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Raymond C, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie WR, de la Bastide M, Dedhia N, Blöcker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kaspryzk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AF, Stupka E, Szustakowski J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrinos A, Morgan MJ, de Jong P, Catanese JJ, Osoegawa K, Shizuya H, Choi S, Chen YJ, International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921

128. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, Di Francesco V, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue B, Sun J, Wang Z, Wang A, Wang X, Wang J, Wei M, Wides R, Xiao C, Yan C, Yao A, Ye J, Zhan M, Zhang W, Zhang H, Zhao Q, Zheng L, Zhong F, Zhong W, Zhu S, Zhao S, Gilbert D, Baumhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An H, Awe A, Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K, Busam D, Carver A, Center A, Cheng ML, Curry L, Danaher S, Davenport L, Desilets R, Dietz S, Dodson K, Doup L, Ferriera S, Garg N, Gluecksmann A, Hart B, Haynes J, Haynes C, Heiner C, Hladun S, Hostin D, Houck J, Howland T, Ibegwam C, Johnson J, Kalush F, Kline L, Koduru S, Love A, Mann F, May D, McCawley S, McIntosh T, McMullen I, Moy M, Moy L, Murphy B, Nelson K, Pfannkoch C, Pratts E, Puri V, Qureshi H, Reardon M, Rodriguez R, Rogers YH, Romblad D, Ruhfel B, Scott R, Sitter C, Smallwood M, Stewart E, Strong R, Suh E, Thomas R, Tint NN, Tse S, Vech C, Wang G, Wetter J, Williams S, Williams M, Windsor S, Winn-Deen E, Wolfe K, Zaveri J, Zaveri K, Abril JF, Guigó R, Campbell MJ, Sjolander KV, Karlak B, Kejariwal A, Mi H, Lazareva B, Hatton T, Narechania A, Diemer K, Muruganujan A, Guo N, Sato S, Bafna V, Istrail S, Lippert R, Schwartz R, Walenz B, Yooseph S, Allen D, Basu A,

Baxendale J, Blick L, Caminha M, Carnes-Stine J, Caulk P, Chiang YH, Coyne M, Dahlke C, Mays A, Dombroski M, Donnelly M, Ely D, Esparham S, Fosler C, Gire H, Glanowski S, Glasser K, Glodek A, Gorokhov M, Graham K, Gropman B, Harris M, Heil J, Henderson S, Hoover J, Jennings D, Jordan C, Jordan J, Kasha J, Kagan L, Kraft C, Levitsky A, Lewis M, Liu X, Lopez J, Ma D, Majoros W, McDaniel J, Murphy S, Newman M, Nguyen T, Nguyen N, Nodell M, Pan S, Peck J, Peterson M, Rowe W, Sanders R, Scott J, Simpson M, Smith T, Sprague A, Stockwell T, Turner R, Venter E, Wang M, Wen M, Wu D, Wu M, Xia A, Zandieh A, Zhu X (2001) The sequence of the human genome. Science 291:1304–1351

129. The 1000 Genomes Project Consortium (2015) A global reference for human genetic variation. Nature 526:68–74
130. EMBI-EBI GWAS catalog, http://www.ebi.ac.uk/gwas/
131. Minster RL, Hawley NL, Su CT, Sun G, Kershaw EE, Cheng H, Buhule OD, Lin J, Reupena MS, Viali S, Tuitele J, Naseri T, Urban Z, Deka R, Weeks DE, McGarvey ST (2016) A thrifty variant in *CREBRF* strongly influences body mass index in Samoans. Nat Genet 48:1049–1054
132. Flannick J, Florez JC (2016) Type 2 diabetes: genetic data sharing to advance complex disease research. Nat Rev Genet 17:535–549

Chapter 5
The Association of Single Nucleotide Polymorphisms with Cancer Risk

Koichi Matsuda

Abstract Cancer is the second leading cause of death and there were 17.5 million cancer cases and 8.7 million deaths worldwide in 2015. Although cancer mortality decreased in the most of countries, cancer cases increased in the most of countries. Recent progress in medical treatment and personalized medicine have significantly improved cancer survival, however prevention and early detection of cancer are the most important approach to reducing cancer mortality. Family history is also associated with a two to fourfold increased risk of cancer in European populations, and 20–40% is expected to be explained by heritable factors. More than 250 studies have identified about 700 significant SNPs. The identification of cancer susceptibility genes contributes to our understanding of disease pathogenesis and risk prediction. Here, we reviewed recent GWAS of prostate, breast, colorectal, lung, liver, gastric, esophageal, bladder, pancreas, ovary, bone, and testicular cancers.

Keywords GWAS · Cancer · Prostate · Breast · Colorectal · Lung · Liver · Gastric · Esophagus · Bladder · Pancreas · Ovary · Bone · Testis

5.1 Introduction

Cancer is the second leading cause of death after cardiovascular disease [133]. In 2015, there were 17.5 million cancer cases and 8.7 million deaths worldwide [55]. Although cancer mortality decreased in the most of countries, cancer cases increased in the most of countries. In 2005, 14% of all deaths were due to cancer, which increased to 16% in 2015 [133]. Prostate cancer, TBL (tracheal, bronchus, and lung) cancer, and colorectal cancer were the most common cancers in men with 1.6 million, 1.4 million, and 0.92 million cases (42% of all cancer cases among men),

K. Matsuda (✉)
Laboratory of Clinical genome Sequencing, Department of Computational biology and medical Sciences, Graduate school of Frontier Sciences, The University of Tokyo, Tokyo, Japan
e-mail: koichima@ims.u-tokyo.ac.jp

T. Tsunoda et al. (eds.), *Genome-Wide Association Studies*,
https://doi.org/10.1007/978-981-13-8177-5_5

respectively. The most common causes of cancer deaths for men were TBL, liver, and stomach cancer with 1.21 million, 577,000, and 535,000 deaths, respectively. For women, the most common incident cancers were breast, colorectal, and TBL cancer, with 2.4 million, 733,000, and 640,000 (46% all incident cases among women), respectively. The leading causes of cancer deaths were breast, TBL, and colorectal cancer, 523,000, 517,000, and 376,000 deaths, respectively. Thus, cancer is a major global public health problem. Recent progress in medical treatment and personalized medicine have significantly improved cancer survival, however prevention and early detection of cancer are the most important approach to reducing cancer mortality.

Environmental carcinogens and their association with cancer were reviewed by the WHO, and more than 100 factors were shown to increase human cancer risk. In addition to these external factors, host genetic factors were shown to increase cancer risk. Family history is also associated with a two to fourfold increased risk of cancer in European populations [59, 173]. An epidemiological study using a Japanese disease biobank consisting of 200,000 patients with 47 common diseases revealed a two to sevenfold higher risk for individuals with a positive family history. Prostate cancer and ovarian cancer showed relatively high odds ratios (ORs) of 7.191 (95% confidential interval (C.I.): 6.284–8.230) and 6.547 (95% C.I.: 4.372–9.804) compared with other diseases (2.300–3.875), indicating the particularly crucial roles of genetic factors in these diseases [74]. In addition, a large scale twin study revealed an effect of heritable factors on various cancers of approximately 20–40% [114]. Because rare genetic defects (mutations) inherited from a parent are estimated to account for less than 6% of total cancers, the remaining 15–35% is expected to be explained by common genetic variations.

In 2002, the first genome-wide association study (GWAS) for myocardial infarction was conducted in a Japanese population and identified a susceptibility locus at 6p21 [138]. Subsequently, GWAS have been successfully applied to a broad range of disease types, and the NHGRI-EBI Catalog of published genome-wide association studies [200] currently lists over 2600 publications and 30,000 SNPs. GWAS have also been extensively applied to cancers, and disease-associated SNPs have been identified for the majority of cancers [116]. Here, we review recent case control studies focused on evaluating single nucleotide polymorphisms (SNPs) and the risks of various cancers.

5.2 Prostate Cancer

Prostate cancer is the most common cancer in men and its incidence has rapidly increased recently [63, 126]. In 2015, there were 1.6 million incident cases of prostate cancer and 366,000 deaths, while there were 974,000 cases in 2005. In 2015, prostate cancer was the cancer with the highest incidence for men in 103 countries,

and the leading cause of cancer deaths for men in 29 countries. Environmental factors, such as a high-fat diet, androgens, physical activity, inflammation, and obesity, may play some important roles in prostate carcinogenesis, but their roles remain unclear [77]. Based on epidemiological evidence and twin studies, a genetic component contributes to its etiology. Approximately 42% of the risk of prostate cancer is accounted for by genetic factors [114]. A family history of prostate cancer doubles the risk of disease development in first-degree relatives [203]. In a Japanese population, a positive family history was associated with as much as a sevenfold increased risk of prostate cancer [74]. Linkage and genetic sequencing studies identified rare moderate- to high-risk gene loci, such as HOXB13 [43] and BRCA1/2 [110, 184], which predispose an individual to prostate cancer when mutated. In addition to these cancer predisposition genes, GWAS have identified more than 100 common SNPs which confer a risk of prostate cancer development with an increasing number of risk alleles [6, 15, 42].

In 2006, a genome-wide linkage scan using 1068 microsatellite markers typed for 871 Icelandic men identified a prostate cancer susceptibility locus on chromosome 8q24 that showed the strongest association (OR of 1.62 and $P = 2.7 \times 10^{-11}$ for DG8S737-8 and OR of 1.51 and $P = 1.0 \times 10^{-11}$ for rs1447295). In 2007, two groups performed the first GWAS of prostate cancer using multiple SNPs as genetic markers and identified a susceptibility locus on chromosome 8q24. In the study by Gudmundsson J et al. in Iceland, 1453 prostate cancer cases and 3064 controls were analyzed using the Illumina HumanHap300 BeadChip, followed by four replication studies [66]. In the study by Yeager et al., 550,000 SNPs were screened in 1172 cases and 1157 controls of European origin and the association at 8q24 was analyzed using four additional sample sets (4296 cases and 4299 controls), confirming the association of the locus at 8q24 with prostate cancer [213]. In 2007, another GWAS of 1501 Icelandic men with prostate cancer and 11,290 controls and follow-up studies identified an association of two SNPs on chromosome 17 with prostate cancer. These two variants are located within a region previously implicated in prostate cancer by family-based linkage studies [67].

Prostate cancer GWAS were also reported in Asian populations. In 2010, a GWAS and replication study of 4584 Japanese men with prostate cancer and 8801 control subjects identified five new loci for prostate cancer at 5p15.13 ($P = 3.9 \times 10^{-18}$), 6q22.1 (GPRC6A/RFX6, $P = 1.6 \times 10^{-12}$), 13q22 ($P = 2.8 \times 10^{-9}$), 2p24.1 (C2orf43, $P = 7.5 \times 10^{-8}$) and 6p21.1 (FOXP4, $P = 7.6 \times 10^{-8}$) [179]. In addition, a Chinese group reported two novel prostate cancer risk loci at 9q31.2 and 19q13.4 in a GWAS of 4484 prostate cancer cases and 8934 controls [211].

In 2014, a comprehensive meta-analysis of 43,303 prostate cancer cases and 43,737 controls from studies of individuals of European, African, Japanese, and Latino ancestry was reported [7]. Twenty-three new susceptibility loci were identified with an association of $P < 5 \times 10^{-8}$. Currently, GWAS have yielded approximately 100 prostate cancer risk SNPs, accounting for 33% of the relative familial risk in European populations. About 30 GWAS studies were reported so far (Table 5.1).

Table 5.1 List of prostate cancer GWAS

Screening (case/control)	Replication (case/control)	Screening_ethnicity	Journal	Year	PUBMEDID	First author	Identified regions
1172/1157	3124/3142	EU	Nat Genet	2007	17401363	Yeager M	8q24.21
1453/3064	1583/2817	EU	Nat Genet	2007	17401366	Gudmundsson J	8q24.21
1501/11290	1992/3058	EU	Nat Genet	2007	17603485	Gudmundsson J	17q12, 17q24.3
59/558		EU	BMC Med Genet	2007	17903305	Murabito JM	
1235/1599	1242/917	EU	J Natl Cancer Inst	2007	18073375	Duggan D	
1172/1157	3941/3964	EU	Nat Genet	2008	18264096	Thomas G	8q24.21, 10q11.22, 11q13.3, 17q12
1854/1894	3268/3366	EU	Nat Genet	2008	18264097	Eeles RA	3p12.1, 6q25.3, 7q21.3, 8q24.21, 10q11.22, 11q13.3, 17q12, 19q13.33, Xp11.22
1854/21372	8239/7590	EU	Nat Genet	2008	18264098	Gudmundsson J	2p15, Xp11.22
(1) 1235/1599	5383/3439	EU	Cancer Res	2009	19117981	Sun J	
1854/1894	19879/18761	EU	Nat Genet	2009	19767753	Eeles RA	(4)
1968/35382	11806/12387	EU	Nat Genet	2009	19767754	Gudmundsson J	3q21.3, 8q24.21, 11q13.3, 19q13.2
1583/3386	3001/5415	JP	Nat Genet	2010	20676098	Takata R	3p11.2, 5p15.33, 6q22.1, 8p21.2, 8q24.21, 10q11.22, 13q22.1, 17q12
(1) 202/100	1122/1167	EU	CANCER EPIDEM BIOMAR	2011	21467234	FitzGerald LM	
3425/3290	1844/3269	AA	Nat Genet	2011	21602798	Haiman CA	17q21.33
2782/4458	7358/6732	EU	Hum Mol Genet	2011	21743057	Schumacher FR	2q37.3, 6q25.3, 8q24.21, 11q13.3, 12q13.13, 17q24.3
6621/6939	30629/29420	EU	Nat Genet	2011	21743467	Kote-Jarai Z	2p11.2, 2q37.3, 3q23, 3q26.2, 5p15.33, 5p12, 6p21.33, 12q13.12, Xq12
316/229	1663/1776	EU	Cancer Biol Ther	2011	22130093	Nam RK	10q26.12, 15q21.1

2076/2099	3437/7134	JP, LA	CANCER EPIDEM BIOMAR	2012	22923026	Cheng I	8q24.21
1417/1008	3067/7926	CN	Nat Genet	2012	23023329	Xu J	8q24.21, 9q31.2, 19q13.42
(2) 11085/11463	24395/24726	EU	Hum Mol Genet	2012	23065704	Amin Al Olama A	19q13.2
5141/54444	5470/7583	EU	Nat Genet	2012	23104005	Gudmundsson J	8q24.21
11085/11463	19662/19715	EU	Nat Genet	2013	23535732	Eeles RA	(5)
90/131	155/182	AF	J Transl Med	2013	23668334	Shan J	
474/458	5096/4972	AF	Hum Genet	2013	24185611	Cook MB	
(3) 931/4120	2571/921	EU	PLoS One	2014	24740154	Lange EM	8q24.21, 11p15.5
1146/1804	1854/1437	EU	CANCER EPIDEM BIOMAR	2014	24753544	Knipe DW	8q24.21, 10q11.22, 17q24.3, 19q13.33
43303/43737		EU, AF, JP, LA	Nat Genet	2014	25217961	Al Olama AA	(6)
7783/38595	6864/5354	EU, AA, EA, LA	Cancer Discov	2015	26034056	Hoffmann TJ	(7)
3000/4394	3605/3919	EA	Nat Commun	2015	26443449	Wang M	5p15.33, 6q22.1, 8p21.2, 8q24.21, 10q11.22, 11p15.4, 14q23.2

EU European, *AA* African american, *AS* Asian, *EA* East Asian, *CN* Chinese, *JP* Japanese, *KR* Korean, *LA* Larino/Hispanic, *AF* African. (1) aggressive, (2) non-aggressive/aggressive, (3) early onset, (4) 2p21, 2q31.1, 3p11.2, 4q22.3, 4q24, 6q25.3, 7q21.3, 8p21.2, 8q24.21, 10q11.22, 11p15.5, 11q13.3, 17q12, 17q24.3, 22q13.2, Xp11.22, (5) 1q21.3, 1q32.1, 2p25.1, 2q37.3, 3q13.2, 4q13.3, 5q35.2, 6p21.32, 6q21, 6q25.2, 7p15.3, 8p21.2, 10q24.32, 11q22.2, 12q24.21, 14q22.1, 14q24.1, 17p13.3, 17q21.32, 18q23, 20q13.33, Xp22.2, (6) 1p36.22, 1q21.3, 1q32.1, 2p25.1, 4q13.3, 6p24.2, 6p22.1, 6p21.32, 6q14.1, 7p12.3, 9p21.3, 10q11.22, 11q23.2, 12q13.11, 14q23.1, 14q24.2, 16q22.2, 20q13.13, 21q22.3, 22q11.21, Xp11.22, Xq13.1, (7) 2p11.2, 4q24, 6q25.3, 7p15.1, 8p21.2, 8q24.21, 10q11.22, 11q13.3, 12q13.12, 12q13.13, 17q12, 17q24.3, 19q13.33, 22q13.2

5.3 Breast Cancer

Breast cancer is the most common cancer in women worldwide with an estimated 2.4 million cases in 2015 [55]. The majority of breast cancer occurred in women, with 2.4 million cases vs 44,000 cases in men, and more than 10% of women will be diagnosed with breast cancer in their lifetime [93]. For women, breast cancer was the leading cause of death (523,000 death in 2015). According to a large scale twin study, genetic factors account for 27% of breast cancer risk [114]. Approximately 10% of patients with breast cancer have a family history of breast cancer [144]. Compared with women without a family history, women with a premenopausal first-degree relative with breast cancer are at a 3.3-fold greater risk. In a Japanese population, a positive family history was associated with a 3.3-fold increased risk of breast cancer [74], indicating that germline transmission significantly contributes to risk [172]. The disease aggregates in families, indicating an important role of genetic factors in breast cancer etiology [127, 217]. This inherited component is driven by rare variants, notably in *BRCA1*, *BRCA2*, *PTEN*, *TP53*, *PALB2*, *STK11*, *ATM* and *CHEK2*, conferring a high lifetime risk of the disease. These cancer predisposition genes account for less than 25% of the familial risk of breast cancer [40]. Therefore, the remaining heritable breast cancer risk (approximately 20%) would be caused by common variants with modest effects. Since 2007, genome-wide association studies (GWAS) have identified approximately 100 common genetic susceptibility loci for breast cancer risk (Table 5.1). In 2007, three groups reported breast cancer susceptibility loci at 10q26 [82], 16q12 [176], 5q11, 8q24, 11p15, and 2q35 [41, 82, 176]. Easton et al. conducted a two-stage genome-wide association study of 4398 breast cancer cases and 4316 controls in a European population, followed by a third stage using 21,860 cases and 22,578 controls and identified five novel independent loci that exhibited strong and consistent evidence of an association with breast cancer ($P < 10^{-7}$) [41]. Four of these loci contain plausible causative genes (*FGFR2, TNRC9, MAP3K1* and *LSP1*).

A GWAS of an Asian population also identified breast cancer susceptibility loci. A GWAS of Chinese women analyzed 607,728 SNPs in 1505 cases and 1522 controls, and 29 SNPs for fast-track replication in an independent set of 1554 cases and 1576 controls. Further analysis identified SNP rs2046210 at 6q25.1, with $P = 2.0 \times 10^{-15}$ and OR of 1.36.

In 2015, a meta-analysis of 11 GWAS comprising 15,748 breast cancer cases and 18,084 controls together with 46,785 cases and 42,892 controls from 41 studies genotyped on a 211,155-marker custom array (iCOGS) identified 15 new loci associated with breast cancer ($P < 5 \times 10^{-8}$) in individuals of European ancestry [129]. To date, nearly 100 genetic risk variants have been identified in these studies that explain approximately 16% of the familial breast cancer risk in European descendants.

In addition to genetic factors related to sporadic breast cancer, genetic modifiers of BRCA1- and BRCA2-related breast cancer were also identified. A genome-wide association study of 1193 individuals with BRCA1 mutations who were diagnosed

with invasive breast cancer under age 40 and 1190 BRCA1 carriers without breast cancer (over age 35) identified five SNPs on chromosome 19p13 that were associated with breast cancer risk ($P = 2.3 \times 10^{-9}$ to $P = 3.9 \times 10^{-7}$) [10]. Genotyping of these SNPs in 6800 population-based breast cancer cases and 6613 controls identified a similar association with estrogen receptor-negative breast cancer (OR = 0.83, P = 0.0003) and an opposite association with estrogen receptor-positive disease (OR = 1.07, P = 0.016). A subsequent GWAS of BRCA1 mutation carriers identified novel loci associated with breast and ovarian cancer risk [30]. A multi-stage GWAS of 11,705 BRCA1 carriers, including 5920 breast cancer cases and 1839 ovarian cancer cases, with a further replication in an additional sample of 2646 BRCA1 carriers identified a novel breast cancer risk modifier locus at 1q32 (rs2290854, $P = 2.7 \times 10^{-8}$, HR = 1.14). BRCA1 breast cancer risk-modifying loci could enable us to estimate breast cancer lifetime risks from 28% to 50% for a low risk group (5% of total BRCA1 carriers) and 81–100% for a high risk group (TOP 5% of total BRCA1 carrier). This estimation may have important implications for the clinical management of BRCA1 carriers. Thirty six studies were reported so far (Table 5.2).

5.4 Colorectal Cancer

Colorectal cancer (CRC) is the third most common cancer and the fourth leading cause of cancer-related death worldwide. There are 1.7 million new CRC cases and 832,000 deaths per year. The odds of developing colon and rectum cancer before age 79 years at the global level was higher for men than for women (1 in 28 men, 1 in 43 women) [55]. A Westernized lifestyle, such as obesity, sedentary behavior, and a high-meat, high-calorie, fat-rich, and fiber-deficient diet, has been linked to an increased colorectal cancer risk [16, 124]. In addition, nearly 15% of patients with CRC have a positive family history of the disease [22, 50], and family history is acknowledged to be one of the strong risk factors. An approximately twofold increased risk of CRC was observed among patients who have a first-degree relative with CRC [90]. In a Japanese population, a positive family history was associated with a 2.4-fold increased risk of colorectal cancer [74]. Although inherited susceptibility is thought to account for ~35% of all CRC cases [114], high-risk germline mutations in *APC*, DNA mismatch repair genes (*MLH1, MSH2, MSH6, PMS2*), *MUTYH, SMAD4, BMPR1A*, and *LKB1* account for <6% of all cases [2]. Therefore, the remaining heritable CRC risk (approximately 30%) would be caused by the combination of common variants with modest effects. More than 20 GWAS of CRC have successfully identified common SNPs associated with CRC risk [20, 31, 39].

In 2007, the first GWAS for colorectal cancer was reported by two groups and identified 8q24 as a susceptibility locus [186, 215]. In these analyses, approximately 1000 cases and control samples (one study used both patients with colorectal cancer and advanced adenoma as the case group) were genotyped for 100,000–550,000 tagged SNPs. A further replication analysis identified a significant association of SNPs at 8q24 with colorectal cancer risk, with $P = 1.27 \times 10^{-14}$ and $P = 3.16 \times 10^{-11}$

Table 5.2 List of breast cancer GWAS

Screening (case/control)	Replication (case/control)	Screening_ethnicity	Journal	Year	PUBMEDID	First author	Identified regions
390/364	26646/24889		Nature	2007	17529967	Easton DF	5q11.2, 8q24.21, 10q26.13, 11p15.5, 16q12.1
(1) 1145/1142	1176/2072	EU	Nat Genet	2007	17529973	Hunter DJ	10q26.13
(2) 1599/11546	4503/7695	EU	Nat Genet	2007	17529974	Stacey SN	2q35, 16q12.1
58/665		EU	BMC Med Genet	2007	17903305	Murabito JM	
(3) 249/299	1193/1166	EU	Proc Natl Acad Sci U S A	2008	18326623	Gold B	6q22.33
30/30		EU, LA, AA	Breast Cancer Res Treat	2008	18463975	Kibriya MG	
1505/1522	6617/3942	CN	Nat Genet	2009	19219042	Zheng W	6q25.1
1145/1142	8625/9657	EU	Nat Genet	2009	19330030	Thomas G	1p11.2, 2q35, 10q26.13, 16q12.1
3659/4897	12576/12223	EU	Nat Genet	2010	20453838	Turnbull C	(12)
2073/2084	17956/15655	CN	PLoS Genet	2010	20585626	Long J	16q12.1
(4) 1193/1190	2974/3012	EU	Nat Genet	2010	20852631	Antoniou AC	19p13.11
2702/5726	7386/7576	EU	Breast Cancer Res Treat	2010	20872241	Li J	10q26.13
(5) 899/804	1264/1222	EU	PLoS Genet	2010	21060860	Gaudet MM	10q26.13
(6) 617/4583	1011/7604	EU	Breast Cancer Res	2010	21062454	Li J	
2839/3507	9041/8980	EU	J Natl Cancer Inst	2011	21263130	Fletcher O	2q35, 3p24.1, 5p12, 8q24.21, 9q31.2, 10q26.13, 16q12.2
302/321	1153/1215	EU	Hum Genet	2011	21424380	Sehrawat B	
2062/2066	15091/14877	EA	Hum Mol Genet	2011	21908515	Cai Q	10q21.2
(6) 2722/6415	2292/16901	AA, EU	Nat Genet	2011	22037553	Haiman CA	5p15.33
2918/2324	16173/18282	CN	PLoS Genet	2012	22383897	Long J	6q25.1
2273/2052	4049/3845	KR	Breast Cancer Res	2012	22452962	Kim HC	2q34

3016/2745	3533/11046	AA	Hum Genet	2012	22923054	Chen F	
(7) 1086/1816	1653/2797	JP	J Hum Genet	2012	22951594	Elgazzar S	
4670/31608	946/8404	EU, AA	Hum Mol Genet	2012	22976474	Siddiq A	6q25.1, 19p13.11, 20q11.22
(8) 823/2795	438/474	EU	Nat Genet	2012	23001122	Orr N	14q24.1, 16q12.1
(9) 477/524	203/263	EU	Hum Genet	2013	23354978	Rinella ES	
3016/2745		AA	PLoS One	2013	23468962	Song C	
10052/12575	45290/41880	EU	Nat Genet	2013	23535729	Michailidou K	(13)
(6) 4193/35194	6514/41455	EU	Nat Genet	2013	23535733	Garcia-Closas M	(14)
(4) 1426/1301	6031/5933	EU	PLoS Genet	2013	23544013	Couch FJ	6q25.1, 19p13.11
2642/2099	2885/3395	JP	PLoS One	2013	24143190	Low SK	10q26.13, 16q12.1, 16q12.2
(10) 1529/3399	2148/1309	EU	Carcinogenesis	2013	24325915	Purrington KS	12p11.22, 19p13.11
(11) 3523/2702	3470/5475	EU	CANCER EPIDEM BIOMAR	2014	24493630	Ahsan H	3p24.1, 5q11.2, 8q24.21, 10q26.13, 11q13.3, 16q12.1
5113/4337	33670/61179	CN, KR	Nat Genet	2014	25038754	Cai Q	1q32.1, 5q14.3, 15q26.1
1497/3213	1643/4971	LA	Nat Commun	2014	25327703	Fejerman L	6q25.1, 16q12.1
89/46		EA	Asian Pac J Cancer Prev	2015	25824743	Haryono SJ	
1367/1658	61686/62327	EU	BMC Cancer	2015	25956309	Palomba G	

EU European, *AA* African american, *AS* Asian, *EA* East Asian, *CN* Chinese, *JP* Japanese, *KR* Korean, *LA* Larino/Hispanic, *AF* African., (1) postmenopausal, (2) ER-positive, (3) non-BRCA1/2 carriers, (4) BRCA1 carriers,(5) BRCA2 carriers, (6) ER negative, (7) hormon receptor positive, (8) male, (9) With strong family history, (10) hormon receptor negative, (11) early onset, (12) 2q35, 5q11.2, 9p21.3, 10q21.2, 10q22.3, 10q26.13, 11q13.3, 16q12.1, (13) 1p36.22, 1p13.2, 1p11.2, 2q14.2, 2q31.1, 2q35, 3p26.1, 3p24.1, 4q24, 4q34.1, 5p15.33, 5p12, 5q11.2, 5q33.3, 6p25.3, 6p23, 6q14.1, 6q25.1, 7q35, 8p12, 8q21.13, 8q24.21, 9q31.2, 10p12.31, 10q21.2, 10q22.3, 10q25.2, 10q26.12, 10q26.13, 11p15.5, 11q13.1, 11q13.3, 11q24.3, 12p13.1, 12p11.22, 12q22, 12q24.21, 13q13.1, 14q13.3, 14q24.1, 14q32.11, 16q12.1, 16q12.2, 16q23.2, 17q22, 18q11.2, 19p13.11, 19q13.31, 21q21.1, 22q12.2, 22q13.1, 14) 1p36.22, 1q32.1, 2p24.1, 5p15.33, 6q25.1, 12p11.22, 16q12.1, 16q12.2, 19p13.11

and ORs of 1.27 and 1.17, respectively. In addition, another study of 1477 colorectal adenoma cases and 2136 controls suggests that susceptibility to CRC is mediated by the development of adenomas (OR = 1.21; $P = 6.89 \times 10^{-5}$).

In 2008, a genome-wide association study analyzing 550,163 tagged SNPs in 940 individuals with familial colorectal tumors (627 CRC and 313 advanced adenomas) and 965 controls and subsequent replication analyses (7473 cases and 5984 controls) identified an association of SNP rs4939827 located on chromosome 18q21.1 (SMAD7) with CRC ($P = 1.0 \times 10^{-12}$). Subsequently, many studies have identified multiple CRC loci at 11q23 [182], 14q22.2, 16q22.1, 19q13.1 and 20p12.3 [178], 15q13.3 [188], 10p14 and 8q23.3 [83], 1q41, 3q26.2, 12q13.13 and 20q13.33 [76]. Most of these studies were conducted using European subjects.

Regarding other ethnic populations, a Japanese group conducted the first GWAS of an Asian population and identified an association of 6q26-q27 with distal colon cancer in 2011 [31]. A GWAS and sub-analyses by tumor location of 1583 Japanese CRC cases and 1898 controls and subsequent replication analyses of 4809 CRC cases and 2973 controls, including Korean subjects, identified a novel locus on 6q26-q27 ($p = 7.92 \times 10^{-9}$, OR of 1.28). In 2013, a GWAS of 2098 Chinese cases and 5749 controls and a replication analysis of East Asians, including up to 5358 cases and 5922 controls, was reported [86]. Three of the four loci were replicated in 26,060 individuals of European descent, with combined P values of 1.22×10^{-10} for 5q31.1, 6.64×10^{-9} for 20p12.3 and 3.06×10^{-8} for 12p13.32. In 2016, Schimit et al. reported a GWAS of CRC in Hispanics (1611 CRC cases and 4330 controls). The authors identified four suggestive associations, although the associations were not statistically significant.

In addition to these analyses, several studies of European and East Asian individuals identified other CRC loci at 14q22.2, 20p12.3 [187], 6p21, 11q13.4, Xp22.2 [38], 2q32.3 [142], 10q24.2 [202], 10q22.3, 10q25.2, 11q12.2, 12p13.31, 17p13.3, 19q13.2 [218], 4q32.2 [164], and 10q25 [196]. Among these loci, the following loci were validated by multiple studies; 5q31.1, 8q23.3, 8q24.21, 10p14, 11q23, 12p13.32, 12q13.13, 14q22.2, 15q13.3, 16q22.1, 16q22, 18q21.1, 19q13.1, 20p12.3, 20q13.33, and Xp22.2. Thus, a further meta-analysis of multiple studies with various ethnic backgrounds would identify new CRC genetic factors and would contribute to the elucidation of the molecular pathogenesis of CRC and improve risk prediction. List of 24 CRC GWAS was shown in Table 5.3.

Lung Cancer

Lung cancer is the most common cause of cancer-related death worldwide, with over 1.7 million deaths annually. Men were more likely to develop lung cancer than women, with 1 in 18 men and 1 in 45 women developing tracheal, bronchus, and lung (TBL) cancer between birth and age 79 years [55]. Between 2005 and 2015, TBL cancer cases increased by 29%. Cigarette smoke, including secondhand smoke, is associated with disease risk and a substantially elevated risk of mortality [220]. Lung cancer types are histologically classified as small cell lung cancer (SCC) and

Table 5.3 List colorectal cancer GWS

Screening (case/control)	Replication (case/control)	Screening_ethnicity	Journal	Year	PUBMEDID	First author	Identified regions
1257/1336	6223/6443	EU	Nat Genet	2007	17618283	Zanke BW	8q24.21
930/960	7334/5246	EU	Nat Genet	2007	17618284	Tomlinson I	8q24.21
930/960	7473/5984	EU	Nat Genet	2007	17934461	Broderick P	18q21.1
981/1002	16476/15351	EU	Nat Genet	2008	18372901	Tenesa A	8q24.21, 11q23.1, 18q21.1
922/927	17872/17526	EU	Nat Genet	2008	18372905	Tomlinson IP	8q23.3, 8q24.21, 10p14
1902/1929	18284/18926	EU	Nat Genet	2008	19011631	Houlston RS	14q22.2, 16q22.1, 19q13.11, 20p12.3
371/1263	4915/8159	EU	Carcinogenesis	2010	20610541	Lascorz J	
3334/4628	14851/15569	EU	Nat Genet	2010	20972440	Houlston RS	1q41, 3q26.2, 12q13.12, 20q13.33
1583/1898	4809/2973	JP	Gut	2011	21242260	Cui R	6q25.3, 8q24.21
2906/3416	8161/9101	EU	Hum Genet	2011	21761138	Peters U	15q13.3
8323/9457	21096/19555	EU	Nat Genet	2012	22634755	Dunlop MG	6p21.2, 11q13.4, Xp22.2
2098/5749	31418/5922	EA	Nat Genet	2012	23263487	Jia WH	5q31.1, 12p13.32, 20p12.3
12696/15113	3056/6658	EU	Gastroenterology	2012	23266556	Peters U	2q32.3, 8q24.21, 18q21.1
882/473	1436/1780	EU	BMC Genomics	2013	23350875	Fernandez-Rozadilla C	
1773/2642	6902/7862	EA	Int J Cancer	2014	24448986	Zhang B	18q21.1
5626/7817	14037/15937	EU	Hum Mol Genet	2014	24737748	Whiffin N	(2)
2098/6172	29849/44035	EA	Nat Genet	2014	24836286	Zhang B	(3)
480/801	1305/2049	EU	PLoS One	2014	24978480	Real LM	
2462/1497	1131/831	EU, EU	Carcinogenesis	2014	25023989	Schmit SL	4q32.2
4520/8500	16823/18211	JP, AA	Nat Commun	2014	25105248	Wang H	10q25.2
7577/9979		EU	Sci Rep	2015	25990418	Al-Tassan NA	(4)

(continued)

Table 5.3 (continued)

Screening (case/control)	Replication (case/control)	Screening_ethnicity	Journal	Year	PUBMEDID	First author	Identified regions
18299/19656	4725/9969	EU	Nat Commun	2015	26151821	Schumacher FR	(5)
(1) 5725/13396		EU	Sci Rep	2015	26621817	Cheng TH	
1611/4330		LA	Carcinogenesis	2016	27207650	Schmit SL	

EU European, *AA* African american, *AS* Asian, *EA* East Asian, *CN* Chinese, *JP* Japanese, *KR* Korean, *LA* Larino/Hispanic, *AF* African. (1) CRC/endometrial, (2) 8q23.3, 8q24.21, 10p14, 10q24.2, 12q13.12, 15q13.3, 20q13.33, (3) 1q41, 5q31.1, 8q23.3, 8q24.21, 10p14, 10q22.3, 10q25.2, 11q12.2, 12p13.32, 12p13.31, 17p13.3, 18q21.1, 19q13.2, 20p12.3, (4) 1p36.12, 8q24.11, 8q24.21, 10p14, 11q13.4, 12q13.12, 15q13.3, 16p13.2, 16q24.1, 18q21.1, 20q13.33, (5) 3p22.1, 3p14.1, 8q23.3, 8q24.21, 10q24.2, 11q23.1, 12q24.12, 12q24.22, 14q22.2, 15q13.3, 18q21.1, 20q13.13

non-small cell lung cancer, which includes adenocarcinoma (ADC) and squamous cell carcinoma (SQC) [72]. SCC and SQC are strongly associated with smoking, whereas ADC is relatively common among female non-smokers [174], indicating differences in the molecular pathogenesis among the histological types. Lung cancer has an important heritable component, and a positive family history is associated with an approximately twofold higher risk of lung cancer [125]. In a Japanese population, a positive family history was associated with a 2.4-fold increased risk of lung cancer [74]. Therefore, identifying genes associated with lung cancer risk may suggest chemoprevention targets or identify groups at high risk. Several GWAS reported that inherited genetic factors increase the risk of lung cancer [80, 81, 107, 130, 168, 169, 198, 199] (Table 5.1).

In 2008, a genome-wide association study of 317,139 single-nucleotide polymorphisms in 1989 lung cancer cases and 2625 controls and replication studies comprising an additional 2513 lung cancer cases and 4752 controls identified a locus on chromosome 15q25 that was strongly associated with lung cancer ($P = 5 \times 10^{-20}$) [81]. The associated region contains several genes that encode nicotinic acetylcholine receptor subunits (*CHRNA5*, *CHRNA3* and *CHRNB4*). A nonsynonymous variant in *CHRNA5* (D398N) is one of the strongest disease associations, providing compelling evidence that a locus at 15q25 predisposes individuals to lung cancer.

In addition, a GWAS of lung cancer comparing 511,919 SNP genotypes in 1952 cases and 1438 controls identified two novel loci at 6p21.33 (*BAT3-MSH5*; $P = 4.97 \times 10^{-10}$) and 5p15.33 (*CLPTM1L*; $P = 7.90 \times 10^{-9}$) [198]. In the analysis of Asian populations, 3q28, 5p15.33, 6p21, and 17q24.2 were shown to be associated with ADC risk in Japanese and/or Korean populations [130, 168]. In addition, loci at 5q32, 10p14, 13q12.12, 20q13.2, and 22q12.2 are associated with lung cancer risk in the Chinese population [36, 80] and loci at 10q25 and 6p21 are associated with susceptibility to lung cancer in Asian females who have never smoked [106]. Loci at 12p13.33 and 12q23.1 are associated with SQC risk in individuals of European ancestry [166] and in the Chinese population [37].

In addition to lung cancer susceptibility loci, GWAS of smoking behavior have also been reported. A meta-analysis of more than 200,000 individuals confirmed an effect of loci at 15q25 (rs1051730, beta = 1.03, $P = 2.8 \times 10^{-73}$), 10q25 (rs1329650, beta = 0.367, $P = 5.7 \times 10^{-10}$), and 9q13 (rs3733829, beta = 0.333, $P = 1.0 \times 10^{-8}$) on smoking quantity. In addition, loci at 11p14.1 (rs6265, OR = 1.06, $P = 1.8 \times 10^{-8}$) and 9q34.2 (rs3025343, OR = 1.12, $P = 3.6 \times 10^{-8}$) were significantly associated with smoking initiation and smoking cessation, respectively [1]. In addition, loci at 19q13 and 8p11.21 were shown to be associated with smoking behavior [185].

More than 20 GWAS have currently identified nearly 30 genetic factors associated with lung cancer predisposition. However, the effect sizes of each variant were relatively small (<1.3 per allele) and the results are not consistent among different ethnic groups. This inconsistency may be partially explained by the differences in allele frequency and genetic/environmental backgrounds. Therefore, additional studies of larger numbers of subjects with different ethnic backgrounds are required to elucidate common and population-specific genetic factors. List of 22 lung cancer GWAS was shown in Table 5.4.

Table 5.4 List of lung cancer GWAS

Screening (case/control)	Replication (case/control)	Screening_ethnicity	Journal	Year	PUBMEDID	First author	Identified regions
338/335	265/356	EU	Cancer Lett	2007	17223258	Spinola M	
1154/1137	2724/3694	EU	Nat Genet	2008	18385676	Amos CI	15q25.1
1926/2522	2513/4752	EU, other	Nature	2008	18385738	Hung RJ	15q25.1
(1) 482/522		EU	Int J Cancer	2008	18729187	Galvan A	
194/219	3878/4831	EU	J Natl Cancer Inst	2008	18780872	Liu P	15q25.1
5095/5200	2448/2983	EU	Nat Genet	2008	18978787	Wang Y	5p15.33, 6p21.33, 15q25.1
2971/3746	2899/5573	EU	Nat Genet	2008	18978790	McKay JD	15q25.1
1952/1438	5608/6767	EU	Cancer Res	2009	19654303	Broderick P	5p15.33, 6p21.33, 15q25.1
(1) 5739/5848	7561/13818	EU	Am J Hum Genet	2009	19836008	Landi MT	5p15.33, 6p22.1, 6p21.33, 15q25.1
377/377	511/1007	EU	Lancet Oncol	2010	20304703	Li Y	
(1) 584/585	2184/2515	EA	PLoS Genet	2010	20700438	Hsiung CA	5p15.33
(1) 1004/1900	1094/9148	JP	Nat Genet	2010	20871597	Miki D	3q28, 5p15.33
(2) 621/1541	804/1470	KR	Hum Mol Genet	2010	20876614	Yoon KA	3q29
2331/3077	6313/6409	CN	Nat Genet	2011	21725308	Hu Z	3q28, 5p15.33, 13q12.12, 22q12.2
(2) 445/497	396/998	KR	Hum Genet	2011	21866343	Ahn MJ	
(1) 1695/5333	4334/8202	JP	Nat Genet	2012	22797724	Shiraishi K	3q28, 5p15.33, 6p21.32, 17q24.2
14900/29485	2338/3077	EU	Hum Mol Genet	2012	22899653	Timofeeva MN	5p15, 6p21, 15q25
(3) 2331/4006	2665/11436	CN	Am J Hum Genet	2012	23103227	Jin G	7p15.3

5510/4544	1099/2913	EA	Nat Genet	2012	23143601	Lan Q	3q28, 5p15.33, 6p21.32, 6q22.1, 10q25.2
(4) 833/3094	2223/6409	CN	PLoS Genet	2013	23341777	Dong J	12q23.1
11348/15861	10246/38295	EU	Nat Genet	2014	24880342	Wang Y	3q28, 13q13.1, 22q12.1
(5) 6877/6277	5878/7046	EA	Hum Mol Genet	2016	26732429	Wang Z	(6)

EU European, *AA* African american, *AS* Asian, *EA* East Asian, *CN* Chinese, *JP* Japanese, *KR* Korean, *LA* Larino/Hispanic, *AF* African. (1) ADC, (2) NSCLC, (3) lung/gastric/esophageal, (4) SCC, (5) lung/ADC, (6) 6p21.1, 9p21.3, 12q13.13, 6p21.1, 9p21.3, 12q13.13

Liver Cancer

More than 400 and 170 million people are estimated to be infected with hepatitis B virus (HBV) and hepatitis C virus (HCV) worldwide, respectively [32, 57]. Persistent HBV/HCV infections cause chronic hepatitis and subsequent fatal liver diseases, such as liver cirrhosis and hepatocellular carcinoma (HCC). HCC is the third most common cause of cancer-related death [140]. In 2015, there were 854,000 cases for liver cancer and 810,000 deaths in the world. Therefore, the treatment of HCV/HBV carriers is an issue of global importance. HBsAg seropositivity rates are as high as 5–12% in Thailand and China, but as low as 0.2–0.5% in North America and Europe [32]. Most HBV carriers were infected through maternal transmission in the neonatal period or infancy [100]. Although some HBV carriers spontaneously eliminate the virus, 2–10% of individuals with chronic hepatitis B are estimated to develop liver cirrhosis each year, and a subset of these individuals suffer from liver failure or hepatocellular carcinoma [147]. Chronic HBV infection seems to be the most important risk factor for HCC [147] [105]. Approximately 80% of individuals with HCC in China have a history of HBV infection [100]. A segregation analysis of familial HCC suggests an interaction between HBV infection and a major genetic locus [108]. In a Japanese population, a positive family history was associated with a 2.3-fold increased risk of liver cancer [74]. According to a two-stage genome-wide association study using 786 Japanese chronic hepatitis B cases and 2201 controls, chronic hepatitis B is significantly associated with *HLA-DPA1* and *HLA-DPB1*. An association of HLA-DP with chronic hepatitis B was confirmed in various ethnic groups, indicating that MHC class 2 variations play important roles in susceptibility and resistance to HBV infection.

A GWAS of HBV-related HCC (348 cases and 359 controls) in a Chinese population led to the identification of one intronic SNP (rs17401966) in *KIF1B* on chromosome 1p36.22 [219]. SNP rs17401966 lies in an approximately 244-kb linkage disequilibrium (LD) block containing *UBE4B, KIF1B,* and *PGD*. *KIF1B* encodes a kinesin superfamily member involved in the transport of organelles and vesicles. Both germline and somatic loss-of-function mutations in the *KIF1Bβ* isoform have been detected in multiple cancers.[214] Furthermore, *KIF1Bβ* was identified as a potential 1p36.2 tumor suppressor in neuroblastoma,[163] suggesting that *KIF1B* is a causative gene on 1p36.2. However, this locus was not validated in the other studies [87, 162].

The second GWAS of 1538 HBV-positive HCC patients and 1465 chronic HBV carriers [113] and subsequent analysis of four independent cohorts totaling 4431 cases and 4725 HBV carriers identified two novel associations at rs9272105 (HLA-DQA1/DRB1, OR = 1.28 and $P = 5.24 \times 10^{-22}$) and 21q21.3 (*GRIK1*, OR = 0.84 and $P = 5.24 \times 10^{-10}$). SNP rs455804 on chromosome 21q21.3 is located within intron 1 of the *GRIK1* gene. The *GRIK1* gene encodes an ionotropic glutamate receptor, GLUR5, which is involved in glutamate signaling. Glutamate plays a central role in the malignant phenotype of glioma, and inhibition of glutamate release and/or glutamate receptor activity suppresses the proliferation and invasion of various cancer

cells. Thus, the association of *GRIK1* with HCC indicates a crucial role of glutamate signaling in HCC development after HBV infection.

In the third study of 11,799 Chinese chronic HBV carriers (GWAS of 2514 chronic HBV carriers; 1161 HCC cases and 1353 controls and a 2-stage validation among 6 independent populations of chronic HBV carriers including 4319 cases and 4966 controls) identified two novel loci: rs7574865 in the *STAT4* gene ($P = 2.48 \times 10^{-10}$, OR = 1.21) and rs9275319 in the HLA-DQ ($P = 2.72 \times 10^{-17}$, OR = 1.49) [87]. The risk allele G at rs7574865 was significantly associated with lower levels of the *STAT4* mRNA in both the HCC and non-tumor tissues of 155 individuals (P = 0.0008 and 0.0002, respectively). In addition, the expression of the *STAT4* mRNA was decreased in HCC tumors compared with paired adjacent non-tumor tissues ($P = 2.33 \times 10^{-14}$). STAT family members are phosphorylated in response to cytokines and growth factors and translocate to the nucleus where they act as transcriptional activators. STAT4 is essential for mediating responses to IL-12 in lymphocytes and regulating the differentiation of T helper cells, and variations in this gene are associated with autoimmune diseases such as systemic lupus erythematosus, rheumatoid arthritis, and inflammatory bowel diseases [118, 136, 149]. Thus, *STAT4* variations would regulate the host immune response and subsequently affect HCC risk among HBV carriers.

HCV infection is present in 20–70% of individuals with HCC [195]. HCV-induced HCC is a multistep and progressive liver disease in which disease progression is influenced by both environmental and genetic risk factors. The impact of host genetic variations on the progression to chronic hepatitis C (CHC) after HCV exposure is well elucidated by GWAS [54, 183]. SNPs in the *IL28B* promoter were shown to be associated with natural HCV clearance ($P = 3 \times 10^{-13}$, OR = 2.6–3.1) [54, 183]. In addition, GWAS identified the associations of *C6orf10* (Japanese), *RNF7* and *MERTK* (European) with liver fibrosis after HCV infection [141, 193].

In 2011, a GWAS of a Japanese population analyzed 432,703 SNPs in 721 HCV-induced HCC cases and 2890 HCV-negative controls. A further analysis of 673 cases and 2596 controls identified a novel locus in the *MICA* promoter on chromosome 6p21.33 (rs2596542, $P = 4.21 \times 10^{-13}$, OR = 1.39) that was significantly associated with HCV-induced HCC. This SNP is not associated with CHC susceptibility (P = 0.61), but is significantly associated with progression from CHC to HCC ($P = 3.13 \times 10^{-8}$) [104]. MICA is a membrane protein that acts as a ligand for NKG2D to activate the anti-tumor effects of natural killer cells and $CD8^+$ T cells [14]. MICA is highly expressed on the cell surface of cancer cells and virus-infected cells. Elevated expression of both the membrane-bound and soluble forms of MICA (sMICA) have been reported in several cancers, including HCC [62, 89]. The sMICA level was elevated among patients without HCC, including patients with chronic hepatitis C, and was not correlated with disease progression. Additionally, risk allele A was correlated with low sMICA levels in subjects with HCV-induced HCC ($P = 1.38 \times 10^{-13}$). Considering the association of risk allele A with low levels of sMICA, individuals who carry the rs2596542 A allele would express low levels of membrane-bound MICA in response to HCV infection, leading to reduced or no activation of natural killer cells and $CD8^+$ T cells in response to virus-infected cells.

Thus, HCV-infected cells with low MICA expression would escape from the immune surveillance system and progress to HCC.

Another Japanese group analyzed 467,538 SNPs in 212 cases and 765 individuals with chronic HCV infection without HCC. An analysis of an independent case control population (710 cases and 1625 controls) identified an association between one intronic SNP in the *DEPDC5* gene on chromosome 22q12.2 with HCC risk (rs1012068, $P = 1.27 \times 10^{-13}$, OR = 1.75) [131]. *DEPDC5* expression is elevated in HCC tissues, and the risk allele is associated with elevated DEPDC5 expression among male subjects. These findings indicate an oncogenic role for DEPDC5 in hepatocellular carcinogenesis. List of eight liver cancer GWAS was shown in Table 5.5.

Gastric Cancer

Gastric cancer is the third leading cause of cancer mortality, and there were 1.3 million cases and 819, 000 deaths worldwide in 2015 [56, 73]. *Helicobacter pylori* infection is the major cause of gastric cancer [64, 192]. Approximately 90% of patients with gastric cancer are infected with *H. pylori*, and the eradication of *H. pylori* by antibiotics in combination with proton pump inhibitors effectively prevents the risk of gastric cancer [51], indicating a causal role for *H. pylori* in disease pathogenesis. Due to the high prevalence of *H. pylori* infections, the incidence of gastric cancer is very high in Japanese populations [55]. However, although more than 50% of individuals are infected with *H. pylori* worldwide [34], only a small subset of infected individuals develops this disease [11], indicating the presence of other factors that modify disease onset. A large scale twin study revealed that 28% of the risk of gastric cancer was explained by genetic factors [114]. In a Japanese population, a positive family history was associated with a 2.44-fold increased risk of gastric cancer [74]. Germline mutations in the *CDH1* gene that encodes the E-cadherin protein were shown to cause hereditary diffuse gastric cancer syndrome, but the overall frequency of E-cadherin germline mutation is a rare event, affecting <3% of the screened population [21]. Previous genome wide association studies (GWAS) identified genetic variations associated with gastric cancer, such as *PSCA* [160], *PLCE1* [4], *MUC1* [160], 3q13.31 and 5p13.1 [167], 6p21.1 [79, 88], and *ATM* [70] (Table 5.1).

In 2008, a GWAS of a Japanese population identified the *PSCA* gene as diffuse gastric cancer locus (rs2294008, OR = 1.62, $P = 1.11 \times 10^{-9}$). SNP rs2294008 is located in the *PSCA* promoter. *PSCA* encodes a glycosylphosphatidylinositol (GPI)-anchored membrane glycoprotein involved in cell renewal and proliferation [65]. PSCA is upregulated in various cancers, including bladder, pancreatic and kidney cancers [158], and PSCA expression is correlated with a higher tumor grade and metastatic properties of prostate cancer [65]. PSCA is also expressed in differentiating gastric epithelial cells and is frequently silenced in gastric cancer and esophageal cancer [158]. In addition, its growth-suppressive effects have also been reported

Table 5.5 List of liver cancer GWAS

Screening (case/control)	Replication (case/control)	Screening_ethnicity	Journal	Year	PUBMEDID	First author	Identified regions
(1) 348/359	2121/1748	CN	Nat Genet	2010	20676096	Zhang H	1p36.22
180/271	206/416	KR	Hepatology	2010	21105107	Clifford RJ	
(2) 721/2890	673/2596	JP	Nat Genet	2011	21499248	Kumar V	6p21.33, 6p21.32
(2) 212/765	710/1625	JP	Nat Genet	2011	21725309	Miki D	22q12.2
1) 95/97	500/728	CN	PLoS One	2011	22174901	Chan KY	
(1) 1538/1465	4431/4725	CN	PLoS Genet	2012	22807686	Li S	6p21.32, 21q21.3
(1) 1161/1353	4319/4966	CN	Nat Genet	2012	23242368	Jiang DK	2q32.2, 6p21.32
(1) 50/50	282/278	CN	Oncol Lett	2015	26870257	Qu LS	

EU European, *AA* African american, *AS* Asian, *EA* East Asian, *CN* Chinese, *JP* Japanese, *KR* Korean, *LA* Larino/Hispanic, *AF* African. (1) HBV, (2) HCV

in these cancer cells. The T allele of rs2294008 encodes a translation initiation codon for the PSCA gene upstream of the known site, resulting in a PSCA protein with an additional nine amino acids at its N-terminus (long PSCA) compared to the reported PSCA protein (short PSCA, 114 amino acids). Long PSCA contains an N-terminal signal peptide, and localizes to the plasma membrane, whereas short PSCA is localized to the cytoplasm. Thus, the PSCA SNP alters the subcellular localization of the PSCA protein and subsequently changes its function. In addition, the *PSCA* SNP is associated with PSCA expression [49]. SNP rs2294008 was also reported to be associated with bladder cancer [210] and duodenal ulcer risk [181]. Interestingly, the T allele of rs2294008 increases gastric and bladder cancer risk, but reduces duodenal ulcer risk. Thus, the growth-promoting effects of PSCA are responsible for the increased risk of gastric cancer and reduced risk of duodenal ulcers among T allele carriers.

The *MUC1* gene on chromosome 1q22 was identified as a GC susceptibility locus by GWAS of Japanese and Chinese populations (rs2070803: $P = 4.33 \times 10^{-13}$; OR = 1.71) [159]. MUC1 exerts an anti-apoptotic function and is considered an oncogene; however, the mucin 1 protein protects gastric epithelial cells from a variety of external insults that cause inflammation and carcinogenesis, such as an *H. pylori* infection. Two major *MUC1* transcripts are expressed in the gastric epithelium: variants 2 and 3. SNP rs4072037 influences the splicing of the primary transcripts. SNP rs4072037 is located in the splicing acceptor site of exon 2 and determines the type of variants; the G and A alleles result in the expression of variants 2 and 3, respectively [135, 159]. Thus, these two functional variations are associated with gastric cancer risk.

PLCE1 is associated with cardia and noncardia gastric cancer in a Chinese population [197]. *PLCE1* SNPs are also associated with the prognosis of Chinese patients, but not Caucasian patients [121, 139]. PLCE1 is a member of the phospholipase family that is involved in cell growth, differentiation and gene expression. PLCE1 is a novel Ras-related protein effector that regulates the actin cytoskeleton and membrane protrusion [137]. PLCE1-knockout mice revealed a crucial role of PLCE1 in Ras oncogene-induced de novo carcinogenesis. Knockout mice showed a delayed onset and markedly reduced incidence of carcinogen-induced squamous skin tumors, and the papillomas that formed in the mice did not undergo malignant progression into carcinomas [13]. $Apc^{Min/+}$ mice, which carry an inactivated allele of the adenomatous polyposis coli gene, exhibited a higher resistance to spontaneous intestinal tumorigenesis on the $PLCE1^{-/-}$ genetic background compared with mice with intact PLCE1 [112]. Low-grade adenomas in the $PLCE1^{-/-}$ $Apc^{Min/+}$ mice exhibited accelerated apoptosis, reduced cellular proliferation, marked attenuation of tumor angiogenesis and a reduction in vascular endothelial growth factor expression. In contrast, high-grade adenomas in these mice exhibited marked attenuation of tumor-associated inflammation without significant differences in apoptosis and proliferation. Therefore, PLCE1 seems to plays crucial roles in intestinal tumorigenesis through two distinct mechanisms, augmentation of angiogenesis and inflammation, depending on the tumor stage.

A locus on chromosome 5p13.1 [167] was identified by a GWAS of the Chinese population and was validated by other studies. The recombination rate analyses and LD analyses of this locus identified a critical region for the association that harbors five genes: PRKAA1, PTGER4, RPL37, SNORD72, and TTC33. Three SNPs with the lowest P-value of less than 1×10^{-10} are located adjacent to PRKAA1, TTC33 and PTGER4; however, the SNP located on the 5′ side of RPL37 also showed a moderate association, indicating that further studies are required to identify a causative gene in this region. Compared with other cancers, only seven loci have been identified to date. Therefore, further analyses are required to elucidate the molecular pathogenesis of gastric cancer and predict risks for *H. pylori* carriers. List of eight gastric cancer GWAS was shown in Table 5.6.

Esophageal Cancer

Esophageal cancer is the sixth most common cause of cancer-related death in the world [55]. There were 483,000 cases and 439,000 deaths worldwide in 2015. Most patients are at advanced stages at the time of diagnosis, and the overall 5-year survival rate is approximately 10–20%, despite the availability of modern surgical techniques combined with various treatment modalities [180]. Because detection of esophageal cancer at earlier stages can improve clinical outcomes, the identification of epidemiologic factors that influence the development of esophageal cancer would facilitate the prevention or early detection of the disease.

Esophageal cancer is prevalent among Asian populations, with marked regional variations in incidence and mortality; for example, a 20-fold difference in incidence is observed between high-risk China and low-risk western Africa [212]. Although the pathogenesis of esophageal cancer has not been completely elucidated, accumulating epidemiological evidence has identified several disease-promoting factors, such as tobacco smoking, heavy alcohol drinking, nutritional deficiencies, and dietary carcinogen exposure [47]. In addition, familial aggregation of esophageal cancer has also been reported, suggesting that some genetic factors might be involved in the pathogenesis of ESCC [78]. In a Japanese population, a positive family history was associated with a 3.8-fold increased risk of esophageal cancer [74]. Thus, genetic and environmental factors play crucial roles in the etiology of ESCC.

Recent advances in genomic research identified many genes associated with disease risk. To date, 7 GWAS studies have reported 19 genetic variations associated with esophageal cancer susceptibility (Table 5.1). In 2009, a 2-step genome-wide association study of 1070 Japanese ESCC cases and 2836 controls identified significant associations of ESCC with *ADH1B* (rs1229984, $P = 6.76 \times 10^{-35}$) and *ALDH2* (rs671, $P = 3.68 \times 10^{-68}$). Individuals who had two genetic factors (*ADH1B* and *ALDH2*) and two lifestyle-related risk factors (smoking and drinking) had a nearly 190-fold higher risk of ESCC than individuals without these factors. Thus, lifestyle

Table 5.6 List of gastric cancer GWAS

Screening (case/control)	Replication (case/control)	Screening_ethnicity	Journal	Year	PUBMEDID	First author	Identified regions
(1) 188/752	1206/1140	JP	Nat Genet	2008	18488030	Sakamoto H	8q24.3
(2) 1625/2100		CN	Nat Genet	2010	20729852	Abnet CC	10q23.33
(1) 606/1264	756/1837	JP	Gastroenterology	2010	21070779	Saeki N	1q22
(3) 1006/2273	3288/3609	CN	Nat Genet	2011	22037551	Shi Y	3q13.31, 5p13.1
(4) 1006/4006	3330/11436	CN	Am J Hum Genet	2012	23103227	Jin G	6p21.1
2043/202533		EU	Nat Genet	2015	26098866	Helgason H	1q22, 11q22.2, 1q22
(5) 2350/2708	7408/7548	EA	Gut	2015	26129866	Hu N	5p13.1, 6p21.1
(3) 2031/4970	3564/4637	CN	Gut	2015	26701879	Wang Z	1q22, 5p13.1, 5q14.3, 8q24.3

EU European, *AA* African american, *AS* Asian, *EA* East Asian, *CN* Chinese, *JP* Japanese, *KR* Korean, *LA* Larino/Hispanic, *AF* African. (1) diffuse, (2) gastric /esophageal, (3) non cardia, (4) gastric/lung/esophageal, (5) cardia/non-cardia

intervention based on genetic risk factors would be effective for the prevention and early detection of ESCC.

Many GWAS of esophageal cancer were conducted in Chinese populations. In 2010, 551,152 SNPs were analyzed in 2240 Chinese gastric cancer cases, 2115 ESCC cases and 3302 controls and multiple variants at chromosomes 10q23 (rs2274223, a nonsynonymous SNP located in *PLCE1*, $P = 3.85 \times 10^{-9}$; OR = 1.34) and 22q12 (rs738722 in CHEK2, $P = 3.85 \times 10^{-9}$; OR = 1.34) were identified [4]. In addition, loci at chromosomes 5q11.2 (*PDE4D*), 6p21 (HLA region), 21q22 (*RUNX1*) [206] and 2q33.1 (*CASP8*), 2q35 (*IGFB2*), 3q27.3 (*ST6GAL1*), 13q33.2 (*SLC10A2*), 16q12.1 (*HEATR3*) [5, 207] are associated with ESCC. Moreover, a GWAS of more than 10,000 samples identified loci at chromosomes 5q31.2 (*TMEM173*) and 17p13.1 (*TP53*) [209].

Esophageal adenocarcinoma is more common in European populations, whereas squamous cell carcinoma is common among Asian populations. Esophageal adenocarcinoma frequently occurs in an intestinal metaplastic epithelium, which is a diagnostic of Barrett's esophagus. A GWAS of esophageal adenocarcinoma cases ($n = 2390$) and Barrett's esophagus cases ($n = 3175$) with 10,120 controls identified three novel associations with loci at chromosome 19p13 (rs10419226: $P = 3.6 \times 10^{-10}$) in the *CRTC1* gene (encoding CREB-regulated transcription coactivator), at chromosome 9q22 (rs11789015: $P = 1.0 \times 10^{-9}$) in the *BARX1* gene, and at chromosome 3p14 (rs2687201: $P = 5.5 \times 10^{-9}$) near the transcription factor *FOXP*. *CRTC1* encodes a CREB-regulated transcription coactivator, and its aberrant activation is associated with oncogenic activity. *BARX1* encodes a transcription factor important for esophageal specification. *FOXP1* regulates esophageal development [111]. However, these loci are associated with both esophageal cancer and Barrett's esophagus, but none of them cleared the genome-wide significant threshold, with the exception of esophageal adenocarcinoma cases, indicating that these variations play roles in the development of Barrett's esophagus, but not adenocarcinoma. List of nine esophageal cancer GWAS was shown in Table 5.7.

Bladder Cancer

Bladder cancer is one of the most frequent cancers (541,000 cases) and causes 188,000 deaths per year worldwide in 2015 [55]. Both environmental and genetic factors are involved in the development of bladder cancer. Tobacco smoking is known to be the most important factor that increases the risk of bladder cancer; and current or former smokers have a two to sixfold higher risk than never-smokers [23, 48]. In addition, occupational exposures to industrial chemicals [29, 58, 101], arsenic contamination in drinking water [25], and infectious diseases [12] also increase the bladder cancer risk. The bladder cancer incidence in males is nearly threefold higher than the incidence in females [23], probably due to the higher prevalence of tobacco smoking and occupational exposure in males. However, familial

Table 5.7 List of esophageal cancer GWAS

Screening (case/control)	Replication (case/control)	Screening_ethnicity	Journal	Year	PUBMEDID	First author	Identified regions
(1) 182/927	782/1898	JP	Gastroenterology	2009	19698717	Cui R	4q23, 12q24.12
(2) 1898/2100		CN	Nat Genet	2010	20729852	Abnet CC	10q23.33, 22q12.1
(1) 1077/1733	7976/11550	CN	Nat Genet	2010	20729853	Wang LD	10q23, 20p13
(1) 2031/2044	3986/4157	CN	Nat Genet	2011	21642993	Wu C	(5)
(1) 2961/3400		CN	Hum Mol Genet.	2012	22323360	Abnet CC	2q33.1
(1) 2031/2044	8092/8620	CN	Nat Genet	2012	22960999	Wu C	(6)
(3) 2031/4006	3006/11436	CN	Am J Hum Genet	2012	23103227	Jin G	
(4) 1516/3209	874/6911	EU	Nat Genet	2013	24121790	Levine DM	3p14, 9q22, 19p13
(4) 5337/5787	9654/10058	CN	Nat Genet	2014	25129146	Wu C	(7)

EU European, *AA* African american, *AS* Asian, *EA* East Asian, *CN* Chinese, *JP* Japanese, *KR* Korean, *LA* Larino/Hispanic, *AF* African. (1) SCC, (2) gastric/esophageal, (3) gastric/lung/esophageal, (4) Barrett's esophagus/ADC, (5) 5q11.2, 5q32, 6p21.1, 10q23.33, 12q24.12, 12q24.13, 15q23, 21q22.12, (6) 3q27.3, 16q12.1, 17p13.3, 17q21.2, 18p11.21, 22q12.1, (7) 2q33.1, 5q31.2, 10q23.33, 17p13.1, 21q22.12

aggregation of bladder cancer has also been reported [3, 134], suggesting the importance of genetic factors in bladder cancer development.

NAT2 and GSTM1 are involved in the detoxification of carcinogens [69], and a *NAT2* slow-metabolizer genotype and a *GSTM1* null genotype are associated with an increased risk of bladder cancer [52, 153]. In addition, recent genome-wide association studies (GWAS) of European populations have identified multiple genetic factors associated with bladder cancer [45, 53, 151, 156, 210].

In 2008, the first GWAS of 1803 bladder cancer cases and 34,336 controls from Iceland and The Netherlands and follow-up studies in seven additional case control groups (2165 cases and 3800 controls) identified rs9642880 on chromosome 8q24, which is located 30 kb upstream of *MYC* (OR = 1.22; $P = 9.34 \times 10^{-12}$), as bladder cancer susceptibility locus [99]. A further analysis of 4739 cases and 45,549 controls revealed an association of rs798766 on chromosome 4p16.3 with bladder cancer (OR = 1.24, $P = 9.9 \times 10^{-12}$) [98]. rs798766 is located in an intron of the TACC3 gene and is 70 kb away from the *FGFR3* gene. The *FGFR3* gene often harbors activating somatic mutations in patients with low-grade, noninvasive bladder cancer. The frequency of the T allele of rs798766 is higher in Ta tumors (non-invasive papillary bladder tumor) that carry an activating mutation in the FGFR3 gene than in Ta tumors with wild-type *FGFR3*, indicating an association between germline variants, somatic mutations of FGFR3 and the risk of bladder cancer.

A GWAS of 969 bladder cancer cases and 957 controls and further replication analysis of three additional US populations identified a missense variant (rs2294008) in the *PSCA* gene that showed a consistent association with bladder cancer in European populations (6667 cases, 39,590 controls with overall P-value of 2.14×10^{-10} and OR of 1.15) [210]. As described for gastric cancer, rs2294008 alters the start codon and is predicted to truncate nine amino acids from the N-terminal signal sequence of the primary PSCA translation product. Another GWAS with a primary scan of 591,637 SNPs in 3532 bladder cancer cases and 5120 controls of European descent followed by a replication strategy that included 8382 cases and 48,275 controls identified three new regions associated with bladder cancer on chromosomes 22q13.1, 19q12 and 2q37.1 [156]. rs1014971, ($P = 8 \times 10^{-12}$) maps to a non-genic region of chromosome 22q13.1, rs8102137 ($P = 2 \times 10^{-11}$) on chromosome 19q12 maps to the *CCNE1* gene and rs11892031 ($P = 1 \times 10^{-7}$) maps to the *UGT1A* cluster on chromosome 2q37.1.

Bladder cancer loci were also identified in Asian populations. According to a GWAS and an independent replication study of 1131 bladder cancer cases and 12,558 non-cancer controls in the Japanese populations, 15q24 is bladder cancer locus (OR = 1.41 and *P*-value of 4.03×10^{-9}). SNP rs8041357, which is in complete linkage disequilibrium ($r^2 = 1$) with rs11543198, is also associated with bladder cancer risk in Europeans ($P = 0.045$ for an additive and $P = 0.025$ for a recessive model), despite the much lower minor allele frequency in Europeans (3.7%) compared with the Japanese individuals (22.2%). rs8041357 is located in the *CYP1A1-CYP1A2* locus, indicating the role of generic variations in a tobacco metabolizing enzyme in the development of bladder cancer. List of ten bladder cancer GWAS was shown in Table 5.8.

Table 5.8 List of bladder cancer GWAS

Screening (case/control)	Replication (case/control)	Screening_ethnicity	Journal	Year	PUBMEDID	First author	Identified regions
1803/34336	2165/3800	EU	Nat Genet	2008	18794855	Kiemeney LA	8q24.21
969/957	5698/38633	EU	Nat Genet	2009	19648920	Wu X	8q24.3
1889/39310	2850/6239	EU	Nat Genet	2010	20348956	Kiemeney LA	4p16.3, 8q24.21
3532/5120	8381/48275	EU	Nat Genet	2010	20972438	Rothman N	(1)
2234/41603	3640/10507	EU	Hum Mol Genet	2011	21750109	Rafnar T	18q12.3
4501/6076	1382/2201	EU	Hum Mol Genet	2011	21824976	Garcia-Closas M	18q12.3
6911/11818	801/1307	EU	Hum Mol Genet	2013	24163127	Figueroa JD	(2)
1670/90180	5241/10456	EU	Hum Mol Genet	2014	24861552	Rafnar T	20p12.2
531/5581	4100/12065	JP	Hum Mol Genet	2014	25281661	Matsuda K	15q24.1
578/1006	2828/3639	CN	Cancer Res	2016	27206850	Wang M	5q12.3

EU European, *AA* African american, *AS* Asian, *EA* East Asian, *CN* Chinese, *JP* Japanese, *KR* Korean, *LA* Larino/Hispanic, *AF* African., (1) 1p13.3, 3q28, 4p16.3, 8p22, 8q24.21, 8q24.3, 19q12, 22q13.1, (2) 3q26.2, 3q28, 4p16.3, 5p15.33, 8p22, 8q24.21, 11p15.5, 12q12, 19q12, 22q13.1

Pancreatic Cancer

Pancreatic cancer is the seventh-leading cause of cancer-related death [55]. There were 426,000 cases and 412,000 deaths worldwide in 2015 [55, 170, 189]. Its 5-year survival rate is less than 10% and no specific symptoms are observed in patients with early stage pancreatic cancer. Therefore, most of the patients were diagnosed at advanced stages, which have a very low possibility of cure for the disease [96]. Epidemiological studies have identified a number of possible risk factors, such as smoking, diabetes, and chronic pancreatitis, which are likely to predispose individuals to the disease. In addition, familial aggregation of the disease has implied a possible involvement of genetic factors in pancreatic cancer [35]; approximately 10% of the patients were reported to have a family history and individuals having first-degree relatives with pancreatic cancer display a two- to fourfold higher risk of the disease [128]. In Japanese populations, a positive family history was associated with a 3.2-fold increased risk of pancreatic cancer [74]. Several hereditary cancer syndromes caused by mutation of *STK11*, *CDKN2A*, and DNA mismatch repair genes [60, 122] and an inherited form of pancreatitis [119, 204] are associated with a high risk of pancreatic cancer. In addition, several GWAS identified common variations associated with pancreatic cancer risk [9].

In 2009, the first GWAS analyzed 558,542 SNPs in 1896 pancreatic cancer cases and 1939 control and a replication analysis using 2457 cases and 2654 controls identified an association between a locus on 9q34 (rs505922 $P = 5.37 \times 10^{-8}$ and OR = 1.20) [8]. Although this SNP did not clear the genome-wide significance threshold ($P = 5 \times 10^{-8}$), this locus was validated in other studies [205]. The protective allele T for rs505922 is in complete LD ($r^2 = 1.0$) with the O allele of the *ABO* locus, indicating that individuals with blood type O have a low risk of pancreatic cancer.

In addition to 9q34.2 (in the *ABO* blood group gene), Petersen et al. reported a GWAS of 3851 pancreatic cancer cases and 3934 unaffected controls and identified three loci on 1q32.1 (*NR5A2*), 5p15.33 (*CLPTM1L-TERT*) and 13q22.1 (in a large non-genic region flanked by *KLF5* and *KLF12*) in 2010 [143]. Subsequent GWAS in European populations identified additional pancreatic cancer susceptibility loci on 5p15.33 (a second independent risk locus in the *CLPTM1L-TERT* gene region), 7q23.2 (*LINC-PINT*), 16q23.1 (*BCAR1*), 13q12.2 (*PDX1*), 22q12.1 (*ZNRF3*), 8q24.1 (nongenic) [205] and 17q24.3 (*LINC00673*), 2p14 (*ETAA1*), 7p14.1 (*SUGCT*), and 3q28 (*TP63*) [27]. Moreover, a GWAS of a Chinese population including 3584 pancreatic cancer cases and 4868 controls identified five significant risk loci on 5p13.1/*DAB2*, 10q26.11/*PRLHR*, 21q21.3/*BACH1*, 21q22.3/*TFF1*, and 22q13.32/*FAM19A5* [208].

The most significant SNP on chr1q32.1 maps to the first intron of the *NR5A2* gene (rs3790844, OR = 0.77, $P = 2.5 \times 10^{-10}$). This gene encodes nuclear receptor subfamily five group A member 2 (NR5A2). NR5A2 is a transcription factor that plays important roles in multiple aspects of pancreatic development and function, including cholesterol synthesis, bile acid homeostasis, steroidogenesis and in

regulating stemness [44]. Likewise, NR5A2 is an important regulator of exocrine function in the adult pancreas, where it regulates the expression of a number of acinar-specific genes [75]. Heterozygous *Nr5a2* mice are viable and exhibit increased rates of pancreatic acinar to ductal metaplasia and impaired recovery after chemically induced acute pancreatitis [46, 194]. Furthermore, *Nr5a2* haploinsufficiency cooperates with pancreatitis in a mouse model driven by oncogenic *KRAS*, increasing the number of preneoplastic pancreatic intraepithelial neoplasia lesions and driving their progression toward pancreatic ductal adenocarcinoma [46, 194]. Thus, NR5A2 appears to be important for maintaining homeostasis in the exocrine pancreas, promotes the regeneration of functional acinar cells from metaplastic duct-like cells after pancreatitis-induced inflammation, and protects the pancreas from *KRAS*-driven pre-neoplastic changes. Based on the mouse studies, the underlying mechanism may involve negative regulation of *NR5A2* gene expression or function, perhaps in combination with inflammation in the pancreas. List of six pancreatic cancer GWAS was shown in Table 5.9.

Ovarian Cancer

Ovarian cancer is the eighth-leading cause of cancer-related death among women [55]. There were 251,000 cases and 161,000 deaths worldwide in 2015. Evidence from twin and family studies suggests that an inherited genetic component contributes to ovarian cancer risk [114, 177]. The relative risk for first degree relatives is 3.1 (95% CI 2.6–3.7) and 6.54 (95% CI 4.372–9.804) among European and Japanese populations, respectively. Rare, high-penetrance alleles of genes such as *BRCA1* and *BRCA2* account for approximately 25–40% of the increased familial risk [85, 177]. In addition, recent GWAS have identified more than 20 common risk variants.

In 2009, the first GWAS of 1817 cases and 2353 controls and replication analysis identified an association of a locus on 9p22 with disease risk (rs3814113 OR = 0.82, $P = 5.1 \times 10^{-19}$) [175]. The association differs by histological subtype and is strongest for serous ovarian cancers (OR 0.77, $P = 4.1 \times 10^{-21}$). Most of the associated SNPs are located within intron 2 of *BNC2* gene. *BNC2* is highly expressed in reproductive tissues (ovary and testis) and may play a role in the differentiation of spermatozoa and oocytes [154]; however, the role of *BNC2* in cancer development is not well understood. Other studies of European populations identified multiple loci on chromosomes 1p36.12, 1p34.3, 2q31.1, 3q25.31, 4q26, 5p15.33, 6p22.1, 6q25.1, 8q21.13, 8q24.21, 9p22.2, 10p12.31, 16q21, 17q12, 17q21.31, 17q21.32, 17q23.2, and 19p13.11 [26, 61, 103, 145, 150]. A three-stage GWAS of a Chinese population identified loci on chromosomes 9q22.33 (rs1413299 in COL15A1, $P = 1.88 \times 10^{-8}$) and 10p11.21 (rs1192691 near ANKRD30A, $P = 2.62 \times 10^{-8}$) [26].

A multi-stage GWAS of 11,705 BRCA1 carriers (of whom 5920 were diagnosed with breast cancer and 1839 were diagnosed with ovarian cancer) with an additional sample of 2646 BRCA1 carriers was conducted to identify cancer risk-modifying

Table 5.9 List of pancreatic cancer GWAS

Screening (case/control)	Replication (case/control)	Screening_ethnicity	Journal	Year	PUBMEDID	First author	Identified regions
1771/1805	2120/2127	EU	Nat Genet	2009	19648918	Amundadottir L	0
3851/3934		EU, CN, other	Nat Genet	2010	20101243	Petersen GM	1q32.1, 13q22.1
991/5209		JP	PLoS One	2010	20686608	Low SK	
981/1991	2603/2877	CN	Nat Genet	2011	22158540	Wu C	(1)
1582/5203	6101/9194	EU	Nat Genet	2014	25086665	Wolpin BM	(2)
7638/7364	2287/4205	EU	Nat Genet	2015	26098869	Childs EJ	(3)

EU European, *AA* African american, *AS* Asian, *EA* East Asian, *CN* Chinese, *JP* Japanese, *KR* Korean, *LA* Larino/Hispanic, *AF* African., (1) 5p13.1, 10q26.11, 13q22.1, 21q21.3, 21q22.3, 22q13.32, (2) 1q32.1, 5p15.33, 7q32.3, 13q12.2, 13q22.1, 16q23.1, 22q12.1, (3) 2p14, 3q28, 5p15.33, 7p14.1, 13q12.2, 13q22.1, 17q24.3

loci among BRCA1 mutation carriers[20]. As a result, two novel ovarian cancer risk modifier loci were identified: 17q21.31 (rs17631303, $P = 1.4 \times 10^{-8}$, HR = 1.27) and 4q32.3 (rs4691139, $P = 3.4 \times 10^{-8}$, HR = 1.20). The 4q32.3 locus was not associated with ovarian cancer risk in the general population or BRCA2 carriers, suggesting a BRCA1-specific association. The 17q21.31 locus was associated with ovarian cancer risk in 8211 BRCA2 carriers ($P = 2 \times 10^{-4}$). Based on the known ovarian cancer risk-modifying loci, the 5% of BRCA1 carriers with the lowest risk have a 28% or less estimated lifetime risk of developing ovarian cancer, whereas the 5% with the highest risk will have a risk of 63% or higher. List of nine ovarian cancer GWAS was shown in Table 5.10.

Bone Malignancy

Primary bone cancer is rare disease with approximately 2500 new cases diagnosed and 1400 deaths each year in the United States [171]. To date, two studies of malignant bone tumors have been reported. Ewing sarcoma is an aggressive and very rare bone tumor with the most unfavorable prognosis of all primary musculoskeletal tumors. Ewing sarcoma is characterized by a fusion transcript of EWSR1 (22q12)/FLI1 (11q24) in approximately 90% of cases. Despite its very low incidence rates (0.155, 0.082 and 0.017 per 10^5 in Europeans, Africans, and Asians, respectively) [84], familial Ewing sarcoma has been reported [91], suggesting the presence of genetic susceptibility factors for Ewing sarcoma. In 2012, a GWAS of 401 cases and 4352 controls in a European population coupled with two independent replication cohorts identified candidate risk loci on chromosomes 1p36.22 ($P = 1.4 \times 10^{-20}$; OR = 2.2), 10q21 ($P = 4.0 \times 10^{-17}$; OR = 1.7) and 15q15 ($P = 6.6 \times 10^{-9}$; OR = 1.5) [146]. SNP rs9430161 on chromosome 1p36.22 is located 25 kb upstream of the *TARDBP* gene, and rs224278 on chromosome 10q21 is located 5 kb upstream of the EGR2 gene. Variants at these loci are associated with the expression levels of *TARDBP* and *EGR2*. *TARDBP* shares structural similarities with *EWSR1* and *FUS*, which encode RNA binding proteins, whereas *EGR2* is a target gene of EWSR1-ETS, suggesting important roles for these genes in the pathogenesis of Ewing sarcoma.

Osteosarcoma is the most common malignant bone tumor. Osteosarcoma frequently occurs in the long bones of children and young adults and is associated with the pubertal growth spurts, suggesting that proliferating osteoblasts or its precursors are the origin of this malignancy. A tall stature and high birth weight are known risk factors. Osteosarcoma is associated with familial tumor syndromes, such as the Li-Fraumeni syndromes, which are caused by mutations in *TP53* gene. Up to 9.5% of young patients with osteosarcoma were shown to carry pathogenic (3.8%) or rare exonic *TP53* variations (5.7%), indicating the important roles of host genetic factors in disease onset [132]. In 2013, a multistage genome-wide association study consisting of 941 cases and 3291 controls of European ancestry identified two loci in the *GRM4* gene on chromosome 6p21.3 (encoding glutamate receptor metabotropic

Table 5.10 List of ovarian cancer GWAS

Screening (case/control)	Replication (case/control)	Screening_ethnicity	Journal	Year	PUBMEDID	First author	Identified regions
1817/2353	6944/9477	EU	Nat Genet	2009	19648919	Song H	9p22.2
1768/2354	8709/51764	EU	Nat Genet	2010	20852632	Goode EL	2q31.1, 8q24.21
1768/2353	8739/10831	EU	Nat Genet	2010	20852633	Bolton KL	19p13
640/41607		EU	Nat Genet	2011	21964575	Rafnar T	17q23.2
3769/4396	39991/1078	EU	Nat Genet	2013	23535730	Pharoah PD	(3)
(1) 683/2044	1706/10258	EU	PLoS Genet	2013	23544013	Couch FJ	4q32.3, 9p22.2, 17q21.31
(2) 1044/1172	1452/2803	CN	Nat Commun	2014	25134534	Chen K	9q22.33, 10p11.21
(2) 4368/9123	14162/42092	EU	Nat Genet	2015	25581431	Kuchenbaecker KB	(4)
(2) 1644/21693		EU	Nat Genet	2015	26075790	Kelemen LE	2q14.1, 2q31.1, 19q13.2

EU European, *AA* African american, *AS* Asian, *EA* East Asian, *CN* Chinese, *JP* Japanese, *KR* Korean, *LA* Larino/Hispanic, *AF* African. (1) BRCA1 carriers, (2) Epithelial, 3) 2q31.1, 3q25.31, 8q21.13, 8q24.21, 9p22.2, 10p12.31, 17q12, 17q21.32, 19p13.11, p36.12, 1p34.3, 2q31.1, 3q25.31, 4q26, 5p15.33, 6p22.1, 8q21.13, 8q24.21, 9p22.2, 10p12.31, 16q21, 17q12, 17q21.31, 17q21.32, 19p13.11

4; rs1906953; $P = 8.1 \times 10^{-9}$) and a locus in the gene desert at 2p25.2 (rs7591996; $P = 1.0 \times 10^{-8}$) [161]. *GRM4* is implicated in intracellular signaling and inhibition of the cyclic AMP (cAMP) signaling cascade. *GRM4* is expressed in human osteosarcoma cells[92] and is associated with a poor prognosis in patients with various human cancers [19, 24], as well as with cancer cell proliferation *in vitro* [120], suggesting that the cAMP pathway is important in osteosarcoma. List of two malignant bone tumor GWAS was shown in Table 5.11.

Testicular Germ Cell Tumors

Testicular germ cell tumors (TGCT) are one of the common causes of cancer in young men, with a mean age at diagnosis of 36 years [18]. Testicular germ cell tumors account for approximately 1% of all male cancers, with an estimated 72,000 new cases and 9000 deaths worldwide [55]. The incidence of the disease varies considerably between ethnic groups. Its incidence is particularly elevated in European populations [109] and has increased rapidly over recent decades. Well-recognized risk factors for TGCT include a history of undescended testis, microlithiasis, infertility and other testicular abnormalities [190]. Family history is also associated with four to tenfold increased risk of TGCT [71]. Multiple GWAS of TGCT have now been conducted, yielding more than 20 independent loci associated with TGCT risk (Table 5.1) [28, 94, 95, 102, 115, 117, 152, 157, 165, 191].

In 2009, the first GWAS were reported from two groups, in which 730/277 cases and 1435/919 controls were genotyped. A subsequent replication analysis identified three loci on chromosome 12p22 within the *KITLG* gene, ($OR = 2.55$, $P = 10^{-31}$), chromosome 5q31.3 ($OR = 1.37$, $P = 3 \times 10^{-13}$) and chromosome 6p21 ($OR = 1.50$, $P = 10^{-13}$) [94, 152].

The *KITLG* gene encodes the ligand for the receptor tyrosine kinase KIT, and several previous reports support an association between the *KITLG/KIT* system and TGCT formation. Somatic missense mutations or amplifications of *KIT* are observed in approximately one-quarter of human seminomas [97]. Germline homozygous mutations in either *KIT* or *KITLG* cause infertility in mice [155], and the *KITLG/*

Table 5.11 List of malignant bone tumor GWAS

Screening (case/control)	Replication (case/control)	Screening_ethnicity	Journal	Year	PUBMEDID	First author	Identified regions
(1) 427/4352	661/1299	EU	Nat Genet	2012	22327514	Postel-Vinay S	1p36.22, 10q21.3, 15q15.1
(2) 694/2703	247/550	EU	Nat Genet	2013	23727862	Savage SA	2p25.2, 6p21.31

EU European, *AA* African american, *AS* Asian, *EA* East Asian, *CN* Chinese, *JP* Japanese, *KR* Korean, *LA* Larino/Hispanic, *AF* African. (1) Ewing sarcoma, (2) Osteosarcoma

KIT system has been shown to regulate the survival, proliferation and migration of germ cells [17]. In addition, a mouse model with a germline heterozygous deletion of the *KITLG* coding sequence exhibits an increased risk of TGCT [68]. A SNP within the same LD block of rs995030 was shown to encode a functional p53-binding site, and the SNP influences the ability of p53 to bind to and regulate the transcription of the *KITLG* gene [216]. These lines of evidence support an association between the *KITLG/KIT* system and TGCT formation. List of eight testicular tumor GWAS was shown in Table 5.12.

5.5 Conclusions

Recently, more than 2500 GWAS have identified over 25,000 genetic factors associated with various phenotypes, such as disease risk, drug response, and quantitative traits. These studies revealed important biological pathways that will contribute to the implementation of personalized medicine. Regarding cancer risk, more than 250 studies have identified about 700 significant SNPs. The identification of cancer susceptibility genes contributes to our understanding of disease pathogenesis and risk prediction. However, more than 70% of the variations exhibited a relatively weak effect, with ORs of less than 1.3. Collectively, these variants explain only less than 20% of the reported cancer heritability [7, 129]. The reason for the missing heritability remains largely unclear [123]. The inability of GWAS to identify a greater proportion of the genetic risk stems from many factors, including genotyping platform limitations in interrogating rare variations [172]. One explanation for this missing heritability is that GWAS are designed to identify common variants (MAF > 0.01) and only poorly interrogate rare variants (MAF < 0.01) [221]. Consequently, genetic research has shifted to examine the association of rare variations with diseases using next generation sequencing. Cybulski et al. applied whole-exome sequencing in *BRCA1*, *BRCA2*, *CHEK2*, *NBN* or *PALB2* mutation-negative breast cancer patients with strong family histories and/or young ages of onset (n = 195) and identified rare recurrent *RECQL* mutations in populations from both Quebec and Poland (P = 0.00004 and 0.008, OR = 49.6 and 5.4) [33]. Mutations in *RECQL* are very rare in the general population (risk allele frequency = 0.00007–0.00021). *RECQL* is implicated in resolving stalled DNA replication forks to prevent double-stranded DNA (dsDNA) breaks [148], and its function is related to other known breast cancer susceptibility genes. Therefore, RECQL is likely to be a novel breast cancer predisposition gene. In addition, whole-exome sequencing of 51 individuals with multiple colonic adenomas from 48 families identified homozygous germline nonsense mutations in the base-excision repair gene *NTHL1* in three unrelated families [201]. Moreover, all three affected women developed an endometrial malignancy or premalignancy. This mutation was exclusively observed in a heterozygous state in controls (minor allele frequency of 0.0036; $n = 2329$), indicating that a homozygous loss-of-function germline mutation in the *NTHL1* gene predisposes individuals to BER-associated adenomatous polyposis and CRC.

Table 5.12 List of testicular tumor GWAS

Screening (case/control)	Replication (case/control)	Screening_ethnicity	Journal	Year	PUBMEDID	First author	Identified regions
730/1435	571/1806	EU	Nat Genet	2009	19483681	Rapley EA	5q31.3, 6p21.31, 12q21.32
277/919	371/860	EU	Nat Genet	2009	19483682	Kanetsky PA	12q21.32
979/4947	664/3456	EU	Nat Genet	2010	20543847	Turnbull C	(1)
349/919	439/960	EU	Hum Mol Genet	2011	21551455	Kanetsky PA	9p24.3
582/1056	3560/8510	EU	Nat Genet	2013	23666239	Chung CC	(2)
986/4946	1064/10082	EU	Nat Genet	2013	23666240	Ruark E	(3)
1326/6687	999/290	EU	Hum Mol Genet	2015	25877299	Kristiansen W	3p24.3, 4q24, 17q12, 19p12
986/4946	5073/14148	EU	Nat Commun	2015	26503584	Litchfield K	3q23, 11q14.1, 16p13.13, 16q24.2

EU European, *AA* African american, *AS* Asian, *EA* East Asian, *CN* Chinese, *JP* Japanese, *KR* Korean, *LA* Larino/Hispanic, *AF* African. (1) 5p15.33, 5q31.3, 9p24.3, 12p13.1, 12q21.32, (2) 4q22.3, 5q31.3, 7p22.3, 12q21.32, 16q23.1, 17q22, (3) 1q22, 1q24.1, 3p24.3, 4q24, 5p15.33, 5q31.1, 5q31.3, 6p21.31, 8q13.3, 9p24.3, 12p13.1, 12q21.32, 16q12.1, 17q22, 21q22.3

Thus, whole exome sequencing is promising strategy to identify rare but highly penetrant cancer predisposition genes. Recently, many investigators have turned to next generation sequencing to study rare variants in complex diseases. However, a larger number of samples is required to validate the results of rare variant association studies compared with common variant association studies [221]. Therefore, the combination of a rare variant association study and a common variant association study would be a useful strategy for investigating the missing heritability of various cancers.

References

1. (2010) Genome-wide meta-analyses identify multiple loci associated with smoking behavior. Nat Genet 42:441–447
2. Aaltonen L, Johns L, Jarvinen H, Mecklin JP, Houlston R (2007) Explaining the familial colorectal cancer risk associated with mismatch repair (MMR)-deficient and MMR-stable tumors. Clin Cancer Res 13:356–361
3. Aben KK, Witjes JA, Schoenberg MP, Hulsbergen-van de Kaa C, Verbeek AL, Kiemeney LA (2002) Familial aggregation of urothelial cell carcinoma. Int J Cancer 98:274–278
4. Abnet CC, Freedman ND, Hu N, Wang Z, Yu K, Shu XO, Yuan JM, Zheng W, Dawsey SM, Dong LM, Lee MP, Ding T, Qiao YL, Gao YT, Koh WP, Xiang YB, Tang ZZ, Fan JH, Wang C, Wheeler W, Gail MH, Yeager M, Yuenger J, Hutchinson A, Jacobs KB, Giffen CA, Burdett L, Fraumeni JF Jr, Tucker MA, Chow WH, Goldstein AM, Chanock SJ, Taylor PR (2010) A shared susceptibility locus in PLCE1 at 10q23 for gastric adenocarcinoma and esophageal squamous cell carcinoma. Nat Genet 42:764–767
5. Abnet CC, Wang Z, Song X, Hu N, Zhou FY, Freedman ND, Li XM, Yu K, Shu XO, Yuan JM, Zheng W, Dawsey SM, Liao LM, Lee MP, Ding T, Qiao YL, Gao YT, Koh WP, Xiang YB, Tang ZZ, Fan JH, Chung CC, Wang C, Wheeler W, Yeager M, Yuenger J, Hutchinson A, Jacobs KB, Giffen CA, Burdett L, Fraumeni JF Jr, Tucker MA, Chow WH, Zhao XK, Li JM, Li AL, Sun LD, Wei W, Li JL, Zhang P, Li HL, Cui WY, Wang WP, Liu ZC, Yang X, Fu WJ, Cui JL, Lin HL, Zhu WL, Liu M, Chen X, Chen J, Guo L, Han JJ, Zhou SL, Huang J, Wu Y, Yuan C, Huang J, Ji AF, Kul JW, Fan ZM, Wang JP, Zhang DY, Zhang LQ, Zhang W, Chen YF, Ren JL, Li XM, Dong JC, Xing GL, Guo ZG, Yang JX, Mao YM, Yuan Y, Guo ET, Zhang W, Hou ZC, Liu J, Li Y, Tang S, Chang J, Peng XQ, Han M, Yin WL, Liu YL, Hu YL, Liu Y, Yang LQ, Zhu FG, Yang XF, Feng XS, Wang Z, Li Y, Gao SG, Liu HL, Yuan L, Jin Y, Zhang YR, Sheyhidin I et al (2012) Genotypic variants at 2q33 and risk of esophageal squamous cell carcinoma in China: a meta-analysis of genome-wide association studies. Hum Mol Genet 21:2132–2141
6. Akamatsu S, Takahashi A, Takata R, Kubo M, Inoue T, Morizono T, Tsunoda T, Kamatani N, Haiman CA, Wan P, Chen GK, Le Marchand L, Kolonel LN, Henderson BE, Fujioka T, Habuchi T, Nakamura Y, Ogawa O, Nakagawa H (2012) Reproducibility, performance, and clinical utility of a genetic risk prediction model for prostate cancer in Japanese. PLoS One 7:e46454
7. Al Olama AA, Kote-Jarai Z, Berndt SI, Conti DV, Schumacher F, Han Y, Benlloch S, Hazelett DJ, Wang Z, Saunders E, Leongamornlert D, Lindstrom S, Jugurnauth-Little S, Dadaev T, Tymrakiewicz M, Stram DO, Rand K, Wan P, Stram A, Sheng X, Pooler LC, Park K, Xia L, Tyrer J, Kolonel LN, Le Marchand L, Hoover RN, Machiela MJ, Yeager M, Burdette L, Chung CC, Hutchinson A, Yu K, Goh C, Ahmed M, Govindasami K, Guy M, Tammela TL, Auvinen A, Wahlfors T, Schleutker J, Visakorpi T, Leinonen KA, Xu J, Aly M, Donovan J, Travis RC, Key TJ, Siddiq A, Canzian F, Khaw KT, Takahashi A, Kubo M, Pharoah P,

Pashayan N, Weischer M, Nordestgaard BG, Nielsen SF, Klarskov P, Roder MA, Iversen P, Thibodeau SN, Mcdonnell SK, Schaid DJ, Stanford JL, Kolb S, Holt S, Knudsen B, Coll AH, Gapstur SM, Diver WR, Stevens VL, Maier C, Luedeke M, Herkommer K, Rinckleb AE, Strom SS, Pettaway C, Yeboah ED, Tettey Y, Biritwum RB, Adjei AA, Tay E, Truelove A, Niwa S, Chokkalingam AP, Cannon-Albright L, Cybulski C, Wokolorczyk D, Kluzniak W, Park J, Sellers T, Lin HY, Isaacs WB, Partin AW, Brenner H, Dieffenbach AK, Stegmaier C, Chen C, Giovannucci EL et al (2014) A meta-analysis of 87,040 individuals identifies 23 new susceptibility loci for prostate cancer. Nat Genet 46:1103–1109

8. Amundadottir L, Kraft P, Stolzenberg-Solomon RZ, Fuchs CS, Petersen GM, Arslan AA, Bueno-De-Mesquita HB, Gross M, Helzlsouer K, Jacobs EJ, Lacroix A, Zheng W, Albanes D, Bamlet W, Berg CD, Berrino F, Bingham S, Buring JE, Bracci PM, Canzian F, Clavel-Chapelon F, Clipp S, Cotterchio M, De Andrade M, Duell EJ, Fox JW Jr, Gallinger S, Gaziano JM, Giovannucci EL, Goggins M, Gonzalez CA, Hallmans G, Hankinson SE, Hassan M, Holly EA, Hunter DJ, Hutchinson A, Jackson R, Jacobs KB, Jenab M, Kaaks R, Klein AP, Kooperberg C, Kurtz RC, Li D, Lynch SM, Mandelson M, Mcwilliams RR, Mendelsohn JB, Michaud DS, Olson SH, Overvad K, Patel AV, Peeters PH, Rajkovic A, Riboli E, Risch HA, Shu XO, Thomas G, Tobias GS, Trichopoulos D, van den Eeden SK, Virtamo J, Wactawski-Wende J, Wolpin BM, Yu H, Yu K, Zeleniuch-Jacquotte A, Chanock SJ, Hartge P, Hoover RN (2009) Genome-wide association study identifies variants in the ABO locus associated with susceptibility to pancreatic cancer. Nat Genet 41:986–990
9. Amundadottir LT (2016) Pancreatic cancer genetics. Int J Biol Sci 12:314–325
10. Antoniou AC, Wang X, Fredericksen ZS, Mcguffog L, Tarrell R, Sinilnikova OM, Healey S, Morrison J, Kartsonaki C, Lesnick T, Ghoussaini M, Barrowdale D, EMBRACE, Peock S, Cook M, Oliver C, Frost D, Eccles D, Evans DG, Eeles R, Izatt L, Chu C, Douglas F, Paterson J, Stoppa-Lyonnet D, Houdayer C, Mazoyer S, Giraud S, Lasset C, Remenieras A, Caron O, Hardouin A, Berthet P, Collaborators GS, Hogervorst FB, Rookus MA, Jager A, Van den Ouweland A, Hoogerbrugge N, van der Luijt RB, Meijers-Heijboer H, Gomez Garcia EB, HEBON, Devilee P, Vreeswijk MP, Lubinski J, Jakubowska A, Gronwald J, Huzarski T, Byrski T, Gorski B, Cybulski C, Spurdle AB, Holland H, Kconfab, Goldgar DE, John EM, Hopper JL, Southey M, Buys SS, Daly MB, Terry MB, Schmutzler RK, Wappenschmidt B, Engel C, Meindl A, Preisler-Adams S, Arnold N, Niederacher D, Sutter C, Domchek SM, Nathanson KL, Rebbeck T, Blum JL, Piedmonte M, Rodriguez GC, Wakeley K, Boggess JF, Basil J, Blank SV, Friedman E, Kaufman B, Laitman Y, Milgrom R, Andrulis IL, Glendon G, Ozcelik H, Kirchhoff T, Vijai J, Gaudet MM, Altshuler D, Guiducci C, Swe B, Loman N, Harbst K, Rantala J, Ehrencrona H, Gerdes AM, Thomassen M, Sunde L et al (2010) A locus on 19p13 modifies risk of breast cancer in BRCA1 mutation carriers and is associated with hormone receptor-negative breast cancer in the general population. Nat Genet 42:885–892
11. Atherton JC (2006) The pathogenesis of Helicobacter pylori-induced gastro-duodenal diseases. Annu Rev Pathol 1:63–96
12. Badawi AF, Mostafa MH, Probert A, O'Connor PJ (1995) Role of schistosomiasis in human bladder cancer: evidence of association, aetiological factors, and basic mechanisms of carcinogenesis. Eur J Cancer Prev 4:45–59
13. Bai Y, Edamatsu H, Maeda S, Saito H, Suzuki N, Satoh T, Kataoka T (2004) Crucial role of phospholipase Cepsilon in chemical carcinogen-induced skin tumor development. Cancer Res 64:8808–8810
14. Bauer S, Groh V, Wu J, Steinle A, Phillips JH, Lanier LL, Spies T (1999) Activation of NK cells and T cells by NKG2D, a receptor for stress-inducible MICA. Science 285:727–729
15. Benafif S, Eeles R (2016) Genetic predisposition to prostate cancer. Br Med Bull 120:75–89
16. Bishehsari F, Mahdavinia M, Vacca M, Malekzadeh R, Mariani-Costantini R (2014) Epidemiological transition of colorectal cancer in developing countries: environmental factors, molecular pathways, and opportunities for prevention. World J Gastroenterol 20:6055–6072

17. Boldajipour B, Raz E (2007) What is left behind – quality control in germ cell migration. Sci STKE 2007:pe16
18. Bray F, Ferlay J, Devesa SS, McGlynn KA, Moller H (2006) Interpreting the international trends in testicular seminoma and nonseminoma incidence. Nat Clin Pract Urol 3:532–543
19. Brocke KS, Staufner C, Luksch H, Geiger KD, Stepulak A, Marzahn J, Schackert G, Temme A, Ikonomidou C (2010) Glutamate receptors in pediatric tumors of the central nervous system. Cancer Biol Ther 9:455–468
20. Broderick P, Carvajal-Carmona L, Pittman AM, Webb E, Howarth K, Rowan A, Lubbe S, Spain S, Sullivan K, Fielding S, Jaeger E, Vijayakrishnan J, Kemp Z, Gorman M, Chandler I, Papaemmanuil E, Penegar S, Wood W, Sellick G, Qureshi M, Teixeira A, Domingo E, Barclay E, Martin L, Sieber O, Kerr D, Gray R, Peto J, Cazier JB, Tomlinson I, Houlston RS (2007) A genome-wide association study shows that common alleles of SMAD7 influence colorectal cancer risk. Nat Genet 39:1315–1317
21. Caldas C, Carneiro F, Lynch HT, Yokota J, Wiesner GL, Powell SM, Lewis FR, Huntsman DG, Pharoah PD, Jankowski JA, Macleod P, Vogelsang H, Keller G, Park KG, Richards FM, Maher ER, Gayther SA, Oliveira C, Grehan N, Wight D, Seruca R, Roviello F, Ponder BA, Jackson CE (1999) Familial gastric cancer: overview and guidelines for management. J Med Genet 36:873–880
22. Carstensen B, Soll-Johanning H, Villadsen E, Sondergaard JO, Lynge E (1996) Familial aggregation of colorectal cancer in the general population. Int J Cancer 68:428–435
23. Castelao JE, Yuan JM, Skipper PL, Tannenbaum SR, Gago-Dominguez M, Crowder JS, Ross RK, Yu MC (2001) Gender- and smoking-related bladder cancer risk. J Natl Cancer Inst 93:538–545
24. Chang HJ, Yoo BC, Lim SB, Jeong SY, Kim WH, Park JG (2005) Metabotropic glutamate receptor 4 expression in colorectal carcinoma and its prognostic significance. Clin Cancer Res 11:3288–3295
25. Chen CJ, Kuo TL, Wu MM (1988) Arsenic and cancers. Lancet 1:414–415
26. Chen K, Ma H, Li L, Zang R, Wang C, Song F, Shi T, Yu D, Yang M, Xue W, Dai J, Li S, Zheng H, Wu C, Zhang Y, Wu X, Li D, Xue F, Li H, Jiang Z, Liu J, Liu Y, Li P, Tan W, Han J, Jie J, Hao Q, Hu Z, Lin D, Ma D, Jia W, Shen H, Wei Q (2014) Genome-wide association study identifies new susceptibility loci for epithelial ovarian cancer in Han Chinese women. Nat Commun 5:4682
27. Childs EJ, Mocci E, Campa D, Bracci PM, Gallinger S, Goggins M, Li D, Neale RE, Olson SH, Scelo G, Amundadottir LT, Bamlet WR, Bijlsma MF, Blackford A, Borges M, Brennan P, Brenner H, Bueno-de-Mesquita HB, Canzian F, Capurso G, Cavestro GM, Chaffee KG, Chanock SJ, Cleary SP, Cotterchio M, Foretova L, Fuchs C, Funel N, Gazouli M, Hassan M, Herman JM, Holcatova I, Holly EA, Hoover RN, Hung RJ, Janout V, Key TJ, Kupcinskas J, Kurtz RC, Landi S, Lu L, Malecka-Panas E, Mambrini A, Mohelnikova-Duchonova B, Neoptolemos JP, Oberg AL, Orlow I, Pasquali C, Pezzilli R, Rizzato C, Saldia A, Scarpa A, Stolzenberg-Solomon RZ, Strobel O, Tavano F, Vashist YK, Vodicka P, Wolpin BM, Yu H, Petersen GM, Risch HA, Klein AP (2015) Common variation at 2p13.3, 3q29, 7p13 and 17q25.1 associated with susceptibility to pancreatic cancer. Nat Genet 47:911–916
28. Chung CC, Kanetsky PA, Wang Z, Hildebrandt MA, Koster R, Skotheim RI, Kratz CP, Turnbull C, Cortessis VK, Bakken AC, Bishop DT, Cook MB, Erickson RL, Fossa SD, Jacobs KB, Korde LA, Kraggerud SM, Lothe RA, Loud JT, Rahman N, Skinner EC, Thomas DC, Wu X, Yeager M, Schumacher FR, Greene MH, Schwartz SM, Mcglynn KA, Chanock SJ, Nathanson KL (2013) Meta-analysis identifies four new loci associated with testicular germ cell tumor. Nat Genet 45:680–685
29. Colt JS, Baris D, Stewart P, Schned AR, Heaney JA, Mott LA, Silverman D, Karagas M (2004) Occupation and bladder cancer risk in a population-based case-control study in New Hampshire. Cancer Causes Control: CCC 15:759–769

30. Couch FJ, Wang X, Mcguffog L, Lee A, Olswold C, Kuchenbaecker KB, Soucy P, Fredericksen Z, Barrowdale D, Dennis J, Gaudet MM, Dicks E, Kosel M, Healey S, Sinilnikova OM, Lee A, Bacot F, Vincent D, Hogervorst FB, Peock S, Stoppa-Lyonnet D, Jakubowska A, KCONFAB, I, Radice P, Schmutzler RK, Swe B, Domchek SM, Piedmonte M, Singer CF, Friedman E, Thomassen M, Ontario Cancer Genetics, N, Hansen TV, Neuhausen SL, Szabo CI, Blanco I, Greene MH, Karlan BY, Garber J, Phelan CM, Weitzel JN, Montagna M, Olah E, Andrulis IL, Godwin AK, Yannoukakos D, Goldgar DE, Caldes T, Nevanlinna H, Osorio A, Terry MB, Daly MB, van Rensburg EJ, Hamann U, Ramus SJ, Toland AE, Caligo MA, Olopade OI, Tung N, Claes K, Beattie MS, Southey MC, Imyanitov EN, Tischkowitz M, Janavicius R, John EM, Kwong A, Diez O, Balmana J, Barkardottir RB, Arun BK, Rennert G, Teo SH, Ganz PA, Campbell I, van der Hout AH, van Deurzen CH, Seynaeve C, Gomez Garcia EB, van Leeuwen FE, Meijers-Heijboer HE, Gille JJ, Ausems MG, Blok MJ, Ligtenberg MJ, Rookus MA, Devilee P, Verhoef S, van Os TA, Wijnen JT, HEBON, EMBRACE, Frost D, Ellis S, Fineberg E, Platte R, Evans DG, Izatt L, Eeles RA, Adlard J et al (2013) Genome-wide association study in BRCA1 mutation carriers identifies novel loci associated with breast and ovarian cancer risk. PLoS Genet 9:e1003212
31. Cui R, Okada Y, Jang SG, Ku JL, Park JG, Kamatani Y, Hosono N, Tsunoda T, Kumar V, Tanikawa C, Kamatani N, Yamada R, Kubo M, Nakamura Y, Matsuda K (2011) Common variant in 6q26-q27 is associated with distal colon cancer in an Asian population. Gut 60:799–805
32. Custer B, Sullivan SD, Hazlet TK, Iloeje U, Veenstra DL, Kowdley KV (2004) Global epidemiology of hepatitis B virus. J Clin Gastroenterol 38:S158–S168
33. Cybulski C, Carrot-Zhang J, Kluzniak W, Rivera B, Kashyap A, Wokolorczyk D, Giroux S, Nadaf J, Hamel N, Zhang S, Huzarski T, Gronwald J, Byrski T, Szwiec M, Jakubowska A, Rudnicka H, Lener M, Masojc B, Tonin PN, Rousseau F, Gorski B, Debniak T, Majewski J, Lubinski J, Foulkes WD, Narod SA, Akbari MR (2015) Germline RECQL mutations are associated with breast cancer susceptibility. Nat Genet 47:643–646
34. Danesh J (1999) Helicobacter pylori infection and gastric cancer: systematic review of the epidemiological studies. Aliment Pharmacol Ther 13:851–856
35. del Chiaro M, Zerbi A, Falconi M, Bertacca L, Polese M, Sartori N, Boggi U, Casari G, Longoni BM, Salvia R (2007) Cancer risk among the relatives of patients with pancreatic ductal adenocarcinoma. Pancreatology 7:459–469
36. Dong J, Hu Z, Wu C, Guo H, Zhou B, Lv J, Lu D, Chen K, Shi Y, Chu M, Wang C, Zhang R, Dai J, Jiang Y, Cao S, Qin Z, Yu D, Ma H, Jin G, Gong J, Sun C, Zhao X, Yin Z, Yang L, Li Z, Deng Q, Wang J, Wu W, Zheng H, Zhou G, Chen H, Guan P, Peng Z, Chen Y, Shu Y, Xu L, Liu X, Liu L, Xu P, Han B, Bai C, Zhao Y, Zhang H, Yan Y, Amos CI, Chen F, Tan W, Jin L, Wu T, Lin D, Shen H (2012) Association analyses identify multiple new lung cancer susceptibility loci and their interactions with smoking in the Chinese population. Nat Genet 44:895–899
37. Dong J, Jin G, Wu C, Guo H, Zhou B, Lv J, Lu D, Shi Y, Shu Y, Xu L, Chu M, Wang C, Zhang R, Dai J, Jiang Y, Yu D, Ma H, Zhao X, Yin Z, Yang L, Li Z, Deng Q, Cao S, Qin Z, Gong J, Sun C, Wang J, Wu W, Zhou G, Chen H, Guan P, Chen Y, Liu X, Liu L, Xu P, Han B, Bai C, Zhao Y, Zhang H, Yan Y, Liu J, Amos CI, Chen F, Tan W, Jin L, Wu T, Hu Z, Lin D, Shen H (2013) Genome-wide association study identifies a novel susceptibility locus at 12q23.1 for lung squamous cell carcinoma in han chinese. PLoS Genet 9:e1003190
38. Dunlop MG, Dobbins SE, Farrington SM, Jones AM, Palles C, Whiffin N, Tenesa A, Spain S, Broderick P, Ooi LY, Domingo E, Smillie C, Henrion M, Frampton M, Martin L, Grimes G, Gorman M, Semple C, Ma YP, Barclay E, Prendergast J, Cazier JB, Olver B, Penegar S, Lubbe S, Chander I, Carvajal-Carmona LG, Ballereau S, Lloyd A, Vijayakrishnan J, Zgaga L, Rudan I, Theodoratou E, Colorectal Tumour Gene Identification, C, Starr JM, Deary I, Kirac I, Kovacevic D, Aaltonen LA, Renkonen-Sinisalo L, Mecklin JP, Matsuda K, Nakamura Y, Okada Y, Gallinger S, Duggan DJ, Conti D, Newcomb P, Hopper J, Jenkins MA, Schumacher F, Casey G, Easton D, Shah M, Pharoah P, Lindblom A, Liu T, Swedish Low-Risk Colorectal

Cancer Study, G, Smith CG, West H, Cheadle JP, Group, C. C, Midgley R, Kerr DJ, Campbell H, Tomlinson IP, Houlston RS (2012a) Common variation near CDKN1A, POLD3 and SHROOM2 influences colorectal cancer risk. Nat Genet 44:770–776

39. Dunlop MG, Dobbins SE, Farrington SM, Jones AM, Palles C, Whiffin N, Tenesa A, Spain S, Broderick P, Ooi LY, Domingo E, Smillie C, Henrion M, Frampton M, Martin L, Grimes G, Gorman M, Semple C, Ma YP, Barclay E, Prendergast J, Cazier JB, Olver B, Penegar S, Lubbe S, Chander I, Carvajal-Carmona LG, Ballereau S, Lloyd A, Vijayakrishnan J, Zgaga L, Rudan I, Theodoratou E, Starr JM, Deary I, Kirac I, Kovacevic D, Aaltonen LA, Renkonen-Sinisalo L, Mecklin JP, Matsuda K, Nakamura Y, Okada Y, Gallinger S, Duggan DJ, Conti D, Newcomb P, Hopper J, Jenkins MA, Schumacher F, Casey G, Easton D, Shah M, Pharoah P, Lindblom A, Liu T, Smith CG, West H, Cheadle JP, Midgley R, Kerr DJ, Campbell H, Tomlinson IP, Houlston RS (2012b) Common variation near CDKN1A, POLD3 and SHROOM2 influences colorectal cancer risk. Nat Genet 44:770–776
40. Easton DF (1999) How many more breast cancer predisposition genes are there? Breast Cancer Res 1:14–17
41. Easton DF, Pooley KA, Dunning AM, Pharoah PD, Thompson D, Ballinger DG, Struewing JP, Morrison J, Field H, Luben R, Wareham N, Ahmed S, Healey CS, Bowman R, Collaborators, S, Meyer KB, Haiman CA, Kolonel LK, Henderson BE, Le Marchand L, Brennan P, Sangrajrang S, Gaborieau V, Odefrey F, Shen CY, Wu PE, Wang HC, Eccles D, Evans DG, Peto J, Fletcher O, Johnson N, Seal S, Stratton MR, Rahman N, Chenevix-Trench G, Bojesen SE, Nordestgaard BG, Axelsson CK, Garcia-Closas M, Brinton L, Chanock S, Lissowska J, Peplonska B, Nevanlinna H, Fagerholm R, Eerola H, Kang D, Yoo KY, Noh DY, Ahn SH, Hunter DJ, Hankinson SE, Cox DG, Hall P, Wedren S, Liu J, Low YL, Bogdanova N, Schurmann P, Dork T, Tollenaar RA, Jacobi CE, Devilee P, Klijn JG, Sigurdson AJ, Doody MM, Alexander BH, Zhang J, Cox A, Brock IW, Macpherson G, Reed MW, Couch FJ, Goode EL, Olson JE, Meijers-Heijboer H, van den Ouweland A, Uitterlinden A, Rivadeneira F, Milne RL, Ribas G, Gonzalez-Neira A, Benitez J, Hopper JL, McCredie M, Southey M, Giles GG, Schroen C, Justenhoven C, Brauch H, Hamann U, Ko YD, Spurdle AB, Beesley J, Chen X, KCONFAB, GROUP, A. M, Mannermaa A, Kosma VM et al (2007) Genome-wide association study identifies novel breast cancer susceptibility loci. Nature 447:1087–1093
42. Eeles R, Goh C, Castro E, Bancroft E, Guy M, Al Olama AA, Easton D, Kote-Jarai Z (2014) The genetic epidemiology of prostate cancer and its clinical implications. Nat Rev Urol 11:18–31
43. Ewing CM, Ray AM, Lange EM, Zuhlke KA, Robbins CM, Tembe WD, Wiley KE, Isaacs SD, Johng D, Wang Y, Bizon C, Yan G, Gielzak M, Partin AW, Shanmugam V, Izatt T, Sinari S, Craig DW, Zheng SL, Walsh PC, Montie JE, Xu J, Carpten JD, Isaacs WB, Cooney KA (2012) Germline mutations in HOXB13 and prostate-cancer risk. N Engl J Med 366:141–149
44. Fayard E, Auwerx J, Schoonjans K (2004) LRH-1: an orphan nuclear receptor involved in development, metabolism and steroidogenesis. Trends Cell Biol 14:250–260
45. Figueroa JD, Ye Y, Siddiq A, Garcia-Closas M, Chatterjee N, Prokunina-Olsson L, Cortessis VK, Kooperberg C, Cussenot O, Benhamou S, Prescott J, Porru S, Dinney CP, Malats N, Baris D, Purdue M, Jacobs EJ, Albanes D, Wang Z, Deng X, Chung CC, Tang W, Bas Bueno-de-Mesquita H, Trichopoulos D, Ljungberg B, Clavel-Chapelon F, Weiderpass E, Krogh V, Dorronsoro M, Travis R, Tjonneland A, Brenan P, Chang-Claude J, Riboli E, Conti D, Gago-Dominguez M, Stern MC, Pike MC, van den Berg D, Yuan JM, Hohensee C, Rodabough R, Cancel-Tassin G, Roupret M, Comperat E, Chen C, De Vivo I, Giovannucci E, Hunter DJ, Kraft P, Lindstrom S, Carta A, Pavanello S, Arici C, Mastrangelo G, Kamat AM, Lerner SP, Barton Grossman H, Lin J, Gu J, Pu X, Hutchinson A, Burdette L, Wheeler W, Kogevinas M, Tardon A, Serra C, Carrato A, Garcia-Closas R, Lloreta J, Schwenn M, Karagas MR, Johnson A, Schned A, Armenti KR, Hosain GM, Andriole G Jr, Grubb R 3rd, Black A, Ryan Diver W, Gapstur SM, Weinstein SJ, Virtamo J, Haiman CA, Landi MT, Caporaso N, Fraumeni JF Jr, Vineis P, Wu X, Silverman DT, Chanock S, Rothman N (2014) Genome-wide association study identifies multiple loci associated with bladder cancer risk. Hum Mol Genet 23:1387–1398

46. Flandez M, Cendrowski J, Canamero M, Salas A, del Pozo N, Schoonjans K, Real FX (2014) Nr5a2 heterozygosity sensitises to, and cooperates with, inflammation in KRas(G12V)-driven pancreatic tumourigenesis. Gut 63:647–655
47. Franceschi S, Bidoli E, Negri E, Zambon P, Talamini R, Ruol A, Parpinel M, Levi F, Simonato L, La Vecchia C (2000) Role of macronutrients, vitamins and minerals in the aetiology of squamous-cell carcinoma of the oesophagus. Int J Cancer 86:626–631
48. Freedman ND, Silverman DT, Hollenbeck AR, Schatzkin A, Abnet CC (2011) Association between smoking and risk of bladder cancer among men and women. JAMA 306:737–745
49. Fu YP, Kohaar I, Rothman N, Earl J, Figueroa JD, Ye Y, Malats N, Tang W, Liu L, Garcia-Closas M, Muchmore B, Chatterjee N, Tarway M, Kogevinas M, Porter-Gill P, Baris D, Mumy A, Albanes D, Purdue MP, Hutchinson A, Carrato A, Tardon A, Serra C, Garcia-Closas R, Lloreta J, Johnson A, Schwenn M, Karagas MR, Schned A, Diver WR, Gapstur SM, Thun MJ, Virtamo J, Chanock SJ, Fraumeni JF Jr, Silverman DT, Wu X, Real FX, Prokunina-Olsson L (2012) Common genetic variants in the PSCA gene influence gene expression and bladder cancer risk. Proc Natl Acad Sci U S A 109:4974–4979
50. Fuchs CS, Giovannucci EL, Colditz GA, Hunter DJ, Speizer FE, Willett WC (1994) A prospective study of family history and the risk of colorectal cancer. N Engl J Med 331:1669–1674
51. Fukase K, Kato M, Kikuchi S, Inoue K, Uemura N, Okamoto S, Terao S, Amagai K, Hayashi S, Asaka M (2008) Effect of eradication of Helicobacter pylori on incidence of metachronous gastric carcinoma after endoscopic resection of early gastric cancer: an open-label, randomised controlled trial. Lancet 372:392–397
52. Garcia A-Closas M, Malats N, Silverman D, Dosemeci M, Kogevinas M, Hein DW, Tard NA, Serra C, Carrato A, Garc A-Closas R (2005) *NAT2* slow acetylation, *GSTM1* null genotype, and risk of bladder cancer: results from the Spanish Bladder Cancer Study and meta-analyses. Lancet 366:649–659
53. Garcia A-Closas M, Ye Y, Rothman N, Figueroa JD, Malats N, Dinney CP, Chatterjee N, Prokunina-Olsson L, Wang Z, Lin J, Real FX, Jacobs KB, Baris D, Thun M, de Vivo I, Albanes D, Purdue MP, Kogevinas M, Kamat AM, Lerner SP, Grossman HB, Gu J, Pu X, Hutchinson A, Fu YP, Burdett L, Yeager M, Tang W, Tardon A, Serra C, Carrato A, Garcia A-Closas R, Lloreta J, Johnson A, Schwenn M, Karagas MR, Schned A, Andriole G Jr, Grubb R 3rd, Black A, Jacobs EJ, Diver WR, Gapstur SM, Weinstein SJ, Virtamo J, Hunter DJ, Caporaso N, Landi MT, Fraumeni JF Jr, Silverman DT, Chanock SJ, Wu X (2011) A genome-wide association study of bladder cancer identifies a new susceptibility locus within SLC14A1, a urea transporter gene on chromosome 18q12.3. Hum Mol Genet 20:4282–4289
54. Ge D, Fellay J, Thompson AJ, Simon JS, Shianna KV, Urban TJ, Heinzen EL, Qiu P, Bertelsen AH, Muir AJ, Sulkowski M, Mchutchison JG, Goldstein DB (2009) Genetic variation in IL28B predicts hepatitis C treatment-induced viral clearance. Nature 461:399–401
55. Global Burden of Disease Cancer C, Fitzmaurice C, Allen C, Barber RM, Barregard L, Bhutta ZA, Brenner H, Dicker DJ, Chimed-Orchir O, Dandona R, Dandona L, Fleming T, Forouzanfar MH, Hancock J, Hay RJ, Hunter-Merrill R, Huynh C, Hosgood HD, Johnson CO, Jonas JB, Khubchandani J, Kumar GA, Kutz M, Lan Q, Larson HJ, Liang X, Lim SS, Lopez AD, Macintyre MF, Marczak L, Marquez N, Mokdad AH, Pinho C, Pourmalek F, Salomon JA, Sanabria JR, Sandar L, Sartorius B, Schwartz SM, Shackelford KA, Shibuya K, Stanaway J, Steiner C, Sun J, Takahashi K, Vollset SE, Vos T, Wagner JA, Wang H, Westerman R, Zeeb H, Zoeckler L, Abd-Allah F, Ahmed MB, Alabed S, Alam NK, Aldhahri SF, Alem G, Alemayohu MA, Ali R, Al-Raddadi R, Amare A, Amoako Y, Artaman A, Asayesh H, Atnafu N, Awasthi A, Saleem HB, Barac A, Bedi N, Bensenor I, Berhane A, Bernabe E, Betsu B, Binagwaho A, Boneya D, Campos-Nonato I, Castaneda-Orjuela C, Catala-Lopez F, Chiang P, Chibueze C, Chitheer A, Choi JY, Cowie B, Damtew S, Das Neves J, Dey S, Dharmaratne S, Dhillon P, Ding E, Driscoll T, Ekwueme D, Endries AY, Farvid M, Farzadfar F, Fernandes J, Fischer F, TT G Hiwot., Gebru A, Gopalani S, et al. (2016) Global, Regional, and National Cancer Incidence, Mortality, years of life lost, years lived with disability, and disability-adjusted life-years for 32 cancer groups, 1990 to 2015: a systematic analysis for the Global Burden of Disease Study. JAMA Oncol

56. Global Burden of Disease Cancer, C, Fitzmaurice C, Dicker D, Pain A, Hamavid H, Moradi-Lakeh M, Macintyre MF, Allen C, Hansen G, Woodbrook R, Wolfe C, Hamadeh RR, Moore A, Werdecker A, Gessner BD, Te Ao B, Mcmahon B, Karimkhani C, Yu C, Cooke GS, Schwebel DC, Carpenter DO, Pereira DM, Nash D, Kazi DS, de Leo D, Plass D, Ukwaja KN, Thurston GD, Yun Jin K, Simard EP, Mills E, Park EK, Catala-Lopez F, Deveber G, Gotay C, Khan G, Hosgood HD 3rd, Santos IS, Leasher JL, Singh J, Leigh J, Jonas JB, Sanabria J, Beardsley J, Jacobsen KH, Takahashi K, Franklin RC, Ronfani L, Montico M, Naldi L, Tonelli M, Geleijnse J, Petzold M, Shrime MG, Younis M, Yonemoto N, Breitborde N, Yip P, Pourmalek F, Lotufo PA, Esteghamati A, Hankey GJ, Ali R, Lunevicius R, Malekzadeh R, Dellavalle R, Weintraub R, Lucas R, Hay R, Rojas-Rueda D, Westerman R, Sepanlou SG, Nolte S, Patten S, Weichenthal S, Abera SF, Fereshtehnejad SM, Shiue I, Driscoll T, Vasankari T, Alsharif U, Rahimi-Movaghar V, Vlassov VV, Marcenes WS, Mekonnen W, Melaku YA, Yano Y, Artaman A, Campos I, Maclachlan J, Mueller U, Kim D, Trillini M, Eshrati B, Williams HC, Shibuya K, Dandona R, Murthy K, Cowie B et al (2015) The Global Burden of Cancer 2013. JAMA Oncol 1:505–527
57. GLOBAL BURDEN OF HEPATITIS, C. W. G (2004) Global burden of disease (GBD) for hepatitis C. J Clin Pharmacol 44:20–29
58. Goebell PJ, Villanueva CM, Rettenmeier AW, Rubben H, Kogevinas M (2004) Environmental exposure, chlorinated drinking water, and bladder cancer. World J Urol 21:424–432
59. Goldgar DE, Easton DF, Cannon-Albright LA, Skolnick MH (1994) Systematic population-based assessment of cancer risk in first-degree relatives of cancer probands. J Natl Cancer Inst 86:1600–1608
60. Goldstein AM, Chan M, Harland M, Gillanders EM, Hayward NK, Avril MF, Azizi E, Bianchi-Scarra G, BISHOP DT, Bressac-de Paillerets B, Bruno W, Calista D, Cannon Albright LA, Demenais F, Elder DE, Ghiorzo P, Gruis NA, Hansson J, Hogg D, Holland EA, Kanetsky PA, Kefford RF, Landi MT, Lang J, Leachman SA, Mackie RM, Magnusson V, Mann GJ, Niendorf K, Newton Bishop J, Palmer JM, Puig S, Puig-Butille JA, De Snoo FA, Stark M, Tsao H, Tucker MA, Whitaker L, Yakobson E, Melanoma Genetics, C (2006) High-risk melanoma susceptibility genes and pancreatic cancer, neural system tumors, and uveal melanoma across GenoMEL. Cancer Res 66:9818–9828
61. Goode EL, Chenevix-Trench G, Song H, Ramus SJ, Notaridou M, Lawrenson K, Widschwendter M, Vierkant RA, Larson MC, Kjaer SK, Birrer MJ, Berchuck A, Schildkraut J, Tomlinson I, Kiemeney LA, Cook LS, Gronwald J, Garcia-Closas M, Gore ME, Campbell I, Whittemore AS, Sutphen R, Phelan C, Anton-Culver H, Pearce CL, Lambrechts D, Rossing MA, Chang-Claude J, Moysich KB, Goodman MT, Dork T, Nevanlinna H, Ness RB, Rafnar T, Hogdall C, Hogdall E, Fridley BL, Cunningham JM, Sieh W, Mcguire V, Godwin AK, Cramer DW, Hernandez D, Levine D, Lu K, Iversen ES, Palmieri RT, Houlston R, Van Altena AM, Aben KK, Massuger LF, Brooks-Wilson A, Kelemen LE, Le ND, Jakubowska A, Lubinski J, Medrek K, Stafford A, Easton DF, Tyrer J, Bolton KL, Harrington P, Eccles D, Chen A, Molina AN, Davila BN, Arango H, Tsai YY, Chen Z, Risch HA, Mclaughlin J, Narod SA, Ziogas A, Brewster W, Gentry-Maharaj A, Menon U, Wu AH, Stram DO, Pike MC, Wellcome Trust Case-Control, C, Beesley J, Webb PM, Australian Cancer, S, Australian Ovarian Cancer Study, G, Ovarian Cancer Association, C, Chen X, Ekici AB, Thiel FC, Beckmann MW, Yang H, Wentzensen N, Lissowska J, Fasching PA, Despierre E, Amant F, Vergote I, Doherty J, Hein R, Wang-Gohrke S, Lurie G et al (2010) A genome-wide association study identifies susceptibility loci for ovarian cancer at 2q31 and 8q24. Nat Genet 42:874–879
62. Groh V, Rhinehart R, Secrist H, Bauer S, Grabstein KH, Spies T (1999) Broad tumor-associated expression and recognition by tumor-derived gamma delta T cells of MICA and MICB. Proc Natl Acad Sci U S A 96:6879–6884
63. Gronberg H (2003) Prostate cancer epidemiology. Lancet 361:859–864
64. Group, H. A. C. C (2001) Gastric cancer and Helicobacter pylori: a combined analysis of 12 case control studies nested within prospective cohorts. Gut 49:347–353

65. Gu Z, Thomas G, Yamashiro J, Shintaku IP, Dorey F, Raitano A, Witte ON, Said JW, Loda M, Reiter RE (2000) Prostate stem cell antigen (PSCA) expression increases with high Gleason score, advanced stage and bone metastasis in prostate cancer. Oncogene 19:1288–1296
66. Gudmundsson J, Sulem P, Manolescu A, Amundadottir LT, Gudbjartsson D, Helgason A, Rafnar T, Bergthorsson JT, Agnarsson BA, Baker A, Sigurdsson A, Benediktsdottir KR, Jakobsdottir M, Xu J, Blondal T, Kostic J, Sun J, Ghosh S, Stacey SN, Mouy M, Saemundsdottir J, Backman VM, Kristjansson K, Tres A, Partin AW, Albers-Akkers MT, Godino-Ivan Marcos J, Walsh PC, Swinkels DW, Navarrete S, Isaacs SD, Aben KK, Graif T, Cashy J, Ruiz-Echarri M, Wiley KE, Suarez BK, Witjes JA, Frigge M, Ober C, Jonsson E, Einarsson GV, Mayordomo JI, Kiemeney LA, Isaacs WB, Catalona WJ, Barkardottir RB, Gulcher JR, Thorsteinsdottir U, Kong A, Stefansson K (2007a) Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24. Nat Genet 39:631–637
67. Gudmundsson J, Sulem P, Steinthorsdottir V, Bergthorsson JT, Thorleifsson G, Manolescu A, Rafnar T, Gudbjartsson D, Agnarsson BA, Baker A, Sigurdsson A, Benediktsdottir KR, Jakobsdottir M, Blondal T, Stacey SN, Helgason A, Gunnarsdottir S, Olafsdottir A, Kristinsson KT, Birgisdottir B, Ghosh S, Thorlacius S, Magnusdottir D, Stefansdottir G, Kristjansson K, Bagger Y, Wilensky RL, Reilly MP, Morris AD, Kimber CH, Adeyemo A, Chen Y, ZHOU J, So WY, Tong PC, Ng MC, Hansen T, Andersen G, Borch-Johnsen K, Jorgensen T, Tres A, Fuertes F, Ruiz-Echarri M, Asin L, Saez B, Van Boven E, Klaver S, Swinkels DW, Aben KK, Graif T, Cashy J, Suarez BK, Van Vierssen Trip O, Frigge ML, Ober C, Hofker MH, Wijmenga C, Christiansen C, Rader DJ, Palmer CN, Rotimi C, Chan JC, Pedersen O, Sigurdsson G, Benediktsson R, Jonsson E, Einarsson GV, Mayordomo JI, Catalona WJ, Kiemeney LA, Barkardottir RB, Gulcher JR, Thorsteinsdottir U, Kong A, Stefansson K (2007b) Two variants on chromosome 17 confer prostate cancer risk, and the one in TCF2 protects against type 2 diabetes. Nat Genet 39:977–983
68. Heaney JD, Lam MY, Michelson MV, Nadeau JH (2008) Loss of the transmembrane but not the soluble kit ligand isoform increases testicular germ cell tumor susceptibility in mice. Cancer Res 68:5193–5197
69. Hein DW (2002) Molecular genetics and function of NAT1 and NAT2: role in aromatic amine metabolism and carcinogenesis. Mutat Res/Fundam Mol Mech Mutagen 506:65–77
70. Helgason H, Rafnar T, Olafsdottir HS, Jonasson JG, Sigurdsson A, Stacey SN, Jonasdottir A, Tryggvadottir L, Alexiusdottir K, Haraldsson A, Le Roux L, Gudmundsson J, Johannsdottir H, Oddsson A, Gylfason A, Magnusson OT, Masson G, Jonsson T, Skuladottir H, Gudbjartsson DF, Thorsteinsdottir U, Sulem P, Stefansson K (2015) Loss-of-function variants in ATM confer risk of gastric cancer. Nat Genet 47:906–910
71. Hemminki K, Li X (2004) Familial risk in testicular cancer as a clue to a heritable and environmental aetiology. Br J Cancer 90:1765–1770
72. Herbst RS, Heymach JV, Lippman SM (2008) Lung cancer. N Engl J Med 359:1367–1380
73. Herrero R, Park JY, Forman D (2014) The fight against gastric cancer – the IARC Working Group report. Best Pract Res Clin Gastroenterol 28:1107–1114
74. Hirata M, Matsuda K (2017) Cross-sectional analysis of BioBank Japan Clinical Data: a large cohort of 200,000 patients with 47 common diseases. J Epidemiol 27:S9–S21
75. Holmstrom SR, Deering T, Swift GH, Poelwijk FJ, Mangelsdorf DJ, Kliewer SA, Macdonald RJ (2011) LRH-1 and PTF1-L coregulate an exocrine pancreas-specific transcriptional network for digestive function. Genes Dev 25:1674–1679
76. Houlston RS, Cheadle J, Dobbins SE, Tenesa A, Jones AM, Howarth K, Spain SL, Broderick P, Domingo E, Farrington S, Prendergast JG, Pittman AM, Theodoratou E, Smith CG, Olver B, Walther A, Barnetson RA, Churchman M, Jaeger EE, Penegar S, Barclay E, Martin L, Gorman M, Mager R, Johnstone E, Midgley R, Niittymaki I, Tuupanen S, Colley J, Idziaszczyk S, Consortium, C, Thomas HJ, Lucassen AM, Evans DG, Maher ER, Consortium, C, GROUP, C. C, GROUP, C. C, Maughan T, Dimas A, Dermitzakis E, Cazier JB, Aaltonen LA, Pharoah P, Kerr DJ, Carvajal-Carmona LG, Campbell H, Dunlop MG, Tomlinson IP (2010) Meta-analysis of three genome-wide association studies identifies susceptibility loci for colorectal cancer at 1q41, 3q26.2, 12q13.13 and 20q13.33. Nat Genet 42:973–977

77. Hsing AW, Chokkalingam AP (2006) Prostate cancer epidemiology. Front Biosci 11:1388–1413
78. Hu N, Wang C, Hu Y, Yang HH, Giffen C, Tang ZZ, Han XY, Goldstein AM, Emmert-Buck MR, Buetow KH, Taylor PR, Lee MP (2005) Genome-wide association study in esophageal cancer using GeneChip mapping 10K array. Cancer Res 65:2542–2546
79. Hu N, Wang Z, Song X, Wei L, Kim BS, Freedman ND, Baek J, Burdette L, Chang J, Chung C, Dawsey SM, Ding T, Gao YT, Giffen C, Han Y, Hong M, Huang J, Kim HS, Koh WP, Liao LM, Mao YM, Qiao YL, Shu XO, Tan W, Wang C, Wu C, Wu MJ, Xiang YB, Yeager M, Yook JH, Yuan JM, Zhang P, Zhao XK, Zheng W, Song K, Wang LD, Lin D, Chanock SJ, Goldstein AM, Taylor PR, Abnet CC (2016) Genome-wide association study of gastric adenocarcinoma in Asia: a comparison of associations between cardia and non-cardia tumours. Gut 65:1611–1618
80. Hu Z, Wu C, Shi Y, Guo H, Zhao X, Yin Z, Yang L, Dai J, Hu L, Tan W, Li Z, Deng Q, Wang J, Wu W, Jin G, Jiang Y, Yu D, Zhou G, Chen H, Guan P, Chen Y, Shu Y, Xu L, Liu X, Liu L, Xu P, Han B, Bai C, Zhao Y, Zhang H, Yan Y, Ma H, Chen J, Chu M, Lu F, Zhang Z, Chen F, Wang X, Jin L, Lu J, Zhou B, Lu D, Wu T, Lin D, Shen H (2011) A genome-wide association study identifies two new lung cancer susceptibility loci at 13q12.12 and 22q12.2 in Han Chinese. Nat Genet 43:792–796
81. Hung RJ, Mckay JD, Gaborieau V, Boffetta P, Hashibe M, Zaridze D, Mukeria A, Szeszenia-Dabrowska N, Lissowska J, Rudnai P, Fabianova E, Mates D, Bencko V, Foretova L, Janout V, Chen C, Goodman G, Field JK, Liloglou T, Xinarianos G, Cassidy A, Mclaughlin J, Liu G, Narod S, Krokan HE, Skorpen F, Elvestad MB, Hveem K, Vatten L, Linseisen J, Clavel-Chapelon F, Vineis P, Bueno-De-Mesquita HB, Lund E, Martinez C, Bingham S, Rasmuson T, Hainaut P, Riboli E, Ahrens W, Benhamou S, Lagiou P, Trichopoulos D, Holcatova I, Merletti F, Kjaerheim K, Agudo A, Macfarlane G, Talamini R, Simonato L, Lowry R, Conway DI, Znaor A, Healy C, Zelenika D, Boland A, Delepine M, Foglio M, Lechner D, Matsuda F, Blanche H, Gut I, Heath S, Lathrop M, Brennan P (2008) A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. Nature 452:633–637
82. Hunter DJ, Kraft P, Jacobs KB, Cox DG, Yeager M, Hankinson SE, Wacholder S, Wang Z, Welch R, Hutchinson A, Wang J, Yu K, Chatterjee N, Orr N, Willett WC, Colditz GA, Ziegler RG, Berg CD, Buys SS, Mccarty CA, Feigelson HS, Calle EE, Thun MJ, Hayes RB, Tucker M, Gerhard DS, Fraumeni JF Jr, Hoover RN, Thomas G, Chanock SJ (2007) A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. Nat Genet 39:870–874
83. Jaeger E, Webb E, Howarth K, Carvajal-Carmona L, Rowan A, Broderick P, Walther A, Spain S, Pittman A, Kemp Z, Sullivan K, Heinimann K, Lubbe S, Domingo E, Barclay E, Martin L, Gorman M, Chandler I, Vijayakrishnan J, Wood W, Papaemmanuil E, Penegar S, Qureshi M, Consortium C, Farrington S, Tenesa A, Cazier JB, Kerr D, Gray R, Peto J, Dunlop M, Campbell H, Thomas H, Houlston R, Tomlinson I (2008) Common genetic variants at the CRAC1 (HMPS) locus on chromosome 15q13.3 influence colorectal cancer risk. Nat Genet 40:26–28
84. Jawad MU, Cheung MC, Min ES, Schneiderbauer MM, Koniaris LG, Scully SP (2009) Ewing sarcoma demonstrates racial disparities in incidence-related and sex-related differences in outcome: an analysis of 1631 cases from the SEER database, 1973-2005. Cancer 115:3526–3536
85. Jervis S, Song H, Lee A, Dicks E, Tyrer J, Harrington P, Easton DF, Jacobs IJ, Pharoah PP, Antoniou AC (2014) Ovarian cancer familial relative risks by tumour subtypes and by known ovarian cancer genetic susceptibility variants. J Med Genet 51:108–113
86. Jia WH, Zhang B, Matsuo K, Shin A, Xiang YB, Jee SH, Kim DH, Ren Z, Cai Q, Long J, Shi J, Wen W, Yang G, Delahanty RJ, Genetics, Epidemiology of Colorectal Cancer, C, Colon Cancer Family, R, Ji BT, Pan ZZ, Matsuda F, Gao YT, Oh JH, Ahn YO, Park EJ, Li HL, Park JW, Jo J, Jeong JY, Hosono S, Casey G, Peters U, Shu XO, Zeng YX, Zheng W (2013) Genome-wide association analyses in East Asians identify new susceptibility loci for colorectal cancer. Nat Genet 45:191–196

87. Jiang DK, Sun J, Cao G, Liu Y, Lin D, Gao YZ, Ren WH, Long XD, Zhang H, Ma XP, Wang Z, Jiang W, Chen TY, Gao Y, Sun LD, Long JR, Huang HX, Wang D, Yu H, Zhang P, Tang LS, Peng B, Cai H, Liu TT, Zhou P, Liu F, Lin X, Tao S, Wan B, Sai-Yin HX, Qin LX, Yin J, Liu L, Wu C, Pei Y, Zhou YF, Zhai Y, Lu PX, Tan A, Zuo XB, Fan J, Chang J, Gu X, Wang NJ, Li Y, Liu YK, Zhai K, Zhang H, Hu Z, Liu J, Yi Q, Xiang Y, Shi R, Ding Q, Zheng W, Shu XO, Mo Z, Shugart YY, Zhang XJ, Zhou G, Shen H, Zheng SL, Xu J, Yu L (2013) Genetic variants in STAT4 and HLA-DQ genes confer risk of hepatitis B virus-related hepatocellular carcinoma. Nat Genet 45:72–75
88. Jin G, Ma H, Wu C, Dai J, Zhang R, Shi Y, Lu J, Miao X, Wang M, Zhou Y, Chen J, Li H, Pan S, Chu M, Lu F, Yu D, Jiang Y, Dong J, Hu L, Chen Y, Xu L, Shu Y, Pan S, Tan W, Zhou B, Lu D, Wu T, Zhang Z, Chen F, Wang X, Hu Z, Lin D, Shen H (2012) Genetic variants at 6p21.1 and 7p15.3 are associated with risk of multiple cancers in Han Chinese. Am J Hum Genet 91:928–934
89. Jinushi M, Takehara T, Tatsumi T, Kanto T, Groh V, Spies T, Kimura R, Miyagi T, Mochizuki K, Sasaki Y, Hayashi N (2003) Expression and role of MICA and MICB in human hepatocellular carcinomas and their regulation by retinoic acid. Int J Cancer 104:354–361
90. Johns LE, Houlston RS (2001) A systematic review and meta-analysis of familial colorectal cancer risk. Am J Gastroenterol 96:2992–3003
91. Joyce MJ, Harmon DC, Mankin HJ, Suit HD, Schiller AL, Truman JT (1984) Ewing's sarcoma in female siblings. A clinical report and review of the literature. Cancer 53:1959–1962
92. Kalariti N, Lembessis P, Koutsilieris M (2004) Characterization of the glutametergic system in MG-63 osteoblast-like osteosarcoma cells. Anticancer Res 24:3923–3929
93. Kamangar F, Dores GM, Anderson WF (2006) Patterns of cancer incidence, mortality, and prevalence across five continents: defining priorities to reduce cancer disparities in different geographic regions of the world. J Clin Oncol 24:2137–2150
94. Kanetsky PA, Mitra N, Vardhanabhuti S, Li M, Vaughn DJ, Letrero R, Ciosek SL, Doody DR, Smith LM, Weaver J, Albano A, Chen C, Starr JR, Rader DJ, Godwin AK, Reilly MP, Hakonarson H, Schwartz SM, Nathanson KL (2009) Common variation in KITLG and at 5q31.3 predisposes to testicular germ cell cancer. Nat Genet 41:811–815
95. Kanetsky PA, Mitra N, Vardhanabhuti S, Vaughn DJ, Li M, Ciosek SL, Letrero R, D'andrea K, Vaddi M, Doody DR, Weaver J, Chen C, Starr JR, Hakonarson H, Rader DJ, Godwin AK, Reilly MP, Schwartz SM, Nathanson KL (2011) A second independent locus within DMRT1 is associated with testicular germ cell tumor susceptibility. Hum Mol Genet 20:3109–3117
96. Kelsen DP, Portenoy R, Thaler H, Tao Y, Brennan M (1997) Pain as a predictor of outcome in patients with operable pancreatic carcinoma. Surgery 122:53–59
97. Kemmer K, Corless CL, Fletcher JA, Mcgreevey L, Haley A, Griffith D, Cummings OW, Wait C, Town A, Heinrich MC (2004) KIT mutations are common in testicular seminomas. Am J Pathol 164:305–313
98. Kiemeney LA, Sulem P, Besenbacher S, Vermeulen SH, Sigurdsson A, Thorleifsson G, Gudbjartsson DF, Stacey SN, Gudmundsson J, Zanon C, Kostic J, Masson G, Bjarnason H, Palsson ST, Skarphedinsson OB, Gudjonsson SA, Witjes JA, Grotenhuis AJ, Verhaegh GW, Bishop DT, Sak SC, Choudhury A, Elliott F, Barrett JH, Hurst CD, de Verdier PJ, Ryk C, Rudnai P, Gurzau E, Koppova K, Vineis P, Polidoro S, Guarrera S, Sacerdote C, Campagna M, Placidi D, Arici C, Zeegers MP, Kellen E, Gutierrez BS, Sanz-Velez JI, Sanchez-Zalabardo M, Valdivia G, Garcia-Prats MD, Hengstler JG, Blaszkewicz M, Dietrich H, Ophoff RA, van den Berg LH, Alexiusdottir K, Kristjansson K, Geirsson G, Nikulasson S, Petursdottir V, Kong A, Thorgeirsson T, Mungan NA, Lindblom A, Van Es MA, Porru S, Buntinx F, Golka K, Mayordomo JI, Kumar R, Matullo G, Steineck G, Kiltie AE, Aben KK, Jonsson E, Thorsteinsdottir U, Knowles MA, Rafnar T, Stefansson K (2010) A sequence variant at 4p16.3 confers susceptibility to urinary bladder cancer. Nat Genet 42:415–419
99. Kiemeney LA, Thorlacius S, Sulem P, Geller F, Aben KK, Stacey SN, Gudmundsson J, Jakobsdottir M, Bergthorsson JT, Sigurdsson A, Blondal T, Witjes JA, Vermeulen SH, Hulsbergen-Van DE Kaa CA, Swinkels DW, Ploeg M, Cornel EB, Vergunst H, Thorgeirsson

TE, Gudbjartsson D, Gudjonsson SA, Thorleifsson G, Kristinsson KT, Mouy M, Snorradottir S, Placidi D, Campagna M, Arici C, Koppova K, Gurzau E, Rudnai P, Kellen E, Polidoro S, Guarrera S, Sacerdote C, Sanchez M, Saez B, Valdivia G, Ryk C, de Verdier P, Lindblom A, Golka K, Bishop DT, Knowles MA, Nikulasson S, Petursdottir V, Jonsson E, Geirsson G, Kristjansson B, Mayordomo JI, Steineck G, Porru S, Buntinx F, Zeegers MP, Fletcher T, Kumar R, Matullo G, Vineis P, Kiltie AE, Gulcher JR, Thorsteinsdottir U, Kong A, Rafnar T, Stefansson K (2008) Sequence variant on 8q24 confers susceptibility to urinary bladder cancer. Nat Genet 40:1307–1312

100. Kobayashi M, Arase Y, Ikeda K, Tsubota A, Suzuki Y, Saitoh S, Kobayashi M, Suzuki F, Akuta N, Someya T, Matsuda M, Sato J, Takagi K, Miyakawa Y, Kumada H (2002) Viral genotypes and response to interferon in patients with acute prolonged hepatitis B virus infection of adulthood in Japan. J Med Virol 68:522–528
101. Kogevinas M, T'Mannetje A, Cordier S, Ranft U, Gonzalez CA, Vineis P, Chang-Claude J, Lynge E, Wahrendorf J, Tzonou A, Jockel KH, Serra C, Porru S, Hours M, Greiser E, Boffetta P (2003) Occupation and bladder cancer among men in Western Europe. Cancer Causes Control: CCC 14:907–914
102. Kristiansen W, Karlsson R, Rounge TB, Whitington T, Andreassen BK, Magnusson PK, Fossa SD, Adami HO, Turnbull C, Haugen TB, Grotmol T, Wiklund F (2015) Two new loci and gene sets related to sex determination and cancer progression are associated with susceptibility to testicular germ cell tumor. Hum Mol Genet 24:4138–4146
103. Kuchenbaecker KB, Ramus SJ, Tyrer J, Lee A, Shen HC, Beesley J, Lawrenson K, Mcguffog L, Healey S, Lee JM, Spindler TJ, Lin YG, Pejovic T, Bean Y, Li Q, Coetzee S, Hazelett D, Miron A, Southey M, Terry MB, Goldgar DE, Buys SS, Janavicius R, Dorfling CM, Van Rensburg EJ, Neuhausen SL, Ding YC, Hansen TV, Jonson L, Gerdes AM, Ejlertsen B, Barrowdale D, Dennis J, Benitez J, Osorio A, Garcia MJ, Komenaka I, Weitzel JN, Ganschow P, Peterlongo P, Bernard L, Viel A, Bonanni B, Peissel B, Manoukian S, Radice P, Papi L, Ottini L, Fostira F, Konstantopoulou I, Garber J, Frost D, Perkins J, Platte R, Ellis S, EMBRACE, Godwin AK, Schmutzler RK, Meindl A, Engel C, Sutter C, Sinilnikova OM, Collaborators, G. S, Damiola F, Mazoyer S, Stoppa-Lyonnet D, Claes K, de Leeneer K, Kirk J, Rodriguez GC, Piedmonte M, O'Malley DM, de la Hoya M, Caldes T, Aittomaki K, Nevanlinna H, Collee JM, Rookus MA, Oosterwijk JC, Breast Cancer Family, R, Tihomirova L, Tung N, Hamann U, Isaccs C, Tischkowitz M, Imyanitov EN, Caligo MA, Campbell IG, Hogervorst FB, HEBON, Olah E, Diez O, Blanco I, Brunet J, Lazaro C, Pujana MA, Jakubowska A, Gronwald J, Lubinski J, Sukiennicki G et al (2015) Identification of six new susceptibility loci for invasive epithelial ovarian cancer. Nat Genet 47:164–171
104. Kumar V, Kato N, Urabe Y, Takahashi A, Muroyama R, Hosono N, Otsuka M, Tateishi R, Omata M, Nakagawa H, Koike K, Kamatani N, Kubo M, Nakamura Y, Matsuda K (2011) Genome-wide association study identifies a susceptibility locus for HCV-induced hepatocellular carcinoma. Nat Genet 43:455–458
105. Lai CL, Ratziu V, Yuen MF, Poynard T (2003) Viral hepatitis B. Lancet 362:2089–2094
106. Lan Q, Hsiung CA, Matsuo K, Hong YC, Seow A, Wang Z, Hosgood HD 3rd, Chen K, Wang JC, Chatterjee N, Hu W, Wong MP, Zheng W, Caporaso N, Park JY, Chen CJ, Kim YH, Kim YT, Landi MT, Shen H, Lawrence C, Burdett L, Yeager M, Yuenger J, Jacobs KB, Chang IS, Mitsudomi T, Kim HN, Chang GC, Bassig BA, Tucker M, Wei F, Yin Z, Wu C, An SJ, Qian B, Lee VH, Lu D, Liu J, Jeon HS, Hsiao CF, Sung JS, Kim JH, Gao YT, Tsai YH, Jung YJ, Guo H, Hu Z, Hutchinson A, Wang WC, Klein R, Chung CC, Oh IJ, Chen KY, Berndt SI, He X, Wu W, Chang J, Zhang XC, Huang MS, Zheng H, Wang J, Zhao X, Li Y, Choi JE, Su WC, Park KH, Sung SW, Shu XO, Chen YM, Liu L, Kang CH, Hu L, Chen CH, Pao W, Kim YC, Yang TY, Xu J, Guan P, Tan W, Su J, Wang CL, Li H, Sihoe AD, Zhao Z, Chen Y, Choi YY, Hung JY, Kim JS, Yoon HI, Cai Q, Lin CC, Park IK, Xu P, Dong J, Kim C, He Q, Perng RP, Kohno T, Kweon SS et al (2012) Genome-wide association analysis identifies new lung cancer susceptibility loci in never-smoking women in Asia. Nat Genet 44:1330–1335

107. Landi MT, Chatterjee N, Yu K, Goldin LR, Goldstein AM, Rotunno M, Mirabello L, Jacobs K, Wheeler W, Yeager M, Bergen AW, Li Q, Consonni D, Pesatori AC, Wacholder S, Thun M, Diver R, Oken M, Virtamo J, Albanes D, Wang Z, Burdette L, Doheny KF, Pugh EW, Laurie C, Brennan P, Hung R, Gaborieau V, Mckay JD, Lathrop M, Mclaughlin J, Wang Y, Tsao MS, Spitz MR, Wang Y, Krokan H, Vatten L, Skorpen F, Arnesen E, Benhamou S, Bouchard C, Metspalu A, Vooder T, Nelis M, Valk K, Field JK, Chen C, Goodman G, Sulem P, Thorleifsson G, Rafnar T, Eisen T, Sauter W, Rosenberger A, Bickeboller H, Risch A, Chang-Claude J, Wichmann HE, Stefansson K, Houlston R, Amos CI, Fraumeni JF Jr, Savage SA, Bertazzi PA, Tucker MA, Chanock S, Caporaso NE (2009) A genome-wide association study of lung cancer identifies a region of chromosome 5p15 associated with risk for adenocarcinoma. Am J Hum Genet 85:679–691
108. Laskus T, Radkowski M, Lupa E, Horban A, Cianciara J, Slusarczyk J (1992) Prevalence of markers of hepatitis viruses in out-patient alcoholics. J Hepatol 15:174–178
109. Le Cornet C, Lortet-Tieulent J, Forman D, Beranger R, Flechon A, Fervers B, Schuz J, Bray F (2014) Testicular cancer incidence to rise by 25% by 2025 in Europe? Model-based predictions in 40 countries using population-based registry data. Eur J Cancer 50:831–839
110. Leongamornlert D, Mahmud N, Tymrakiewicz M, Saunders E, Dadaev T, Castro E, Goh C, Govindasami K, Guy M, O'Brien L, Sawyer E, Hall A, Wilkinson R, Easton D, Collaborators, U, Goldgar D, Eeles R, Kote-Jarai Z (2012) Germline BRCA1 mutations increase prostate cancer risk. Br J Cancer 106:1697–1701
111. Levine DM, Ek WE, Zhang R, Liu X, Onstad L, Sather C, Lao-Sirieix P, Gammon MD, Corley DA, Shaheen NJ, Bird NC, Hardie LJ, Murray LJ, Reid BJ, Chow WH, Risch HA, Nyren O, Ye W, Liu G, Romero Y, Bernstein L, Wu AH, Casson AG, Chanock SJ, Harrington P, Caldas I, Debiram-Beecham I, Caldas C, Hayward NK, Pharoah PD, Fitzgerald RC, Macgregor S, Whiteman DC, Vaughan TL (2013) A genome-wide association study identifies new susceptibility loci for esophageal adenocarcinoma and Barrett's esophagus. Nat Genet 45:1487–1493
112. Li M, Edamatsu H, Kitazawa R, Kitazawa S, Kataoka T (2009) Phospholipase Cepsilon promotes intestinal tumorigenesis of Apc(Min/+) mice through augmentation of inflammation and angiogenesis. Carcinogenesis 30:1424–1432
113. Li S, Qian J, Yang Y, Zhao W, Dai J, Bei JX, Foo JN, Mclaren PJ, Li Z, Yang J, Shen F, Liu L, Yang J, Li S, Pan S, Wang Y, Li W, Zhai X, Zhou B, Shi L, Chen X, Chu M, Yan Y, Wang J, Cheng S, Shen J, Jia W, Liu J, Yang J, Wen Z, Li A, Zhang Y, Zhang G, Luo X, Qin H, Chen M, Wang H, Jin L, Lin D, Shen H, He L, De Bakker PI, Wang H, Zeng YX, Wu M, Hu Z, Shi Y, Liu J, Zhou W (2012) GWAS identifies novel susceptibility loci on 6p21.32 and 21q21.3 for hepatocellular carcinoma in chronic hepatitis B virus carriers. PLoS Genet 8:e1002791
114. Lichtenstein P, Holm NV, Verkasalo PK, Iliadou A, Kaprio J, Koskenvuo M, Pukkala E, Skytthe A, Hemminki K (2000) Environmental and heritable factors in the causation of cancer--analyses of cohorts of twins from Sweden, Denmark, and Finland. N Engl J Med 343:78–85
115. Litchfield K, Holroyd A, Lloyd A, Broderick P, Nsengimana J, Eeles R, Easton DF, Dudakia D, Bishop DT, Reid A, Huddart RA, Grotmol T, Wiklund F, Shipley J, Houlston RS, Turnbull C (2015a) Identification of four new susceptibility loci for testicular germ cell tumour. Nat Commun 6:8690
116. Litchfield K, Shipley J, Turnbull C (2015b) Common variants identified in genome-wide association studies of testicular germ cell tumour: an update, biological insights and clinical application. Andrology 3:34–46
117. Litchfield K, Sultana R, Renwick A, Dudakia D, Seal S, Ramsay E, Powell S, Elliott A, Warren-Perry M, Eeles R, Peto J, Kote-Jarai Z, Muir K, Nsengimana J, UKTCC, Stratton MR, Easton DF, Bishop DT, Huddart RA, Rahman N, Turnbull C, UKTCC (2015c) Multistage genome-wide association study identifies new susceptibility locus for testicular germ cell tumour on chromosome 3q25. Hum Mol Genet 24:1169–1176

118. Liu JZ, Van Sommeren S, Huang H, Ng SC, Alberts R, Takahashi A, Ripke S, Lee JC, Jostins L, Shah T, Abedian S, Cheon JH, Cho J, Daryani NE, Franke L, Fuyuno Y, Hart A, Juyal RC, Juyal G, Kim WH, Morris AP, Poustchi H, Newman WG, Midha V, Orchard TR, Vahedi H, Sood A, Sung JJ, Malekzadeh R, Westra HJ, Yamazaki K, Yang SK, International Multiple Sclerosis Genetics, C, International, I. B. D. G. C, Barrett JC, Franke A, Alizadeh BZ, Parkes M, Bk T, Daly MJ, Kubo M, Anderson CA, Weersma RK (2015) Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. Nat Genet 47:979–986
119. Lowenfels AB, Maisonneuve P, Dimagno EP, Elitsur Y, Gates LK Jr, Perrault J, Whitcomb DC (1997) Hereditary pancreatitis and the risk of pancreatic cancer. International Hereditary Pancreatitis Study Group. J Natl Cancer Inst 89:442–446
120. Luksch H, Uckermann O, Stepulak A, Hendruschk S, Marzahn J, Bastian S, Staufner C, Temme A, Ikonomidou C (2011) Silencing of selected glutamate receptor subunits modulates cancer growth. Anticancer Res 31:3181–3192
121. Luo D, Gao Y, Wang S, Wang M, Wu D, Wang W, Xu M, Zhou J, Gong W, Tan Y, Zhang Z (2011) Genetic variation in PLCE1 is associated with gastric cancer survival in a Chinese population. J Gastroenterol 46:1260–1266
122. Lynch HT, Fusaro RM, Lynch JF, Brand R (2008) Pancreatic cancer and the FAMMM syndrome. Familial Cancer 7:103–112
123. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, Mccarthy MI, Ramos EM, Cardon LR, Chakravarti A (2009) Finding the missing heritability of complex diseases. Nature 461:747–753
124. Marley AR, Nan H (2016) Epidemiology of colorectal cancer. Int J Mol Epidemiol Genet 7:105–114
125. Matakidou A, Eisen T, Houlston RS (2005) Systematic review of the relationship between family history and lung cancer risk. Br J Cancer 93:825–833
126. Matsuda T, Saika K (2009) Comparison of time trends in prostate cancer incidence (1973-2002) in Asia, from cancer incidence in five continents, Vols IV-IX. Jpn J Clin Oncol 39:468–469
127. Mavaddat N, Antoniou AC, Easton DF, Garcia-Closas M (2010) Genetic susceptibility to breast cancer. Mol Oncol 4:174–191
128. McWilliams RR, Rabe KG, Olswold C, de Andrade M, Petersen GM (2005) Risk of malignancy in first-degree relatives of patients with pancreatic carcinoma. Cancer 104:388–394
129. Michailidou K, Beesley J, Lindstrom S, Canisius S, Dennis J, Lush MJ, Maranian MJ, Bolla MK, Wang Q, Shah M, Perkins BJ, Czene K, Eriksson M, Darabi H, Brand JS, Bojesen SE, Nordestgaard BG, Flyger H, Nielsen SF, Rahman N, Turnbull C, BOCS, Fletcher O, Peto J, Gibson L, Dos-Santos-Silva I, Chang-Claude J, Flesch-Janys D, Rudolph A, Eilber U, Behrens S, Nevanlinna H, Muranen TA, Aittomaki K, Blomqvist C, Khan S, Aaltonen K, Ahsan H, Kibriya MG, Whittemore AS, John EM, Malone KE, Gammon MD, Santella RM, Ursin G, Makalic E, Schmidt DF, Casey G, Hunter DJ, Gapstur SM, Gaudet MM, Diver WR, Haiman CA, Schumacher F, Henderson BE, Le Marchand L, Berg CD, Chanock SJ, Figueroa J, Hoover RN, Lambrechts D, Neven P, Wildiers H, Van Limbergen E, Schmidt MK, Broeks A, Verhoef S, Cornelissen S, Couch FJ, Olson JE, Hallberg E, Vachon C, Waisfisz Q, Meijers-Heijboer H, Adank MA, van der Luijt RB, Li J, Liu J, Humphreys K, Kang D, Choi JY, Park SK, Yoo KY, Matsuo K, Ito H, Iwata H, Tajima K, Guenel P, Truong T, Mulot C, Sanchez M, Burwinkel B, Marme F, Surowy H, Sohn C, Wu AH, Tseng CC, van den Berg D, Stram DO, Gonzalez-Neira A et al (2015) Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer. Nat Genet 47:373–380
130. Miki D, Kubo M, Takahashi A, Yoon KA, Kim J, Lee GK, Zo JI, Lee JS, Hosono N, Morizono T, Tsunoda T, Kamatani N, Chayama K, Takahashi T, Inazawa J, Nakamura Y, Daigo Y (2010) Variation in TP63 is associated with lung adenocarcinoma susceptibility in Japanese and Korean populations. Nat Genet 42:893–896

131. Miki D, Ochi H, Hayes CN, Abe H, Yoshima T, Aikata H, Ikeda K, Kumada H, Toyota J, Morizono T, Tsunoda T, Kubo M, Nakamura Y, Kamatani N, Chayama K (2011) Variation in the DEPDC5 locus is associated with progression to hepatocellular carcinoma in chronic hepatitis C virus carriers. Nat Genet 43:797–800
132. Mirabello L, Yeager M, Mai PL, Gastier-Foster JM, Gorlick R, Khanna C, Patino-Garcia A, Sierrasesumaga L, Lecanda F, Andrulis IL, Wunder JS, Gokgoz N, Barkauskas DA, Zhang X, Vogt A, Jones K, Boland JF, Chanock SJ, Savage SA (2015) Germline TP53 variants and susceptibility to osteosarcoma. J Natl Cancer Inst 107
133. MORTALITY, G. B. D, CAUSES OF DEATH, C (2016) Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980-2015: a systematic analysis for the Global Burden of Disease Study 2015. Lancet 388:1459–1544
134. Murta-Nascimento C, Silverman DT, Kogevinas M, Garc A-Closas M, Rothman N, Tard NA, Garc A-Closas R, Serra C, Carrato A, Villanueva C (2007) Risk of bladder cancer associated with family history of cancer: do low-penetrance polymorphisms account for the increase in risk? Cancer Epidemiol Biomark Prev 16:1595–1600
135. Ng W, Loh A, Teixeira A, Pereira S, Swallow D (2008) Genetic regulation of MUC1 alternative splicing in human tissues. Br J Cancer 99:978–985
136. Okada Y, Wu D, Trynka G, Raj T, Terao C, Ikari K, Kochi Y, Ohmura K, Suzuki A, Yoshida S, Graham RR, Manoharan A, Ortmann W, Bhangale T, Denny JC, Carroll RJ, Eyler AE, Greenberg JD, Kremer JM, Pappas DA, Jiang L, Yin J, Ye L, Su DF, Yang J, Xie G, Keystone E, Westra HJ, Esko T, Metspalu A, Zhou X, Gupta N, Mirel D, Stahl EA, Diogo D, Cui J, Liao K, Guo MH, Myouzen K, Kawaguchi T, Coenen MJ, Van Riel PL, Van de Laar MA, Guchelaar HJ, Huizinga TW, Dieude P, Mariette X, Bridges SL Jr, Zhernakova A, Toes RE, Tak PP, Miceli-Richard C, Bang SY, Lee HS, Martin J, Gonzalez-Gay MA, Rodriguez-Rodriguez L, Rantapaa-Dahlqvist S, Arlestig L, Choi HK, Kamatani Y, Galan P, Lathrop M, CONSORTIUM, R, CONSORTIUM, G, Eyre S, Bowes J, Barton A, de Vries N, Moreland LW, Criswell LA, Karlson EW, Taniguchi A, Yamada R, Kubo M, Liu JS, Bae SC, Worthington J, Padyukov L, Klareskog L, Gregersen PK, Raychaudhuri S, Stranger BE, de Jager PL, Franke L, Visscher PM, Brown MA, Yamanaka H, Mimori T, Takahashi A, Xu H, Behrens TW, Siminovitch KA, Momohara S, Matsuda F, Yamamoto K, Plenge RM (2014) Genetics of rheumatoid arthritis contributes to biology and drug discovery. Nature 506:376–381
137. Ou L, Guo Y, Luo C, Wu X, Zhao Y, Cai X (2010) RNA interference suppressing PLCE1 gene expression decreases invasive power of human bladder cancer T24 cell line. Cancer Genet Cytogenet 200:110–119
138. Ozaki K, Ohnishi Y, Iida A, Sekine A, Yamada R, Tsunoda T, Sato H, Hori M, Nakamura Y, Tanaka T (2002) Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction. Nat Genet 32:650–654
139. Palmer AJ, Lochhead P, Hold GL, Rabkin CS, Chow WH, Lissowska J, Vaughan TL, Berry S, Gammon M, Risch H, El-Omar EM (2012) Genetic variation in C20orf54, PLCE1 and MUC1 and the risk of upper gastrointestinal cancers in Caucasian populations. Eur J Cancer Prev 21:541–544
140. Parkin DM, Bray F, Ferlay J, Pisani P (2005) Global cancer statistics, 2002. CA Cancer J Clin 55:74–108
141. Patin E, Kutalik Z, Guergnon J, Bibert S, Nalpas B, Jouanguy E, Munteanu M, Bousquet L, Argiro L, Halfon P, Boland A, Mullhaupt B, Semela D, Dufour JF, Heim MH, Moradpour D, Cerny A, Malinverni R, Hirsch H, Martinetti G, Suppiah V, Stewart G, Booth DR, George J, Casanova JL, Brechot C, Rice CM, Talal AH, Jacobson IM, Bourliere M, Theodorou I, Poynard T, Negro F, Pol S, Bochud PY, Abel L, SWISS HEPATITIS, C. C. S. G, INTERNATIONAL HEPATITIS, C. G. C, FRENCH, A. H. C. E. P. G. S. G (2012) Genome-wide association study identifies variants associated with progression of liver fibrosis from HCV infection. Gastroenterology 143:1244–52 e1-12

142. Peters U, Jiao S, Schumacher FR, Hutter CM, Aragaki AK, Baron JA, Berndt SI, Bezieau S, Brenner H, Butterbach K, Caan BJ, Campbell PT, Carlson CS, Casey G, Chan AT, Chang-Claude J, Chanock SJ, Chen LS, Coetzee GA, Coetzee SG, Conti DV, Curtis KR, Duggan D, Edwards T, Fuchs CS, Gallinger S, Giovannucci EL, Gogarten SM, Gruber SB, Haile RW, Harrison TA, Hayes RB, Henderson BE, Hoffmeister M, Hopper JL, Hudson TJ, Hunter DJ, Jackson RD, Jee SH, Jenkins MA, Jia WH, Kolonel LN, Kooperberg C, Kury S, Lacroix AZ, Laurie CC, Laurie CA, Le Marchand L, Lemire M, Levine D, Lindor NM, Liu Y, Ma J, Makar KW, Matsuo K, Newcomb PA, Potter JD, Prentice RL, Qu C, Rohan T, Rosse SA, Schoen RE, Seminara D, Shrubsole M, Shu XO, Slattery ML, Taverna D, Thibodeau SN, Ulrich CM, White E, Xiang Y, Zanke BW, Zeng YX, Zhang B, Zheng W, Hsu L, Colon Cancer Family, R, The, G. & EPIDEMIOLOGY OF COLORECTAL CANCER, C (2013) Identification of genetic susceptibility loci for colorectal tumors in a genome-wide meta-analysis. Gastroenterology 144:799–807 e24
143. Petersen GM, Amundadottir L, Fuchs CS, Kraft P, Stolzenberg-Solomon RZ, Jacobs KB, Arslan AA, Bueno-de-Mesquita HB, Gallinger S, Gross M, Helzlsouer K, Holly EA, Jacobs EJ, Klein AP, Lacroix A, Li D, Mandelson MT, Olson SH, Risch HA, Zheng W, Albanes D, Bamlet WR, Berg CD, Boutron-Ruault MC, Buring JE, Bracci PM, Canzian F, Clipp S, Cotterchio M, de Andrade M, Duell EJ, Gaziano JM, Giovannucci EL, Goggins M, Hallmans G, Hankinson SE, Hassan M, Howard B, Hunter DJ, Hutchinson A, Jenab M, Kaaks R, Kooperberg C, Krogh V, Kurtz RC, Lynch SM, Mcwilliams RR, Mendelsohn JB, Michaud DS, Parikh H, Patel AV, Peeters PH, Rajkovic A, Riboli E, Rodriguez L, Seminara D, Shu XO, Thomas G, Tjonneland A, Tobias GS, Trichopoulos D, van den Eeden SK, Virtamo J, Wactawski-Wende J, Wang Z, Wolpin BM, Yu H, Yu K, Zeleniuch-Jacquotte A, Fraumeni JF Jr, Hoover RN, Hartge P, Chanock SJ (2010) A genome-wide association study identifies pancreatic cancer susceptibility loci on chromosomes 13q22.1, 1q32.1 and 5p15.33. Nat Genet 42:224–228
144. Pharoah PD, Day NE, Duffy S, Easton DF, Ponder BA (1997) Family history and the risk of breast cancer: a systematic review and meta-analysis. Int J Cancer 71:800–809
145. Pharoah PD, Tsai YY, Ramus SJ, Phelan CM, Goode EL, Lawrenson K, Buckley M, Fridley BL, Tyrer JP, Shen H, Weber R, Karevan R, Larson MC, Song H, Tessier DC, Bacot F, Vincent D, Cunningham JM, Dennis J, Dicks E, Australian Cancer, S, Australian Ovarian Cancer Study, G, Aben KK, Anton-Culver H, Antonenkova N, Armasu SM, Baglietto L, Bandera EV, Beckmann MW, Birrer MJ, Bloom G, Bogdanova N, Brenton JD, Brinton LA, Brooks-Wilson A, Brown R, Butzow R, Campbell I, Carney ME, Carvalho RS, Chang-Claude J, Chen YA, Chen Z, Chow WH, Cicek MS, Coetzee G, Cook LS, Cramer DW, Cybulski C, Dansonka-Mieszkowska A, Despierre E, Doherty JA, Dork T, Du Bois A, Durst M, Eccles D, Edwards R, Ekici AB, Fasching PA, Fenstermacher D, Flanagan J, Gao YT, Garcia-Closas M, Gentry-Maharaj A, Giles G, Gjyshi A, Gore M, Gronwald J, Guo Q, Halle MK, Harter P, Hein A, Heitz F, Hillemanns P, Hoatlin M, Hogdall E, Hogdall CK, Hosono S, Jakubowska A, Jensen A, Kalli KR, Karlan BY, Kelemen LE, Kiemeney LA, Kjaer SK, Konecny GE, Krakstad C, Kupryjanczyk J, Lambrechts D, Lambrechts S, Le ND, Lee N, Lee J, Leminen A, Lim BK, Lissowska J, Lubinski J, Lundvall L, Lurie G, Massuger LF et al (2013) GWAS meta-analysis and replication identifies three new susceptibility loci for ovarian cancer. Nat Genet 45:362–370. 370e1-2
146. Postel-Vinay S, Veron AS, Tirode F, Pierron G, Reynaud S, Kovar H, Oberlin O, Lapouble E, Ballet S, Lucchesi C, Kontny U, Gonzalez-NEIRA A, Picci P, Alonso J, Patino-Garcia A, De Paillerets BB, Laud K, Dina C, Froguel P, Clavel-Chapelon F, Doz F, Michon J, Chanock SJ, Thomas G, Cox DG, Delattre O (2012) Common variants near TARDBP and EGR2 are associated with susceptibility to Ewing sarcoma. Nat Genet 44:323–327
147. Pungpapong S, Kim WR, Poterucha JJ (2007) Natural history of hepatitis B virus infection: an update for clinicians. Mayo Clin Proc 82:967–975
148. Puranam KL, Blackshear PJ (1994) Cloning and characterization of RECQL, a potential human homologue of the *Escherichia coli* DNA helicase RecQ. J Biol Chem 269:29838–29845

149. Radstake TR, Gorlova O, Rueda B, Martin JE, Alizadeh BZ, Palomino-Morales R, Coenen MJ, Vonk MC, Voskuyl AE, Schuerwegh AJ, Broen JC, Van Riel PL, Van't Slot R, Italiaander A, Ophoff RA, Riemekasten G, Hunzelmann N, Simeon CP, Ortego-Centeno N, Gonzalez-Gay MA, Gonzalez-Escribano MF, Spanish Scleroderma, G, Airo P, Van Laar J, Herrick A, Worthington J, Hesselstrand R, Smith V, De Keyser F, Houssiau F, Chee MM, Madhok R, Shiels P, Westhovens R, Kreuter A, Kiener H, de Baere E, Witte T, Padykov L, Klareskog L, Beretta L, Scorza R, Lie BA, Hoffmann-Vold AM, Carreira P, Varga J, Hinchcliff M, Gregersen PK, Lee AT, Ying J, Han Y, Weng SF, Amos CI, Wigley FM, Hummers L, Nelson JL, Agarwal SK, Assassi S, Gourh P, Tan FK, Koeleman BP, Arnett FC, Martin J, Mayes MD (2010) Genome-wide association study of systemic sclerosis identifies CD247 as a new susceptibility locus. Nat Genet 42:426–429
150. Rafnar T, Gudbjartsson DF, Sulem P, Jonasdottir A, Sigurdsson A, Jonasdottir A, Besenbacher S, Lundin P, Stacey SN, Gudmundsson J, Magnusson OT, le Roux L, Orlygsdottir G, Helgadottir HT, Johannsdottir H, Gylfason A, Tryggvadottir L, Jonasson JG, de Juan A, Ortega E, Ramon-Cajal JM, Garcia-Prats MD, Mayordomo C, Panadero A, Rivera F, Aben KK, van Altena AM, Massuger LF, Aavikko M, Kujala PM, STAFF, S, Aaltonen LA, Olafsdottir K, Bjornsson J, Kong A, Salvarsdottir A, Saemundsson H, Olafsson K, Benediktsdottir KR, Gulcher J, Masson G, Kiemeney LA, Mayordomo JI, Thorsteinsdottir U, Stefansson K (2011a) Mutations in BRIP1 confer high risk of ovarian cancer. Nat Genet 43:1104–1107
151. Rafnar T, Vermeulen SH, Sulem P, Thorleifsson G, Aben KK, Witjes JA, Grotenhuis AJ, Verhaegh GW, Hulsbergen-van de Kaa CA, Besenbacher S, Gudbjartsson D, Stacey SN, Gudmundsson J, Johannsdottir H, Bjarnason H, Zanon C, Helgadottir H, Jonasson JG, Tryggvadottir L, Jonsson E, Geirsson G, Nikulasson S, Petursdottir V, Bishop DT, Chung-Sak S, Choudhury A, Elliott F, Barrett JH, Knowles MA, de Verdier PJ, Ryk C, Lindblom A, Rudnai P, Gurzau E, Koppova K, Vineis P, Polidoro S, Guarrera S, Sacerdote C, Panadero A, Sanz-Velez JI, Sanchez M, Valdivia G, Garcia-Prats MD, Hengstler JG, Selinski S, Gerullis H, Ovsiannikov D, Khezri A, Aminsharifi A, Malekzadeh M, van den Berg LH, Ophoff RA, Veldink JH, Zeegers MP, Kellen E, Fostinelli J, Andreoli D, Arici C, Porru S, Buntinx F, Ghaderi A, Golka K, Mayordomo JI, Matullo G, Kumar R, Steineck G, Kiltie AE, Kong A, Thorsteinsdottir U, Stefansson K, Kiemeney LA (2011b) European genome-wide association study identifies SLC14A1 as a new urinary bladder cancer susceptibility gene. Hum Mol Genet 20:4268–4281
152. Rapley EA, Turnbull C, al Olama AA, Dermitzakis ET, Linger R, Huddart RA, Renwick A, Hughes D, Hines S, Seal S, Morrison J, Nsengimana J, Deloukas P, COLLABORATION, U. K. T. C, Rahman N, Bishop DT, Easton DF, Stratton MR (2009) A genome-wide association study of testicular germ cell tumor. Nat Genet 41:807–810
153. Risch A, Wallace DM, Bathers S, Sim E (1995) Slow N-acetylation genotype is a susceptibility factor in occupational and smoking related bladder cancer. Hum Mol Genet 4:231–236
154. Romano RA, Li H, Tummala R, Maul R, Sinha S (2004) Identification of Basonuclin2, a DNA-binding zinc-finger protein expressed in germ tissues and skin keratinocytes. Genomics 83:821–833
155. Roskoski R Jr (2005) Signaling by Kit protein-tyrosine kinase--the stem cell factor receptor. Biochem Biophys Res Commun 337:1–13
156. Rothman N, Garcia-Closas M, Chatterjee N, Malats N, Wu X, Figueroa JD, Real FX, van den Berg D, Matullo G, Baris D, Thun M, Kiemeney LA, Vineis P, de Vivo I, Albanes D, Purdue MP, Rafnar T, Hildebrandt MA, Kiltie AE, Cussenot O, Golka K, Kumar R, Taylor JA, Mayordomo JI, Jacobs KB, Kogevinas M, Hutchinson A, Wang Z, Fu YP, Prokunina-Olsson L, Burdett L, Yeager M, Wheeler W, Tardon A, Serra C, Carrato A, Garcia-Closas R, Lloreta J, Johnson A, Schwenn M, Karagas MR, Schned A, Andriole G Jr, Grubb R 3rd, Black A, Jacobs EJ, Diver WR, Gapstur SM, Weinstein SJ, Virtamo J, Cortessis VK, Gago-Dominguez M, Pike MC, Stern MC, Yuan JM, Hunter DJ, Mcgrath M, Dinney CP, Czerniak B, Chen M, Yang H, Vermeulen SH, Aben KK, Witjes JA, Makkinje RR, Sulem P, Besenbacher S, Stefansson K, Riboli E, Brennan P, Panico S, Navarro C, Allen NE, Bueno-de-Mesquita HB,

Trichopoulos D, Caporaso N, Landi MT, Canzian F, Ljungberg B, Tjonneland A, Clavel-Chapelon F, Bishop DT, Teo MT, Knowles MA, Guarrera S, Polidoro S, Ricceri F, Sacerdote C, Allione A, Cancel-Tassin G, Selinski S, Hengstler JG, Dietrich H, Fletcher T, Rudnai P, Gurzau E, Koppova K, Bolick SC, Godfrey A, Xu Z et al (2010) A multi-stage genome-wide association study of bladder cancer identifies multiple susceptibility loci. Nat Genet 42:978–984

157. Ruark E, Seal S, Mcdonald H, Zhang F, Elliot A, Lau K, Perdeaux E, Rapley E, Eeles R, Peto J, Kote-Jarai Z, Muir K, Nsengimana J, Shipley J, COLLABORATION, U. K. T. C, Bishop DT, Stratton MR, Easton DF, Huddart RA, Rahman N, Turnbull C (2013) Identification of nine new susceptibility loci for testicular cancer, including variants near DAZL and PRDM14. Nat Genet 45:686–689
158. Saeki N, Gu J, Yoshida T, Wu X (2010) Prostate stem cell antigen: a Jekyll and Hyde molecule? Clin Cancer Res. United States: 2010 Aacr 16:3533–3538
159. Saeki N, Saito A, Choi IJ, Matsuo K, Ohnami S, Totsuka H, Chiku S, Kuchiba A, Lee YS, Yoon KA, Kook MC, Park SR, KIM YW, Tanaka H, Tajima K, Hirose H, Tanioka F, Matsuno Y, Sugimura H, Kato S, Nakamura T, Nishina T, Yasui W, Aoyagi K, Sasaki H, Yanagihara K, Katai H, Shimoda T, Yoshida T, Nakamura Y, Hirohashi S, Sakamoto H (2011) A functional single nucleotide polymorphism in mucin 1, at chromosome 1q22, determines susceptibility to diffuse-type gastric cancer. Gastroenterology 140:892–902
160. Sakamoto H, Yoshimura K, Saeki N, Katai H, Shimoda T, Matsuno Y, Saito D, Sugimura H, Tanioka F, Kato S, Matsukura N, Matsuda N, Nakamura T, Hyodo I, Nishina T, Yasui W, Hirose H, Hayashi M, Toshiro E, Ohnami S, Sekine A, Sato Y, Totsuka H, Ando M, Takemura R, Takahashi Y, Ohdaira M, Aoki K, Honmyo I, Chiku S, Aoyagi K, Sasaki H, Yanagihara K, Yoon KA, Kook MC, Lee YS, Park SR, Kim CG, Choi IJ, Yoshida T, Nakamura Y, Hirohashi S (2008) Genetic variation in PSCA is associated with susceptibility to diffuse-type gastric cancer. Nat Genet 40:730–740
161. Savage SA, Mirabello L, Wang Z, Gastier-Foster JM, Gorlick R, Khanna C, Flanagan AM, Tirabosco R, Andrulis IL, Wunder JS, Gokgoz N, Patino-Garcia A, Sierrasesumaga L, Lecanda F, Kurucu N, Ilhan IE, Sari N, Serra M, Hattinger C, Picci P, Spector LG, Barkauskas DA, Marina N, de Toledo SR, Petrilli AS, Amary MF, Halai D, Thomas DM, Douglass C, Meltzer PS, Jacobs K, Chung CC, Berndt SI, Purdue MP, Caporaso NE, Tucker M, Rothman N, Landi MT, Silverman DT, Kraft P, Hunter DJ, Malats N, Kogevinas M, Wacholder S, Troisi R, Helman L, Fraumeni JF Jr, Yeager M, Hoover RN, Chanock SJ (2013) Genome-wide association study identifies two susceptibility loci for osteosarcoma. Nat Genet 45:799–803
162. Sawai H, Nishida N, Mbarek H, Matsuda K, Mawatari Y, Yamaoka M, Hige S, Kang JH, Abe K, Mochida S, Watanabe M, Kurosaki M, Asahina Y, Izumi N, Honda M, Kaneko S, Tanaka E, Matsuura K, Itoh Y, Mita E, Korenaga M, Hino K, Murawaki Y, Hiasa Y, Ide T, Ito K, Sugiyama M, Ahn SH, Han KH, Park JY, Yuen MF, Nakamura Y, Tanaka Y, Mizokami M, Tokunaga K (2012) No association for Chinese HBV-related hepatocellular carcinoma susceptibility SNP in other East Asian populations. BMC Med Genet 13:47
163. Schlisio S, Kenchappa RS, Vredeveld LC, George RE, Stewart R, Greulich H, Shahriari K, Nguyen NV, Pigny P, Dahia PL, Pomeroy SL, Maris JM, Look AT, Meyerson M, Peeper DS, Carter BD, Kaelin WG Jr (2008) The kinesin KIF1Bbeta acts downstream from EglN3 to induce apoptosis and is a potential 1p36 tumor suppressor. Genes Dev 22:884–893
164. Schmit SL, Schumacher FR, Edlund CK, Conti DV, Raskin L, Lejbkowicz F, Pinchev M, Rennert HS, Jenkins MA, Hopper JL, Buchanan DD, Lindor NM, Le Marchand L, Gallinger S, Haile RW, Newcomb PA, Huang SC, Rennert G, Casey G, Gruber SB (2014) A novel colorectal cancer risk locus at 4q32.2 identified from an international genome-wide association study. Carcinogenesis 35:2512–2519
165. Schumacher, F. R., Wang, Z., Skotheim, R. I., Koster, R., Chung, C. C., Hildebrandt, M. A., Kratz, C. P., Bakken, A. C., Bishop, D. T., Cook, M. B., Erickson, R. L., Fossa, S. D., Greene, M. H., Jacobs, K. B., Kanetsky, P. A., Kolonel, L. N., Loud, J. T., Korde, L. A., Le Marchand, L., Lewinger, J. P., Lothe, R. A., Pike, M. C., Rahman, N., Rubertone, M. V., Schwartz, S. M.,

Siegmund, K. D., Skinner, E. C., Turnbull, C., Van Den Berg, D. J., Wu, X., Yeager, M., Nathanson, K. L., Chanock, S. J., Cortessis, V. K., Mcglynn, K. A. 2013. Testicular germ cell tumor susceptibility associated with the UCK2 locus on chromosome 1q23. Hum Mol Genet, 22, 2748–2753

166. Shi J, Chatterjee N, Rotunno M, Wang Y, Pesatori AC, Consonni D, Li P, Wheeler W, Broderick P, Henrion M, Eisen T, Wang Z, Chen W, Dong Q, Albanes D, Thun M, Spitz MR, Bertazzi PA, Caporaso NE, Chanock SJ, Amos CI, Houlston RS, Landi MT (2012) Inherited variation at chromosome 12p13.33, including RAD52, influences the risk of squamous cell lung carcinoma. Cancer Discov 2:131–139
167. Shi Y, Hu Z, Wu C, Dai J, Li H, Dong J, Wang M, Miao X, Zhou Y, Lu F, Zhang H, Hu L, Jiang Y, Li Z, Chu M, Ma H, Chen J, Jin G, Tan W, Wu T, Zhang Z, Lin D, Shen H (2011) A genome-wide association study identifies new susceptibility loci for non-cardia gastric cancer at 3q13.31 and 5p13.1. Nat Genet 43:1215–1218
168. Shiraishi K, Kunitoh H, Daigo Y, Takahashi A, Goto K, Sakamoto H, Ohnami S, Shimada Y, Ashikawa K, Saito A, Watanabe S, Tsuta K, Kamatani N, Yoshida T, Nakamura Y, Yokota J, Kubo M, Kohno T (2012) A genome-wide association study identifies two new susceptibility loci for lung adenocarcinoma in the Japanese population. Nat Genet 44:900–903
169. Shiraishi K, Okada Y, Takahashi A, Kamatani Y, Momozawa Y, Ashikawa K, Kunitoh H, Matsumoto S, Takano A, Shimizu K, Goto A, Tsuta K, Watanabe S, Ohe Y, Watanabe Y, Goto Y, Nokihara H, Furuta K, Yoshida A, Goto K, Hishida T, Tsuboi M, Tsuchihara K, Miyagi Y, Nakayama H, Yokose T, Tanaka K, Nagashima T, Ohtaki Y, Maeda D, Imai K, Minamiya Y, Sakamoto H, Saito A, Shimada Y, Sunami K, Saito M, Inazawa J, Nakamura Y, Yoshida T, Yokota J, Matsuda F, Matsuo K, Daigo Y, Kubo M, Kohno T (2016) Association of variations in HLA class II and other loci with susceptibility to EGFR-mutated lung adenocarcinoma. Nat Commun 7:12451
170. Siegel R, Ma J, Zou Z, Jemal A (2014) Cancer statistics, 2014. CA Cancer J Clin 64:9–29
171. Siegel R, Naishadham D, Jemal A (2012) Cancer statistics, 2012. CA Cancer J Clin 62:10–29
172. Skol AD, Sasaki MM, Onel K (2016) The genetics of breast cancer risk in the post-genome era: thoughts on study design to move past BRCA and towards clinical relevance. Breast Cancer Res 18:99
173. Slattery ML, Kerber RA (1994) Family history of cancer and colon cancer risk: the Utah population database. J Natl Cancer Inst 86:1618–1626
174. Sobue T, Suzuki T, Fujimoto I, Matsuda M, Doi O, Mori T, Furuse K, Fukuoka M, Yasumitsu T, Kuwahara O et al (1994) Case-control study for lung cancer and cigarette smoking in Osaka, Japan: comparison with the results from Western Europe. Jpn J Cancer Res 85:464–473
175. Song H, Ramus SJ, Tyrer J, Bolton KL, Gentry-Maharaj A, Wozniak E, Anton-Culver H, Chang-Claude J, Cramer DW, Dicioccio R, Dork T, Goode EL, Goodman MT, Schildkraut JM, Sellers T, Baglietto L, Beckmann MW, Beesley J, Blaakaer J, Carney ME, Chanock S, Chen Z, Cunningham JM, Dicks E, Doherty JA, Durst M, Ekici AB, Fenstermacher D, Fridley BL, Giles G, Gore ME, De Vivo I, Hillemanns P, Hogdall C, Hogdall E, Iversen ES, Jacobs IJ, Jakubowska A, Li D, Lissowska J, Lubinski J, Lurie G, Mcguire V, Mclaughlin J, Medrek K, Moorman PG, Moysich K, Narod S, Phelan C, Pye C, Risch H, Runnebaum IB, Severi G, Southey M, Stram DO, Thiel FC, Terry KL, Tsai YY, Tworoger SS, Van Den Berg DJ, Vierkant RA, Wang-Gohrke S, Webb PM, Wilkens LR, Wu AH, Yang H, Brewster W, Ziogas A, Australian Cancer, S, Australian Ovarian Cancer Study, G, Ovarian Cancer Association, C, Houlston R, Tomlinson I, Whittemore AS, Rossing MA, Ponder BA, Pearce CL, Ness RB, Menon U, Kjaer SK, Gronwald J, Garcia-Closas M, Fasching PA, Easton DF, Chenevix-Trench G, Berchuck A, Pharoah PD, Gayther SA (2009) A genome-wide association study identifies a new ovarian cancer susceptibility locus on 9p22.2. Nat Genet 41:996–1000
176. Stacey SN, Manolescu A, Sulem P, Rafnar T, Gudmundsson J, Gudjonsson SA, Masson G, Jakobsdottir M, Thorlacius S, Helgason A, Aben KK, Strobbe LJ, Albers-Akkers MT, Swinkels DW, Henderson BE, Kolonel LN, Le Marchand L, Millastre E, Andres R, Godino J, Garcia-Prats MD, Polo E, Tres A, Mouy M, Saemundsdottir J, Backman VM, Gudmundsson

L, Kristjansson K, Bergthorsson JT, Kostic J, Frigge ML, Geller F, Gudbjartsson D, Sigurdsson H, Jonsdottir T, Hrafnkelsson J, Johannsson J, Sveinsson T, Myrdal G, Grimsson HN, Jonsson T, Von Holst S, Werelius B, Margolin S, Lindblom A, Mayordomo JI, Haiman CA, Kiemeney LA, Johannsson OT, Gulcher JR, Thorsteinsdottir U, Kong A, Stefansson K (2007) Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer. Nat Genet 39:865–869

177. Stratton JF, Pharoah P, Smith SK, Easton D, Ponder BA (1998) A systematic review and meta-analysis of family history and risk of ovarian cancer. Br J Obstet Gynaecol 105:493–499
178. Study, C, Houlston RS, Webb E, Broderick P, Pittman AM, DI Bernardo MC, Lubbe S, Chandler I, Vijayakrishnan J, Sullivan K, Penegar S, Colorectal Cancer Association Study, C, Carvajal-Carmona L, Howarth K, Jaeger E, Spain SL, Walther A, Barclay E, Martin L, Gorman M, Domingo E, Teixeira AS, CO, R. G. I. C, Kerr D, Cazier JB, Niittymaki I, Tuupanen S, Karhu A, Aaltonen LA, Tomlinson IP, Farrington SM, Tenesa A, Prendergast JG, Barnetson RA, Cetnarskyj R, Porteous ME, Pharoah PD, Koessler T, Hampe J, Buch S, Schafmayer C, Tepel J, Schreiber S, Volzke H, Chang-Claude J, Hoffmeister M, Brenner H, Zanke BW, Montpetit A, Hudson TJ, Gallinger S, Campbell H, Dunlop MG (2008) Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. Nat Genet 40:1426–1435
179. Takata R, Akamatsu S, Kubo M, Takahashi A, Hosono N, Kawaguchi T, Tsunoda T, Inazawa J, Kamatani N, Ogawa O, Fujioka T, Nakamura Y, Nakagawa H (2010) Genome-wide association study identifies five new susceptibility loci for prostate cancer in the Japanese population. Nat Genet 42:751–754
180. tamoto E, Tada M, Murakawa K, Takada M, Shindo G, Teramoto K, Matsunaga A, Komuro K, Kanai M, Kawakami A, Fujiwara Y, Kobayashi N, Shirata K, Nishimura N, Okushiba S, Kondo S, Hamada J, Yoshiki T, Moriuchi T, Katoh H (2004) Gene-expression profile changes correlated with tumor progression and lymph node metastasis in esophageal cancer. Clin Cancer Res 10:3629–3638
181. Tanikawa C, Urabe Y, Matsuo K, Kubo M, Takahashi A, Ito H, Tajima K, Kamatani N, Nakamura Y, Matsuda K (2012) A genome-wide association study identifies two susceptibility loci for duodenal ulcer in the Japanese population. Nat Genet 44(430–4):S1–S2
182. Tenesa A, Farrington SM, Prendergast JG, Porteous ME, Walker M, Haq N, Barnetson RA, Theodoratou E, Cetnarskyj R, Cartwright N, Semple C, Clark AJ, Reid FJ, Smith LA, Kavoussanakis K, Koessler T, Pharoah PD, Buch S, schafmayer C, Tepel J, Schreiber S, Volzke H, Schmidt CO, Hampe J, Chang-Claude J, Hoffmeister M, Brenner H, Wilkening S, Canzian F, Capella G, Moreno V, Deary IJ, Starr JM, Tomlinson IP, Kemp Z, Howarth K, Carvajal-Carmona L, Webb E, Broderick P, Vijayakrishnan J, Houlston RS, Rennert G, Ballinger D, Rozek L, Gruber SB, Matsuda K, Kidokoro T, Nakamura Y, Zanke BW, Greenwood CM, Rangrej J, Kustra R, Montpetit A, Hudson TJ, Gallinger S, Campbell H, Dunlop MG (2008) Genome-wide association scan identifies a colorectal cancer susceptibility locus on 11q23 and replicates risk loci at 8q24 and 18q21. Nat Genet 40:631–637
183. Thomas DL, Thio CL, Martin MP, Qi Y, Ge D, O'Huigin C, kidd J, Kidd K, Khakoo SI, Alexander G, Goedert JJ, Kirk GD, Donfield SM, Rosen HR, Tobler LH, Busch MP, Mchutchison JG, Goldstein DB, Carrington M (2009) Genetic variation in IL28B and spontaneous clearance of hepatitis C virus. Nature 461:798–801
184. Thompson D, Easton D, Breast Cancer Linkage, C (2001) Variation in cancer risks, by mutation position, in BRCA2 mutation carriers. Am J Hum Genet 68:410–419
185. Thorgeirsson TE, Gudbjartsson DF, Surakka I, Vink JM, Amin N, Geller F, Sulem P, Rafnar T, Esko T, Walter S, Gieger C, Rawal R, Mangino M, Prokopenko I, Magi R, Keskitalo K, Gudjonsdottir IH, Gretarsdottir S, Stefansson H, Thompson JR, Aulchenko YS, Nelis M, Aben KK, Den Heijer M, Dirksen A, Ashraf H, Soranzo N, Valdes AM, Steves C, Uitterlinden AG, Hofman A, Tonjes A, Kovacs P, Hottenga JJ, Willemsen G, Vogelzangs N, Doring A, Dahmen N, Nitz B, Pergadia ML, Saez B, De Diego V, Lezcano V, Garcia-Prats MD, Ripatti S, Perola M, Kettunen J, Hartikainen AL, Pouta A, Laitinen J, Isohanni M, Huei-Yi S, Allen

M, Krestyaninova M, Hall AS, Jones GT, Van Rij AM, Mueller T, Dieplinger B, Haltmayer M, Jonsson S, Matthiasson SE, Oskarsson H, Tyrfingsson T, Kiemeney LA, Mayordomo JI, Lindholt JS, Pedersen JH, Franklin WA, Wolf H, Montgomery GW, Heath AC, Martin NG, madden PA, Giegling I, Rujescu D, Jarvelin MR, Salomaa V, Stumvoll M, Spector TD, Wichmann HE, Metspalu A, Samani NJ, Penninx BW, Oostra BA, Boomsma DI, Tiemeier H, Van Duijn CM, Kaprio J, Gulcher JR, Mccarthy MI, Peltonen L, Thorsteinsdottir U, Stefansson K (2010) Sequence variants at CHRNB3-CHRNA6 and CYP2A6 affect smoking behavior. Nat Genet 42:448–453

186. Tomlinson I, Webb E, Carvajal-Carmona L, Broderick P, Kemp Z, Spain S, Penegar S, Chandler I, Gorman M, Wood W, Barclay E, Lubbe S, Martin L, Sellick G, Jaeger E, Hubner R, Wild R, Rowan A, Fielding S, Howarth K, Consortium, C, Silver A, Atkin W, Muir K, Logan R, Kerr D, Johnstone E, Sieber O, Gray R, Thomas H, Peto J, Cazier JB, Houlston R (2007) A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. Nat Genet 39:984–988
187. Tomlinson IP, Carvajal-Carmona LG, Dobbins SE, Tenesa A, Jones AM, Howarth K, Palles C, Broderick P, Jaeger EE, Farrington S, Lewis A, Prendergast JG, Pittman AM, Theodoratou E, Olver B, Walker M, Penegar S, Barclay E, Whiffin N, Martin L, Ballereau S, Lloyd A, Gorman M, Lubbe S, Consortium, C, Collaborators, C, Consortium, E, Howie B, Marchini J, Ruiz-Ponte C, Fernandez-Rozadilla C, Castells A, Carracedo A, Castellvi-Bel S, Duggan D, Conti D, Cazier JB, Campbell H, Sieber O, Lipton L, Gibbs P, Martin NG, Montgomery GW, Young J, Baird PN, Gallinger S, Newcomb P, Hopper J, Jenkins MA, Aaltonen LA, Kerr DJ, Cheadle J, Pharoah P, Casey G, Houlston RS, Dunlop MG (2011) Multiple common susceptibility variants near BMP pathway loci GREM1, BMP4, and BMP2 explain part of the missing heritability of colorectal cancer. PLoS Genet 7:e1002105
188. Tomlinson IP, Webb E, Carvajal-Carmona L, Broderick P, Howarth K, Pittman AM, Spain S, Lubbe S, Walther A, Sullivan K, Jaeger E, Fielding S, Rowan A, Vijayakrishnan J, Domingo E, Chandler I, Kemp Z, Qureshi M, Farrington SM, Tenesa A, Prendergast JG, Barnetson RA, Penegar S, Barclay E, Wood W, Martin L, Gorman M, Thomas H, Peto J, Bishop DT, Gray R, Maher ER, Lucassen A, Kerr D, Evans DG, Consortium, C, Schafmayer C, Buch S, Volzke H, Hampe J, Schreiber S, John U, Koessler T, Pharoah P, Van Wezel T, Morreau H, Wijnen JT, Hopper JL, Southey MC, Giles GG, Severi G, Castellvi-Bel S, Ruiz-Ponte C, Carracedo A, Castells A, Consortium E, Forsti A, Hemminki K, Vodicka P, Naccarati A, Lipton L, Ho JW, Cheng KK, Sham PC, Luk J, Agundez JA, Ladero JM, De La Hoya M, Caldes T, Niittymaki I, Tuupanen S, Karhu A, Aaltonen L, Cazier JB, Campbell H, Dunlop MG, Houlston RS (2008) A genome-wide association study identifies colorectal cancer susceptibility loci on chromosomes 10p14 and 8q23.3. Nat Genet 40:623–630
189. Torre LA, Bray F, Siegel RL, Ferlay J, Lortet-Tieulent J, Jemal A (2015) Global cancer statistics, 2012. CA Cancer J Clin 65:87–108
190. Trabert B, Zugna D, Richiardi L, Mcglynn KA, Akre O (2013) Congenital malformations and testicular germ cell tumors. Int J Cancer 133:1900–1904
191. Turnbull C, Rapley EA, Seal S, Pernet D, Renwick A, Hughes D, Ricketts M, Linger R, Nsengimana J, Deloukas P, Huddart RA, Bishop DT, Easton DF, Stratton MR, Rahman N, Collaboration, U. K. T. C (2010) Variants near DMRT1, TERT and ATF7IP are associated with testicular germ cell cancer. Nat Genet 42:604–607
192. Uemura N, Okamoto S, Yamamoto S, Matsumura N, Yamaguchi S, Yamakido M, Taniyama K, Sasaki N, Schlemper RJ (2001) Helicobacter pylori infection and the development of gastric cancer. N Engl J Med 345:784–789
193. Urabe Y, Ochi H, Kato N, Kumar V, Takahashi A, Muroyama R, Hosono N, Otsuka M, Tateishi R, Lo PH, Tanikawa C, omata M, Koike K, Miki D, Abe H, Kamatani N, Toyota J, Kumada H, Kubo M, Chayama K, Nakamura Y, Matsuda K (2013) A genome-wide association study of HCV-induced liver cirrhosis in the Japanese population identifies novel susceptibility loci at the MHC region. J Hepatol 58:875–882

194. Von Figura G, Morris JPT, Wright CV, Hebrok M (2014) Nr5a2 maintains acinar cell differentiation and constrains oncogenic Kras-mediated pancreatic neoplastic initiation. Gut 63:656–664
195. Vong S, Bell BP (2004) Chronic liver disease mortality in the United States, 1990-1998. Hepatology 39:476–483
196. Wang H, Burnett T, Kono S, Haiman CA, Iwasaki M, Wilkens LR, Loo LW, Van Den Berg D, Kolonel LN, Henderson BE, Keku TO, Sandler RS, Signorello LB, Blot WJ, Newcomb PA, Pande M, Amos CI, West DW, Bezieau S, Berndt SI, Zanke BW, Hsu L, Genetics, Epidemiology of Colorectal Cancer, C, Lindor NM, Haile RW, Hopper JL, Jenkins MA, Gallinger S, Casey G, Colon Cancer Family, R, Stenzel SL, Schumacher FR, Peters U, Gruber SB, Colorectal Transdisciplinary, S, Tsugane S, Stram DO, Le Marchand L (2014) Trans-ethnic genome-wide association study of colorectal cancer identifies a new susceptibility locus in VTI1A. Nat Commun 5:4613
197. Wang LD, Zhou FY, Li XM, Sun LD, Song X, Jin Y, Li JM, Kong GQ, Qi H, Cui J, Zhang LQ, Yang JZ, Li JL, Li XC, Ren JL, Liu ZC, Gao WJ, Yuan L, Wei W, Zhang YR, Wang WP, Sheyhidin I, Li F, Chen BP, Ren SW, Liu B, Li D, Ku JW, Fan ZM, Zhou SL, Guo ZG, Zhao XK, Liu N, Ai YH, Shen FF, Cui WY, Song S, Guo T, Huang J, yuan C, Huang J, Wu Y, Yue WB, Feng CW, Li HL, Wang Y, Tian JY, Lu Y, Yuan Y, Zhu WL, Liu M, Fu WJ, Yang X, Wang HJ, Han SL, Chen J, Han M, Wang HY, Zhang P, Li XM, Dong JC, Xing GL, Wang R, Guo M, Chang ZW, Liu HL, Guo L, Yuan ZQ, Liu H, Lu Q, Yang LQ, Zhu FG, Yang XF, Feng XS, Wang Z, Li Y, Gao SG, Qige Q, Bai LT, Yang WJ, Lei GY, Shen ZY, Chen LQ, Li EM, Xu LY, Wu ZY, Cao WK, Wang JP, Bao ZQ, Chen JL, Ding GC, Zhuang X, Zhou YF, Zheng HF, Zhang Z, Zuo XB, Dong ZM, Fan DM, He X, Wang J et al (2010) Genome-wide association study of esophageal squamous cell carcinoma in Chinese subjects identifies susceptibility loci at PLCE1 and C20orf54. Nat Genet 42:759–763
198. Wang Y, Broderick P, Webb E, Wu X, Vijayakrishnan J, Matakidou A, Qureshi M, Dong Q, Gu X, Chen WV, Spitz MR, Eisen T, Amos CI, Houlston RS (2008) Common 5p15.33 and 6p21.33 variants influence lung cancer risk. Nat Genet 40:1407–1409
199. Wang Z, Seow WJ, Shiraishi K, Hsiung CA, Matsuo K, Liu J, Chen K, Yamji T, Yang Y, Chang IS, Wu C, Hong YC, Burdett L, Wyatt K, Chung CC, Li SA, Yeager M, Hutchinson A, Hu W, Caporaso N, Landi MT, Chatterjee N, Song M, Fraumeni JF Jr, Kohno T, Yokota J, Kunitoh H, Ashikawa K, Momozawa Y, Daigo Y, Mitsudomi T, Yatabe Y, Hida T, Hu Z, Dai J, Ma H, Jin G, Song B, Wang Z, Cheng S, Yin Z, Li X, Ren Y, Guan P, Chang J, Tan W, Chen CJ, Chang GC, Tsai YH, Su WC, Chen KY, Huang MS, Chen YM, Zheng H, Li H, Cui P, Guo H, XU P, Liu L, Iwasaki M, Shimazu T, Tsugane S, Zhu J, Jiang G, Fei K, Park JY, Kim YH, Sung JS, Park KH, Kim YT, Jung YJ, Kang CH, Park IK, Kim HN, Jeon HS, Choi JE, Choi YY, Kim JH, Oh IJ, Kim YC, Sung SW, Kim JS, Yoon HI, Kweon SS, Shin MH, Seow A, Chen Y, Lim WY, Liu J, Wong MP, Lee VH, Bassig BA, Tucker M, Berndt SI, Chow WH, JI BT, Wang J, Xu J, Sihoe AD, Ho JC et al (2016) Meta-analysis of genome-wide association studies identifies multiple lung cancer susceptibility loci in never-smoking Asian women. Hum Mol Genet 25:620–629
200. Welter D, Macarthur J, Morales J, Burdett T, Hall P, Junkins H, Klemm A, Flicek P, Manolio T, Hindorff L, Parkinson H (2014) The NHGRI GWAS catalog, a curated resource of SNP-trait associations. Nucleic Acids Res 42:D1001–D1006
201. Weren RD, Ligtenberg MJ, Kets CM, De Voer RM, Verwiel ET, Spruijt L, Van Zelst-Stams WA, Jongmans MC, Gilissen C, Hehir-Kwa JY, Hoischen A, Shendure J, Boyle EA, Kamping EJ, Nagtegaal ID, Tops BB, Nagengast FM, Geurts Van Kessel A, Van Krieken JH, Kuiper RP, Hoogerbrugge N (2015) A germline homozygous mutation in the base-excision repair gene NTHL1 causes adenomatous polyposis and colorectal cancer. Nat Genet 47:668–671
202. Whiffin N, Hosking FJ, Farrington SM, Palles C, Dobbins SE, Zgaga L, Lloyd A, Kinnersley B, Gorman M, Tenesa A, Broderick P, Wang Y, Barclay E, Hayward C, Martin L, Buchanan DD, Win AK, Hopper J, Jenkins M, Lindor NM, Newcomb PA, Gallinger S, Conti D, Schumacher F, Casey G, Liu T, SWEDISH LOW-RISK COLORECTAL CANCER STUDY,

G, Campbell H, Lindblom A, Houlston RS, Tomlinson IP, Dunlop MG (2014) Identification of susceptibility loci for colorectal cancer in a genome-wide meta-analysis. Hum Mol Genet 23:4729–4737

203. Whittemore AS, Wu AH, Kolonel LN, John EM, Gallagher RP, Howe GR, West DW, Teh CZ, Stamey T (1995) Family history and prostate cancer risk in black, white, and Asian men in the United States and Canada. Am J Epidemiol 141:732–740
204. Witt H, Luck W, Hennies HC, Classen M, Kage A, Lass U, Landt O, Becker M (2000) Mutations in the gene encoding the serine protease inhibitor, Kazal type 1 are associated with chronic pancreatitis. Nat Genet 25:213–216
205. Wolpin BM, Rizzato C, Kraft P, Kooperberg C, Petersen GM, Wang Z, Arslan AA, Beane-Freeman L, Bracci PM, Buring J, Canzian F, Duell EJ, Gallinger S, Giles GG, Goodman GE, Goodman PJ, Jacobs EJ, Kamineni A, Klein AP, Kolonel LN, Kulke MH, Li D, Malats N, Olson SH, Risch HA, Sesso HD, Visvanathan K, White E, Zheng W, Abnet CC, Albanes D, Andreotti G, Austin MA, Barfield R, Basso D, Berndt SI, Boutron-Ruault MC, Brotzman M, Buchler MW, Bueno-De-Mesquita HB, Bugert P, Burdette L, Campa D, Caporaso NE, Capurso G, Chung C, Cotterchio M, Costello E, Elena J, Funel N, Gaziano JM, Giese NA, Giovannucci EL, Goggins M, Gorman MJ, Gross M, Haiman CA, Hassan M, Helzlsouer KJ, Henderson BE, Holly EA, Hu N, Hunter DJ, Innocenti F, Jenab M, Kaaks R, Key TJ, Khaw KT, Klein EA, Kogevinas M, Krogh V, Kupcinskas J, Kurtz RC, Lacroix A, Landi MT, Landi S, Le Marchand L, Mambrini A, Mannisto S, Milne RL, Nakamura Y, Oberg AL, Owzar K, Patel AV, Peeters PH, Peters U, Pezzilli R, Piepoli A, Porta M, Real FX, Riboli E, Rothman N, Scarpa A, Shu XO, Silverman DT, Soucek P, Sund M, Talar-Wojnarowska R, Taylor PR, Theodoropoulos GE et al (2014) Genome-wide association study identifies multiple susceptibility loci for pancreatic cancer. Nat Genet 46:994–1000
206. Wu C, Hu Z, He Z, Jia W, Wang F, Zhou Y, Liu Z, Zhan Q, Liu Y, Yu D, Zhai K, Chang J, Qiao Y, Jin G, Liu Z, Shen Y, Guo C, Fu J, Miao X, Tan W, Shen H, Ke Y, Zeng Y, Wu T, Lin D (2011a) Genome-wide association study identifies three new susceptibility loci for esophageal squamous-cell carcinoma in Chinese populations. Nat Genet 43:679–684
207. Wu C, Kraft P, Zhai K, Chang J, Wang Z, Li Y, Hu Z, He Z, Jia W, Abnet CC, Liang L, Hu N, Miao X, Zhou Y, Liu Z, Zhan Q, Liu Y, Qiao Y, Zhou Y, Jin G, Guo C, Lu C, Yang H, Fu J, Yu D, Freedman ND, Ding T, Tan W, Goldstein AM, Wu T, Shen H, Ke Y, Zeng Y, Chanock SJ, Taylor PR, Lin D (2012) Genome-wide association analyses of esophageal squamous cell carcinoma in Chinese identify multiple susceptibility loci and gene-environment interactions. Nat Genet 44:1090–1097
208. Wu C, Miao X, Huang L, Che X, Jiang G, Yu D, Yang X, Cao G, Hu Z, Zhou Y, Zuo C, Wang C, Zhang X, Zhou Y, Yu X, Dai W, Li Z, Shen H, Liu L, Chen Y, Zhang S, Wang X, Zhai K, Chang J, Liu Y, Sun M, Cao W, Gao J, Ma Y, Zheng X, Cheung ST, Jia Y, Xu J, Tan W, Zhao P, Wu T, Wang C, Lin D (2011b) Genome-wide association study identifies five loci associated with susceptibility to pancreatic cancer in Chinese populations. Nat Genet 44:62–66
209. Wu C, Wang Z, Song X, Feng XS, Abnet CC, He J, Hu N, Zuo XB, Tan W, Zhan Q, Hu Z, He Z, Jia W, Zhou Y, Yu K, Shu XO, Yuan JM, Zheng W, Zhao XK, Gao SG, Yuan ZQ, Zhou FY, Fan ZM, Cui JL, Lin HL, Han XN, Li B, Chen X, Dawsey SM, Liao L, Lee MP, Ding T, Qiao YL, Liu Z, Liu Y, Yu D, Chang J, Wei L, Gao YT, Koh WP, Xiang YB, Tang ZZ, Fan JH, Han JJ, Zhou SL, Zhang P, Zhang DY, Yuan Y, Huang Y, Liu C, Zhai K, Qiao Y, Jin G, Guo C, Fu J, Miao X, Lu C, Yang H, Wang C, Wheeler WA, Gail M, Yeager M, Yuenger J, Guo ET, Li AL, Zhang W, Li XM, Sun LD, Ma BG, Li Y, Tang S, Peng XQ, Liu J, Hutchinson A, Jacobs K, Giffen C, Burdette L, Fraumeni JF Jr, Shen H, Ke Y, Zeng Y, Wu T, Kraft P, Chung CC, Tucker MA, Hou ZC, Liu YL, Hu YL, Liu Y, Wang L, Yuan G, Chen LS, Liu X, Ma T, Meng H, Sun L, Li XM, Li XM, Ku JW, Zhou YF et al (2014) Joint analysis of three genome-wide association studies of esophageal squamous cell carcinoma in Chinese populations. Nat Genet 46:1001–1006

210. Wu X, Ye Y, Kiemeney LA, Sulem P, Rafnar T, Matullo G, Seminara D, Yoshida T, Saeki N, Andrew AS, Dinney CP, Czerniak B, Zhang ZF, Kiltie AE, bishop DT, Vineis P, Porru S, Buntinx F, Kellen E, Zeegers MP, Kumar R, Rudnai P, Gurzau E, Koppova K, Mayordomo JI, Sanchez M, Saez B, Lindblom A, De Verdier P, Steineck G, Mills GB, Schned A, Guarrera S, Polidoro S, Chang SC, Lin J, Chang DW, Hale KS, Majewski T, Grossman HB, Thorlacius S, Thorsteinsdottir U, Aben KK, Witjes JA, Stefansson K, Amos CI, Karagas MR, Gu J (2009) Genetic variation in the prostate stem cell antigen gene PSCA confers susceptibility to urinary bladder cancer. Nat Genet 41:991–995
211. Xu J, Mo Z, Ye D, Wang M, Liu F, Jin G, Xu C, Wang X, Shao Q, Chen Z, Tao Z, Qi J, Zhou F, Wang Z, Fu Y, He D, Wei Q, Guo J, Wu D, Gao X, Yuan J, Wang G, Xu Y, Wang G, Yao H, Dong P, Jiao Y, Shen M, Yang J, Ou-Yang J, Jiang H, Zhu Y, Ren S, Zhang Z, Yin C, Gao X, Dai B, Hu Z, Yang Y, Wu Q, Chen H, Peng P, Zheng Y, Zheng X, Xiang Y, Long J, Gong J, Na R, Lin X, Yu H, Wang Z, Tao S, feng J, Sun J, Liu W, Hsing A, Rao J, Ding Q, Wiklund F, Gronberg H, Shu XO, Zheng W, Shen H, Jin L, Shi R, Lu D, Zhang X, Sun J, Zheng s L, Sun Y (2012) Genome-wide association study in Chinese men identifies two new prostate cancer risk loci at 9q31.2 and 19q13.4. Nat Genet 44:1231–1235
212. Yang L, Parkin DM, Li L, Chen Y (2003) Time trends in cancer mortality in China: 1987–1999. Int J Cancer 106:771–783
213. Yeager M, Orr N, Hayes RB, Jacobs KB, Kraft P, Wacholder S, Minichiello MJ, Fearnhead P, Yu K, Chatterjee N, wang Z, welch R, Staats BJ, Calle EE, Feigelson HS, Thun MJ, Rodriguez C, Albanes D, Virtamo J, Weinstein S, Schumacher FR, Giovannucci E, Willett WC, Cancel-Tassin G, Cussenot O, Valeri A, Andriole GL, Gelmann EP, Tucker M, Gerhard DS, Fraumeni JF Jr, Hoover R, Hunter DJ, Chanock SJ, Thomas G (2007) Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. Nat Genet 39:645–649
214. Yeh IT, Lenci RE, Qin Y, Buddavarapu K, Ligon AH, Leteurtre E, Do Cao C, Cardot-Bauters C, Pigny P, Dahia PL (2008) A germline mutation of the KIF1B beta gene on 1p36 in a family with neural and nonneural tumors. Hum Genet 124:279–285
215. Zanke BW, Greenwood CM, Rangrej J, Kustra R, Tenesa A, Farrington SM, Prendergast J, Olschwang S, Chiang T, Crowdy E, Ferretti V, Laflamme P, Sundararajan S, Roumy S, Olivier JF, Robidoux F, Sladek R, Montpetit A, Campbell P, Bezieau S, O'shea AM, Zogopoulos G, cotterchio M, Newcomb P, Mclaughlin J, Younghusband B, Green R, Green J, Porteous ME, Campbell H, Blanche H, Sahbatou M, Tubacher E, Bonaiti-Pellie C, Buecher B, Riboli E, Kury S, Chanock SJ, Potter J, Thomas G, Gallinger S, Hudson TJ, Dunlop MG (2007) Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24. Nat Genet 39:989–994
216. Zeron-Medina J, Wang X, Repapi E, campbell MR, Su D, Castro-Giner F, Davies B, Peterse EF, Sacilotto N, Walker GJ, Terzian T, Tomlinson IP, Box NF, Meinshausen N, De Val S, Bell DA, Bond GL (2013) A polymorphic p53 response element in KIT ligand influences cancer risk and has undergone natural selection. Cell 155:410–422
217. Zhang B, Beeghly-Fadiel A, Long J, Zheng W (2011) Genetic variants associated with breast-cancer risk: comprehensive research synopsis, meta-analysis, and epidemiological evidence. Lancet Oncol 12:477–488
218. Zhang B, Jia WH, Matsuda K, Kweon SS, Matsuo K, Xiang YB, Shin A, Jee SH, Kim DH, Cai Q, Long J, Shi J, Wen W, yang G, Zhang Y, Li C, Li B, Guo Y, Ren Z, Ji BT, Pan ZZ, Takahashi A, Shin MH, Matsuda F, Gao YT, Oh JH, Kim S, AHN YO, Genetics, Epidemiology of Colorectal Cancer, C, Chan AT, Chang-Claude J, Slattery ML, Colorectal Transdisciplinary, S, Gruber SB, Schumacher FR, Stenzel SL, Colon Cancer Family R, Casey G, Kim HR, Jeong JY, Park JW, Li HL, Hosono S, Cho SH, Kubo M, Shu XO, Zeng YX, Zheng W (2014) Large-scale genetic study in East Asians identifies six new loci associated with colorectal cancer risk. Nat Genet 46:533–542

219. Zhang H, Zhai Y, Hu Z, Wu C, Qian J, Jia W, Ma F, Huang W, Yu L, Yue W, Wang Z, Li P, Zhang Y, Liang R, Wei Z, Cui Y, Xie W, Cai M, Yu X, Yuan Y, Xia X, Zhang X, Yang H, Qiu W, Yang J, Gong F, Chen M, Shen H, Lin D, Zeng YX, He F, Zhou G (2010) Genome-wide association study identifies 1p36.22 as a new susceptibility locus for hepatocellular carcinoma in chronic hepatitis B virus carriers. Nat Genet 42:755–758
220. Zheng W, Mclerran DF, Rolland BA, Fu Z, Boffetta P, He J, Gupta PC, Ramadas K, Tsugane S, Irie F, Tamakoshi A, Gao YT, Koh WP, Shu XO, Ozasa K, Nishino Y, Tsuji I, Tanaka H, Chen CJ, Yuan JM, Ahn YO, Yoo KY, Ahsan H, Pan WH, Qiao YL, Gu D, Pednekar MS, Sauvaget C, Sawada N, Sairenchi T, Yang G, Wang R, Xiang YB, Ohishi W, Kakizaki M, Watanabe T, Oze I, You SL, Sugawara Y, Butler LM, Kim DH, Park SK, Parvez F, Chuang SY, Fan JH, Shen CY, Chen Y, Grant EJ, Lee JE, Sinha R, Matsuo K, Thornquist M, Inoue M, Feng Z, Kang D, Potter JD (2014) Burden of total and cause-specific mortality related to tobacco smoking among adults aged >/= 45 years in Asia: a pooled analysis of 21 cohorts. PLoS Med 11:e1001631
221. Zuk O, Schaffner SF, Samocha K, Do R, Hechter E, Kathiresan S, Daly MJ, Neale BM, Sunyaev SR, Lander ES (2014) Searching for missing heritability: designing rare variant association studies. Proc Natl Acad Sci U S A 111:E455–E464

Chapter 6
Genetics of Infectious Diseases

Yosuke Omae and Katsushi Tokunaga

Abstract Genome-wide association studies (GWASs) have been performed in the field of human genetics to identify disease- or phenotype-related genetic variants. Infectious diseases are caused by bacteria, viruses, parasites, or fungi and these pathogens are considered as one of the environmental factors of disease onset. The first GWAS in infectious disease was reported in 2007 for acquired immunodeficiency syndrome (AIDS). More than 80 GWASs have since been reported in various infectious diseases and successfully revealed genetic risk factors. In this chapter, we will review GWAS reports published between 2007 and 2016 for three major global infectious diseases (AIDS, malaria, and tuberculosis), hepatitis (B and C), and other infectious diseases. We will also discuss the currently proposed mechanisms based on GWAS findings.

Keywords GWAS · Infectious disease · AIDS · HIV · Malaria · *Plasmodium* · Tuberculosis · *Mycobacterium* · Hepatitis · HBV · HCV · Leprosy · Meningococcus · *Helicobacter pylori* · Pneumococcus · *Salmonella* · *Staphylococcus* · Bacteremia · Sepsis · Dengue · Herpes zoster · HPV · Influenza · Leishmaniasis · HLA

6.1 Introduction

Infectious diseases are caused by pathogenic microorganisms, such as bacteria, viruses, parasites, or fungi, and can spread from one person to another. Infectious diseases are leading causes of human mortality and morbidity. To identify host genetic risk factors for infectious diseases, candidate gene approaches and family-based approaches have been applied and the contribution of several human genetic factors has been suggested [1]. Human leukocyte antigen (HLA) is one of the major

Y. Omae · K. Tokunaga (✉)
Genome Medical Science Project (Toyama), National Center for Global Health and Medicine (NCGM), Tokyo, Japan
e-mail: tokunaga@m.u-tokyo.ac.jp

T. Tsunoda et al. (eds.), *Genome-Wide Association Studies*,
https://doi.org/10.1007/978-981-13-8177-5_6

genetic factors reported from candidate gene studies in various infectious diseases [1]. HLA is a gene complex encoding the major histocompatibility complex (MHC) in humans and an important regulator of the immune response. *HLA* genes show great polymorphisms which allow them to fine-tune the adaptive immune system against foreign pathogens. In brief, antigen-presenting cells (APCs) engulf a pathogen through phagocytosis and load small peptides digested from pathogen proteins onto HLA antigens (referred to as peptide presentation) [2]. HLA class I molecules (e.g., HLA-A, -B, and -C) present peptides to CD8-positive or cytotoxic T cells and class II molecules (e.g., HLA-DP, -DQ, -DR) present peptides to CD4-positive or helper T cells so that pathogens can be destroyed by the immune system [2]. *HLA* gene polymorphisms can alter peptide presentation and pathogen clearance by the immune system, resulting in various susceptibilities to pathogen infection.

Several non-HLA genes have also been reported in previous candidate gene studies and family-based studies. However, a limitation of candidate gene approaches is that *a priori* knowledge is necessary and researchers may overlook the major determinant for the diseases, which was unexpected at the time of analysis. Conversely, a limitation of family-based approaches is that identified genetic variants in familial disease onset cases can be rare among sporadic cases. Genome-wide association studies (GWASs) can overcome these limitations by using a hypothesis-free and comprehensive approach in analyses. In GWASs, we use single nucleotide polymorphisms (SNPs) as markers for disease causative genetic variants, genotype multiple SNPs simultaneously by using a DNA-microarray, and compare the differences in genotype frequencies between a case group and a control group or assess the correlation between genotype frequencies and clinical parameters.

In this chapter, we will review genetic factors detected to be associated with infectious disease by GWASs. We focused on associations that passed statistically significant thresholds after considering multiplicity or the standard genome-wide significance p-value threshold ($\alpha = 5.00E\text{-}08$). We hope this review and the gene list will be of some help to the readers for better understanding of genetics in infectious diseases.

6.2 Three Major Global Infectious Diseases (Table 6.1)

Acquired Immune Deficiency Syndrome (AIDS)

AIDS is caused by human immunodeficiency virus (HIV) infection. AIDS interferes with the host immune system and increases the risk of opportunistic infections after a prolonged period with no symptoms. Great variability in the susceptibility to HIV-1 infection and in the subsequent disease course is known, and the first GWAS for HIV-1 was reported in 2007 as the first GWAS in infectious diseases [3]. The Goldstein group conducted an association analysis between plasma circulating virus levels (viral load) and genotype frequency of SNPs in 486 HIV-1 infected European

Table 6.1 List of significantly associated genes and polymorphisms in GWAS of three major global infectious diseases

Gene	Polymorphism	Study design	Cases/Controls (Population)	P	Effect size (Odds ratio, if not specified)	Ref.
AIDS (Viral load)						
HLA-B/HCP5	rs2395029, B∗5701	Viral load	Case 486 (European)	9.36E-12	9.6% of total variation	[3]
HLA-C	rs9264942	Viral load	Case 486 (European)	3.77E-09	6.5% of total variation	[3]
CCR5	rs333 (delta 32)	Viral load	Case 2362 (European)	1.70E-10	1.7% of total variation	[10]
HLA-B	B∗57:03	Viral load	Case 515 (African-American)	5.60E-10	~10% of total variation	[5]
HLA-B, HLA-A	Peptide binding groove	Viral load	Case 6315 (European)	2.00E-83	12.3% of total variation	[7]
CCR5/CCRL2	rs1015164, delta 32, Hap-P1	Viral load	Case 6315 (European)	1.50E-19	2.2% of total variation	[7]
AIDS (Disease progression)						
ZNRD1/ RNF39	rs9261174	Disease progression	Case 1071 (European)	1.80E-08	5.8% of total variation[a]	[10]
PARD3B	rs11884476	Disease progression	Case 755 (European/ American)	3.37E-09	Hazard ratio=0.30	[17]
HLA-B/HCP5	rs2395029	AIDS-nonprogressors vs HIV-uninfected	275/1352 (European)	6.79E-10	3.47 (2.39–5.04)[b]	[11]
CXCR6	rs2234358	AIDS-nonprogressors vs HIV-uninfected	276/697 (European/ American)	2.10E-08	1.77 (1.44–2.18)[b]	[16]
HLA-B	Position 97 (67, 70)	AIDS-controllers vs AIDS-progressors	974/2648 (European)	4.00E-45	No data	[12]

(continued)

Table 6.1 (continued)

Gene	Polymorphism	Study design	Cases/Controls (Population)	P	Effect size (Odds ratio, if not specified)	Ref.
HLA-C	rs9264942	AIDS-controllers vs AIDS-progressors	974/2648 (European)	2.80E-35	2.9 (No 95% CI data)[b]	[12]
MICA	rs4418214	AIDS-controllers vs AIDS-progressors	974/2648 (European)	1.40E-34	4.4 (No 95% CI data)[b]	[12]
PSORS1C3	rs3131018	AIDS-controllers vs AIDS-progressors	974/2648 (European)	4.20E-16	2.1 (No 95% CI data)[b]	[12]
AIDS (HIV acquisition)						
CYP7B1	rs6996198	HIV-infected cases vs HIV-uninfected controls	1739/1397 (European)	7.76E-08	0.70 (No 95% CI data)[c]	[26]
CCR5	delta32	HIV-infected cases vs HIV-uninfected controls	6334/7247 (European)	5.00E-09	0.2 (No 95% CI data)[d]	[27]
Malaria (Severe)						
HBB	rs11036238	Cases vs Controls (in children)	2045/3078 (Gambian)	3.70E-11	0.63 (0.55–0.72)	[34]
ABO	rs8176719	Cases vs Controls	2645/3050 (Ghanaian)	1.10E-20	1.67 (1.50–1.86)	[35]
ATP2B4	rs2365860	Cases vs Controls	2645/3050 (Ghanaian)	1.50E-08	0.63 (0.55–0.74)	[35]
MARVELD3	rs2334880	Cases vs Controls	2645/3050 (Ghanaian)	3.90E-08	1.24 (1.15–1.34)	[35]
FREM3/GYPE	rs184895969	Cases vs Controls	~10,000/15,000 (African)	9.50E-11	0.67 (0.60–0.76)	[36]
Tuberculosis						
18q11.2	rs4331426	Cases vs Controls	3632/7501 (Ghanaian/Gambian)	6.80E-09	1.19 (1.13–1.27)	[42]
11p13	rs2057178	Cases vs Controls	2127/5636 (Ghanaian)	2.63E-09	0.77 (0.71–0.84)	[43]
ASAP1	rs10956514	Cases vs Controls	6396/8038 (Russian)	1.00E-10	0.85 (0.81–0.89)	[44]
HLA-DQA1/-DRB1	rs557011, rs9271378	Cases vs Controls	9654/29,4043 (European)	2.00E-15	1.18 (1.13–1.23)	[45]

(continued)

Table 6.1 (continued)

Gene	Polymorphism	Study design	Cases/Controls (Population)	P	Effect size (Odds ratio, if not specified)	Ref.
MAFB	rs6071980	Young cases vs Controls	393/1255 (Thai/ Japanese)	6.69E-08	1.73 (1.42–2.11)	[49]
IL12B	rs4921437	HIV-positive TB cases vs HIV-positive controls	267/314 (Ugandan/ Tanzanian)	2.11E-08	0.37 (0.27–0.53)	[50]

CI confidence interval
[a]Data in reference [3]
[b]Protective for disease progression
[c]Data before meta-analysis
[d]Data under a recessive model

individuals. They replicated the association of *HLA-B* identified in previous candidate gene approaches and revealed a novel association of *HLA-C* with HIV-1 viral load [3]. Association of these *HLA* class I genes with HIV-1 viral load was confirmed by subsequent GWASs from independent studies in European [4], African American [5], and Chinese [6] populations. The most significant SNP in the European GWAS (rs2395029) is located near the HLA complex 5 (*HCP5*) gene but is in almost complete linkage disequilibrium (LD) with the *HLA-B*57:01* allele in the European population [3]. Conversely, a GWAS in African Americans revealed a strong association between *HLA-B*57:03* and HIV-1 viral load [5], suggesting that *HLA-B*57* group alleles affect viral load variations. An analysis of amino acid residues within the *HLA* loci further indicated that the major genetic associations observed between the *HLA* locus and HIV control were due to polymorphisms in amino acids located at the *HLA-B* peptide binding groove [7]. This finding suggests that presentation of specific viral epitopes is dependent on the structure of the HLA peptide binding groove and may alter the efficiency of the cytotoxic T cell response. Another significantly associated SNP (rs9264942) is located 35-kb upstream of the *HLA-C* gene and the protective allele of rs9264942 was strongly associated with higher expression of the *HLA-C* gene [3]. The Carrington group revealed that higher expression of *HLA-C* correlated with stronger cytotoxic T cell responses and contributed to viral control [8]. Later, the group attributed the molecular mechanism for variability in *HLA-C* expression to a polymorphism in the 3′ untranslated region of *HLA-C*, which is in strong LD with rs9264942 and regulates the binding of a microRNA, hsa-miR-148, to its target site [9]. Decreased binding of this microRNA can increase the expression of *HLA-C* and promote cytotoxic T cell responses. Furthermore, an expanded GWAS for viral load in 2362 European HIV-1 infected cases identified a significant association of C-C chemokine receptor type 5 (*CCR5*) with viral load, which was reported in previous candidate gene studies [10]. The Goldstein group estimated that common variants located in *HLA* class I molecules and the *CCR5* region explained the majority (25%) of the host genetic contribution to the variation in HIV-1 viral load using GWAS data from 6315 European individuals [7].

A GWAS to identify genetic variants associated with AIDS progression after HIV-1 infection was also first conducted by the Goldstein group using the drop in CD4-positive T cell counts as the indicator of AIDS progression [3]. Their later expanded GWAS identified a significant association of zinc ribbon domain-containing 1/ring finger protein 39 (*ZNRD1/RNF39*) with AIDS progression [10]. Furthermore, a GWAS comparing HIV-infected non-progressors to AIDS and HIV-uninfected controls identified a significant association of *HLA-B* polymorphism with AIDS progression, which emphasized the role of *HLA-B* in control of disease progression soon after infection [11]. A GWAS in a multiethnic cohort of HIV-1 controllers and AIDS progressors revealed associations of specific amino acids in the *HLA-B* peptide binding groove with progression phenotype and independent associations for *HLA-C* and major histocompatibility class I polypeptide-related sequence A (*MICA*) gene variants [12]. The associations of *HLA-B*, *HLA-C*, and *MICA* with AIDS progression have been confirmed by subsequent GWASs in independent European cohorts [13, 14]. MICA functions as a ligand for natural killer group 2 member D (NKG2D), which is present on CD8-positive T cells and natural killer (NK) cells. The engagement of MICA with NKG2D activates cytolytic responses against infected cells and tumor cells [15]. Different expression levels of soluble MICA due to genetic polymorphism can alter the infection status of HIV-1. Other loci including psoriasis susceptibility 1 candidate 3 (*PSORS1C3*) [12], C-X-C motif chemokine receptor 6 (*CXCR6*) [16], and par-3 family cell polarity regulator beta (*PARD3B*) [17] showed significant associations with AIDS progression, although their associations and other suggestive associations [18–20] required further confirmation in independent cohort studies.

Compared to the GWAS for viral load and disease progression, GWASs for HIV-1 acquisition through comparison of HIV-infected individuals and HIV-uninfected individuals have experienced difficulty in identifying significant associations of genetic variants with HIV [21–25]. Thus far, only two multi-cohort GWASs have revealed significant associations of genetic variants with HIV-1 acquisition [26, 27]. The largest GWAS comprised 6300 cases and 7200 controls and identified an association between 32-base-pair deletion in *CCR5* gene and HIV-1 acquisition [27]. The necessity of large sample numbers to identify significant associations suggests that genetic influence on HIV acquisition is smaller than that on viral load or disease progression.

GWASs for other aspects of HIV-1 infection have focused on the development of cross-reactive neutralizing antibodies [28], *in vitro* replication of HIV-1 in monocyte-derived macrophages [29], and death as disease course of AIDS [30]. However, no significant associations have been identified from these studies, possibly due to their limited sample sizes. Of note, the Fellay group considered the interaction between host and pathogen in infectious diseases and proposed a unique genome-to-genome analysis of associations between human genetic variation, HIV-1 sequence diversity, and viral control [31]. Their method identified associations between SNPs within the *HLA* region and 48 amino acid variants in HIV and will be worth noting

for future identification of AIDS genetic risk factors through consideration of the viral genome.

Malaria

Malaria is caused by pathogenic parasites in the genus *Plasmodium*, which are transmitted through the bites of infected mosquitoes. Infection by malaria parasite may result in a wide variety of symptoms, ranging from absent or very mild symptoms to severe disease, including malarial anemia with hemolysis and cerebral malaria with neurological symptoms. Five species of *Plasmodium* can infect humans and *Plasmodium falciparum* is the major cause of severe malaria. Malaria affects mortality and morbidity in endemic areas of sub-Saharan Africa. Molecular genetic studies before the first GWAS suggested that hemoglobinopathies including sickle cell trait and glucose-6-phosphate dehydrogenase (G6PD) deficiency confer a survival advantage against severe malaria and have subsequently increased in frequency through natural selection over generations [32]. Genetic variance component analysis considering the correlation between the disease incidence and the degree of genetic relationship estimated that host genetic factors accounted for nearly 25% of the risk of severe malaria [33]. The first GWAS in malaria reported in 2009 conducted a two-stage GWAS including 958 Gambian children with severe malaria and 1382 controls, followed by a replication study in an independent sampling of 1087 cases and 2376 controls. The GWAS identified the *HBB* region, which encodes beta globin of hemoglobin [34]. It is already well known that the sickle cell allele, hemoglobin S, confers resistance to *P. falciparum* [32]. Subsequent GWASs in 2012 and 2015 further identified ATPase plasma membrane Ca2+ transporting 4 (*ATP2B4*) [35], MARVEL domain containing 3 (*MARVELD3*) [35], and FRAS1 related extracellular matrix 3/glycophorin E (*FREM3/GYPE*) genes [36] as novel malaria resistance genes in addition to the *ABO* gene, which encodes for histo-blood group ABO system transferase and has already been identified as a malaria risk gene from candidate gene studies [37]. Associations between *HBB*, *ABO*, and *ATP2B4* loci and severe malaria were confirmed in a multicenter study comprising 11,890 malaria cases and 17,441 controls from 12 locations in Africa, Asia, and Oceania [38]. ATP2B4 is a major calcium pump in the plasma membrane of erythrocytes [39], suggesting that an alteration of its structure or expression can disturb homeostasis of intra-erythrocytic calcium concentrations and affect the development and structure of the parasite at intra-erythrocytic stages. Interestingly, an association between G6PD deficiency and malaria has not been detected in reported GWASs, however, the multicenter study revealed opposing effects on cerebral malaria and severe malarial anemia, which was consistent across different populations [38]. As G6PD deficiency has been observed to suppress *P. vivax* infection more effectively than *P. falciparum* infection [40], parasitic genetic variation is proposed as a source of observed differences in clinical outcome. Further investigation for causative parasitic variation in malaria is warranted.

Tuberculosis (TB)

TB is caused by the pathogenic bacterium, *Mycobacterium tuberculosis* (MTB), which is transmitted through the air from person to person. MTB most commonly grows in the lungs and symptoms during the active infection phase include a bad cough, chest pain, and coughing up blood or sputum, which can be fatal if not treated. TB is one of the major causes of infectious disease-related mortality worldwide. Although one-third of the world population is infected by MTB, only 5–15% of infected people develop active TB; the remaining 90% of infected people remain in a dormant stage throughout their life [41], suggesting the contribution of host genetic factors to TB onset. The first GWAS for TB was reported in 2010. The Hill group reported the association between the chromosome 18q11.2 locus and TB onset among 3632 cases and 7501 controls in an African population [42]. The Hill group further expanded the candidate SNPs through the application of the whole genome SNP imputation method and reported another significant association in the chromosome 11p13 locus [43]. A GWAS in Russian with pulmonary TB reported an association of ArfGAP with SH3 domain, Ankyrin repeat and PH domain 1 (*ASAP1*) gene with TB onset, and decreased macrophage migration was observed in individuals homozygous for the risk allele [44]. The association of the *HLA* class II locus, which was reported in previous candidate gene studies, with TB onset was also identified from a GWAS in European populations [45]. However, the effect sizes of these identified genes were weak (odds ratio <1.5) and reproducibility of these identified genetic factors was controversial among independent GWASs conducted in Indonesia, South Africa, Morocco, Thailand, and Japan [46–49]. Therefore, identification of common genetic risk factors remains challenging for TB.

The Mahasirimongkol group and our group have focused on differences in TB onset. Primary TB patients show symptoms within 5 years after infection, whereas recurrent TB patients generally progress >5 years after infection. We proposed that this difference in TB onset can be a determinant of genetic risk factors, and identified a genetic risk factor specific for young age onset (assuming primary TB) and shared among Thai and Japanese populations [49]. The identified SNP is located near MAF BZIP transcription factor B (*MAFB*), which functions as a transcription factor that determines the fate of monocyte/macrophage differentiation. Another unique approach focusing on HIV-positive individuals who are at high risk for disease progression has been proposed. A GWAS comparing 267 HIV-infected TB cases and 314 HIV-infected non-TB controls successfully identified a common variant near the interleukin-12B (*IL12B*) gene, which is involved in cell-mediated immunity against intracellular bacteria [50].

These reports suggest the importance of considering the clinical heterogeneity of TB to identify shared genetic risk factors for TB. MTB is also known to demonstrate differences between its genome structure and *in vitro* phenotype [51]. We recently conducted a GWAS based on lineage information of MTB and revealed that pathogen lineage can affect the risk of host polymorphisms [52]. The risk of *HLA* class II alleles also differed according to the specific lineage in MTB [53]. Further consid-

eration of pathogen heterogeneity may also help facilitate identification of shared risk genetic factors for TB onset.

6.3 Chronic Hepatitis Virus Infection (Table 6.2)

Hepatitis C

Hepatitis C is a liver disease caused by blood-borne infection of hepatitis C virus (HCV), and chronic HCV infection leads to liver cirrhosis or liver cancer. The genetic basis of HCV infection was identified through analysis of differences in anti-HCV treatment response [54–56]. It was well known that many patients will not be cured by treatment with pegylated interferon-α (PEG-IFN-α) and ribavirin combination therapy, which is a standard treatment for HCV patients [57]. GWASs comparing treatment responders and non-responders were reported from three independent research groups at approximately the same time in 2009. All three GWASs revealed an unexpectedly strong association of *IL28B* (also called IFN-λ3) with HCV treatment response in different populations [54–56]. *IL28B* was also found to be associated with the baseline (pre-treatment) viral load and contributes to the host viral clearance [58–60]. Lower *IL28B* expression levels were observed in individuals carrying risk alleles [55, 56]. IL-28B is a cytokine distantly related to type I IFNs and forms a gene cluster with *IL28A* and *IL29*, which comprise the type III IFN family. These type III IFNs were reported to be induced by viral infection and have antiviral activity [61, 62], however, a surprisingly strong effect during HCV clearance is observed. All three type III IFNs interact with a unique heterodimeric class II cytokine receptor consisting of IL-10Rβ, which is a receptor shared with other cytokine receptors, and IL-28Rα, which is a receptor specific to these IFNs [61, 62]. Thus, they may serve as an alternative to type I IFNs, which are well-known regulators of antiviral response, in providing immunity to viral infection through a common downstream signaling system. Moreover, IFN-λ signaling has been proposed as a potential target for novel antiviral drug development. The association of *HLA* class II alleles with viral clearance and chronic HCV infection has been reported in addition to *IL28B* [60, 63]. In addition, the first GWASs focused on individuals infected by HCV genotype 1, the most common type in developed countries with the lowest treatment response among several HCV genotypes, and the association of *IL28B* was confirmed in HCV genotype 4 but not in HCV genotype 2 or 3 [59]. HCV viral genome analysis further revealed that variation in the HCV genome core region is associated with poor response to IFN therapy, indicating that both host and viral genetic factors contribute to the IFN response [64].

Besides the viral clearance phenotype, host genetic factors related to the progression of HCV-induced liver diseases have also been explored among chronic hepatitis C cases. Initial GWASs on HCV-induced hepatocellular carcinoma (HCC) were reported in Japanese populations. One GWAS compared HCV-induced HCC

Table 6.2 List of significantly associated genes and polymorphisms in GWAS of hepatitis B and C

Gene	Polymorphism	Study design	Cases/Controls (Population)	P	Effect size (Odds ratio)	Ref.
HCV (infection status)						
IL28B	rs12979860	Treatment-induced viral clearance	Case 1137 (diverse ethnic groups)[a]	1.37E-28	7.3 (5.1–10.4)[b]	[54]
IL28B	rs8099917	Treatment-induced viral clearance	128/186 (Japanese)	2.68E-32	27.1 (14.6–50.3)	[55]
IL28B	rs8099917	Treatment-induced viral clearance	261/294 (Australian/European)	9.25E-09	1.98 (1.57–2.52)	[56]
IL28B	rs8099917	Spontaneous viral clearance	347/1015 (European)	6.07E-09	2.31 (1.74–3.06)	[59]
IL28B	rs12979860	Spontaneous viral clearance	919/1482 (European, African, others)	2.17E-30	0.45 (No 95% CI data)[c]	[60]
HLA-DQB1	rs4273729	Spontaneous viral clearance	919/1482 (European, African, others)	1.71E-16	0.59 (No 95% CI data)	[60]
HLA-DQB1	rs9275572	Chronic infection cases vs healthy controls	6218/29,894 (Japanese)	3.59E-16	0.79 (0.75–0.84)	[63]
HCV (Disease progression)						
MICA	rs2596542	HCC cases vs controls	1394/5486 (Japanese)	4.21E-13	1.39 (1.27–1.52)	[65]
DEPDC5	rs1012068	HCC cases vs non-developers	922/2390 (Japanese)	1.27E-13	1.75 (1.51–2.03)	[66]
TLL1	rs17047200	HCC cases after eradication vs non-developers	253/543 (Japanese)	2.66E-08	2.37 (1.74–3.23)	[67]
RNF7	rs16851720	Liver fibrosis, progression	Case 1636 (European)	8.90E-09	0.23 (0.15–0.31)[d]	[68]
MERTK	rs4374383	Liver fibrosis, blood transfusion	Case 319 (European)	1.10E-09	0.19 (0.10–0.37)[e]	[68]
C6orf10	rs910049	Liver cirrhosis cases vs non-developers	1618/4854 (Japanese)	9.15E-11	1.46 (1.28–1.58)	[69]
BTNL2/HLA-DRA	rs3135363	Liver cirrhosis cases vs non-developers	1618/4854 (Japanese)	1.45E-10	1.37 (1.24–1.51)	[69]

(continued)

Table 6.2 (continued)

Gene	Polymorphism	Study design	Cases/Controls (Population)	P	Effect size (Odds ratio)	Ref.
HLA-DQA1	∗06:01 allele	Liver cirrhosis cases vs non-developers	682/1045 (Japanese)	4.53E-04	2.80 (1.38–3.32)	[69]
HLA-DQA1/-DRB1	rs9461776	Vasculitis cases vs non-developers	448/626 (European)	7.10E-09	2.02 (No 95% CI data)	[70]
HBV (Infection status)						
HLA-DPA1/-DPB1	rs9277535, rs3077	Chronic hepatitis B cases vs healthy controls	1890/3350 (Japanese/Thai)	6.34E-39	0.57 (0.52–0.62)	[72]
HLA-DQA1/-DQB1	rs7453920, rs2856718	Chronic hepatitis B cases vs healthy controls	2662/6486 (Japanese)	3.99E-37	1.56 (1.45–1.67)	[73]
HLA-DPA1/-DPB1	rs9277535, rs3077	HBV carriers vs HBV resolvers	651/434 (Japanese/Korean)	1.89E-12	0.43 (0.34–0.54)	[74]
HLA-DQA1/-DQB1	rs7453920	HBV carriers vs HBV resolvers	5181/6610 (Han Chinese)	4.93E-37	0.53 (0.48–0.59)	[75]
HLA-C	rs3130542	HBV carriers vs HBV resolvers	5181/6610 (Han Chinese)	9.49E-14	1.33 (1.23–1.44)	[75]
UBE2L3	rs4821116	HBV carriers vs HBV resolvers	5181/6610 (Han Chinese)	1.71E-12	0.82 (0.77–0.87)	[75]
EHMT2	rs652888	HBV carriers vs healthy controls	1371/2938 (Korean)	7.07E-13	1.38 (1.22–1.57)	[79]
TCF19	rs1419881	HBV carriers vs healthy controls	1371/2938 (Korean)	1.26E-18	0.73 (0.66–0.81)	[79]
HLA-DPA3	rs9366816	HBV carriers vs controls (in males)	1065/1623 (Han Taiwanese)	2.58E-10	1.43 (1.28–1.60)	[76]
CFB	rs12614	Chronic hepatitis B cases vs controls	9114/9257 (Chinese)	1.28E-34	1.89 (1.69–2.08)	[80]
HLA-DOA	rs38352	Chronic hepatitis B cases vs controls	9114/9257 (Chinese)	1.04E-23	1.26 (1.20–1.31)	[80]
HLA-C	rs2853953	Chronic hepatitis B cases vs controls	9114/9257 (Chinese)	5.06E-20	1.47 (1.35–1.59)	[80]

(continued)

Table 6.2 (continued)

Gene	Polymorphism	Study design	Cases/Controls (Population)	P	Effect size (Odds ratio)	Ref.
NOTCH4	rs422951	Chronic hepatitis B cases vs controls	9114/9257 (Chinese)	5.33E-16	1.27 (1.20–1.35)	[80]
CD40	rs1883832	Chronic hepatitis B cases vs controls	9114/9257 (Chinese)	2.95E-15	1.19 (1.14–1.25)	[80]
INTS10	rs7000921	HBV carriers vs HBV resolvers	5156/4413 (Chinese)	3.20E-12	0.78 (0.73–0.84)	[81]
HBV (Disease progression)						
KIF1B	rs17401966	HCC cases vs non-developers	2310/1789 (Chinese)	3.40E-19	0.62 (0.56–0.69)	[82]
HLA-DQA1/-DRB1	rs9272105	HCC cases vs non-developers	5969/6190 (Chinese)	5.24E-22	1.28 (1.22–1.35)	[85]
GRIK1	rs455804	HCC cases vs non-developers	5969/6190 (Chinese)	5.24E-10	0.84 (0.80–0.89)	[85]
HLA-DQB1/–DQA1	rs9275319	HCC cases vs non-developers	5480/6319 (Chinese)	8.65E-19	1.51 (1.38–1.66)	[86]
STAT4	rs7574865	HCC cases vs non-developers	5480/6319 (Chinese)	1.66E-11	1.22 (1.15–1.29)	[86]
FDX1	rs2724432	Liver cirrhosis+HCC cases vs non-developers	76/343 (Arab)	4.29E-08	3.01 (2.21–5.30)	[88]

CI confidence interval
[a]European-American, African-American and Hispanic
[b]Data for C allele, recessive model in European population
[c]Data for T allele
[d]Data under an additive model
[e]Data under a recessive model

patients and HCV-negative controls and identified an association with the *MICA* gene [65]. Another GWAS compared HCV-induced HCC patients and chronic HCV individuals without HCC and identified the DEP domain-containing 5 (*DEPDC5*) gene; its function has been poorly understood, however, its expression was increased in HCC cases [66]. A GWAS for HCC after eradication of HCV infection was conducted using 123 cases and 332 controls, followed by a replication study in a Japanese population. This GWAS identified a SNP in tolloid-like 1 (*TLL1*) and expression of *TLL1* was increased in animal models of liver injury and liver tissues

of patients with fibrosis compared with respective controls [67]. Several other host genetic factors were found to be significantly associated with the progression of other HCV-induced liver diseases, such as ring finger protein 7 (*RNF7*) and MER receptor tyrosine kinase (*MERTK*) with liver fibrosis status [68], and chromosome 6 open reading frame 10 (*C6orf10*), butyrophilin-like 2 (*BTNL2*)/*HLA-DRA* and *HLA-DQA1* with liver cirrhosis [69]. GWASs for cryoglobulin-related vasculitis, an autoimmune and B cell lymphoproliferative disorder, and lichen planus, a chronic inflammatory mucocutaneous disease, among patients with HCV infection have also been established [70, 71]. These reports suggest the association of host immune system including HLA with several aspects of disease progression after chronic HCV infection.

Hepatitis B

Hepatitis B is caused by hepatitis B virus (HBV) infection through contact with the blood or other body fluids of an infected person, and chronic HBV infection leads to a high risk of death from cirrhosis and liver diseases such as HCC. Clinical outcomes after exposure to HBV are known to be highly variable and approximately 15% of HBV-infected people become chronic carriers, 75% of who live in Southeast Asia and East Pacific areas. The first GWAS for HBV infection was conducted in 786 Japanese chronic hepatitis B cases and 2201 controls, followed by replication studies in three additional Japanese and Thai cohorts consisting of 1300 cases and 2100 controls. The first GWAS revealed an association of SNPs in the *HLA-DPA1/HLA-DPB1* locus with susceptibility to chronic HBV infection [72]. An expanded GWAS including 2662 Japanese chronic hepatitis B cases and 6486 healthy controls further revealed an association of *HLA-DQA1/HLA-DQB1* with susceptibility to chronic HBV infection [73]. Subsequent GWASs based on the viral clearance status comparing HBV carriers and HBV resolved individuals confirmed the association of *HLA-DP* and *HLA-DQ* with susceptibility to chronic HBV infection in Japanese [74], Korean [74], Han Chinese [75], and Han Taiwanese [76] populations. Furthermore, trans-ethnic association analysis of *HLA-DPA1/HLA-DPB1* alleles and haplotypes identified susceptibility and resistance alleles to chronic HBV infection in Asian populations including Japanese, Korean, Hong Kong, and Thai [77] populations. However, few studies have examined and replicated the association of *HLA-DPA1/HLA-DPB1* with chronic HBV infection in non-Asian populations. One possible explanation is that risk SNPs identified in Asian population have low minor allele frequency in non-Asian populations. HLA-DPA1 and HLA-DPB1 form a heterodimer consisting of an alpha and a beta chain of class II HLA molecules on the surface of APCs [78]. Polymorphisms in *HLA* may result in different binding affinities between HLA-DP subtypes and extracellular antigens and alter the pathogenesis of HBV infection. In addition to the *HLA* class II locus, associations of *HLA-C* and ubiquitin-conjugating enzyme E2 L3 (*UBE2L3*) with chronic HBV infection were identified in Han Chinese [75]. Thus far, GWASs in Korean and Chinese

populations revealed seven additional risk loci for chronic HBV infection, including euchromatic histone-lysine-methyltransferase 2 (*EHMT2*), transcription factor 19 (*TCF19*) [79], *HLA-DPA3* [76], complement factor B (*CFB*), *HLA-DOA*, neurogenic locus notch homolog (*NOTCH4*), *CD40* [80], and integrator complex subunit 10 (*INTS10*) [81], although additional studies are warranted to further confirm these findings.

Host genetic variants related to HBV-induced liver disease progression among chronic HBV cases have been examined in Chinese populations. The first GWAS was conducted using 355 HBV-induced HCC cases and 360 HBV carriers without HCC among Southern Chinese individuals, followed by replication studies in five additional independent Chinese cohorts. The GWAS identified one SNP at the intronic region of kinesin family member 1B (*KIF1B*) [82], although this association has not been confirmed in other populations [83, 84]. Subsequent GWASs identified a significant association of *HLA-DQA1/HLA-DRB1,* glutamate ionotropic receptor kainate type subunit 1 (*GRIK1*) [85], *HLA-DQB1/HLA-DQA1*, and signal transducer and activator of transcription 4 (*STAT4*) with HBV-induced HCC [86]. Trans-ethnic association analysis of *HLA-DPA1/HLA-DPB1* alleles in Asian populations confirmed the association of class II *HLA* alleles [77]. These results suggest that *HLA* class II is strongly associated with both chronic HBV infection and HBV-induced progression of liver disease. Another GWAS in Southern Chinese individuals suggested an association of different host genetic factors with HBV-induced HCC, although these results must be confirmed [87]. Additionally, a GWAS on liver cirrhosis progression identified a novel candidate risk allele in the upstream region of the ferredoxin 1 (*FDX1*) gene among chronic HBV carriers in an Arab population [88].

Variations in HBV genotype are well recognized and HBV genotype and mutations were reported to be associated with HBV-related HCC risk of host genetic variants [89]. This result suggests important interactions between host genome variation and virus genome variation, which is consistent with HIV and HCV infections mentioned above.

6.4 Other Infectious Diseases (Table 6.3)

Leprosy

Leprosy is caused by the pathogenic bacterium *M. leprae* and progresses to peripheral neuropathy and permanent progressive deformity if not treated. Although both leprosy and TB originate from infection by *Mycobacterium* species, GWASs for leprosy have achieved outstanding success compared to those for TB. The first GWAS for leprosy was reported from the Zhang and Liu group in 2009 based on 706 affected cases and 1225 unaffected controls from a Han Chinese population. They detected a strong signal at the *HLA-DR-DQ* locus on chromosome 6p21,

Table 6.3 List of significantly associated genes and polymorphisms in GWAS of other infectious diseases

Gene	Polymorphism	Study design	Cases/ Controls (Population)	P	Effect size (Odds ratio)	Ref.
Leprosy						
C13orf31(LACC1)	rs3764147	Cases vs Controls	3960/7180 (Chinese)	3.72E-54	1.68 (1.57–1.80)	[90]
NOD2	rs9302752	Cases vs Controls	3960/7180 (Chinese)	3.77E-40	1.59 (1.49–1.71)	[90]
LACC1/CCDC122	rs3088362	Cases vs Controls	3960/7180 (Chinese)	1.36E-31	1.52 (1.41–1.63)	[90]
HLA-DR-DQ	rs602875	Cases vs Controls	3960/7180 (Chinese)	5.35E-27	0.67 (0.62–0.72)	[90]
TNFSF15	rs6478108	Cases vs Controls	3960/7180 (Chinese)	3.39E-21	1.37 (1.28–1.46)	[90]
RIPK2	rs42490	Cases vs Controls	3960/7180 (Chinese)	1.38E-16	0.76 (0.71–0.81)	[90]
HLA-DRB1/–DQA1	rs9270650	Cases vs Controls	377/370 (Indian)	4.90E-14	2.30 (1.85–2.86)	[94]
TLR1	rs5743618	Cases vs Controls	434/460 (Indian/ Turkey)	1.70E-09	0.37 (0.26–0.51)	[94]
RAB32	rs2275606	Cases vs Controls	4407/10,880 (Chinese)	3.94E-14	1.30 (1.21–1.39)	[91]
IL23R	rs3762318	Cases vs Controls	4407/10,880 (Chinese)	3.27E-11	0.69 (0.62–0.77)	[91]
CIITA	rs77061563	Cases vs Controls	8313/16,017 (Chinese)	6.23E-15	0.84 (0.80–0.88)	[92]
CCDC88B	rs663743	Cases vs Controls	8313/16,017 (Chinese)	8.84E-14	1.24 (1.17–1.31)	[92]
EGR2	rs58600253	Cases vs Controls	8313/16,017 (Chinese)	3.02E-12	1.22 (1.15–1.29)	[92]
CDH18	rs73058713	Cases vs Controls	8313/16,017 (Chinese)	9.54E-09	1.19 (1.12–1.27)	[92]

(continued)

Table 6.3 (continued)

Gene	Polymorphism	Study design	Cases/ Controls (Population)	P	Effect size (Odds ratio)	Ref.
DEC1	rs10817758	Cases vs Controls	8313/16,017 (Chinese)	1.15E-08	1.13 (1.08–1.18)	[92]
BATF3	rs2221593	Cases vs Controls	8313/16,017 (Chinese)	3.09E-08	1.15 (1.10–1.22)	[92]
CTSB	rs10100465	Cases vs Controls	8156/15,610 (Chinese)	2.85E-11	0.85 (No 95% CI data)	[93]
MED30	rs55894533	Cases vs Controls	8156/15,610 (Chinese)	5.07E-11	1.15 (No 95% CI data)	[93]
BBS9	rs4720118	Cases vs Controls	8156/15,610 (Chinese)	3.85E-10	1.16 (No 95% CI data)	[93]
SYN2	rs6807915	Cases vs Controls	8156/15,610 (Chinese)	1.94E-08	0.89 (No 95% CI data)	[93]
Meningococcus						
CFH/CFHR3	rs426736	Cases vs Controls	1443/6079 (European)	4.60E-13	0.63 (0.55–0.71)	[100]
H. pylori						
TLR1	rs10004195	Anti-H. pylori IgG titer high vs lowanti-H. pylori IgG titer high vs low	2623/7862 (European)	1.40E-18	0.70 (0.65–0.76)	[105]
FCGR2A/ FCGR2B	rs368433	Anti-H. pylori IgG titer high vs lowanti-H. pylori IgG titer high vs low	2623/7862 (European)	2.10E-08	0.73 (0.65–0.81)	[105]
Pneumococcus						
AC011288.2 (lincRNA)	rs140817150	Cases vs Controls	542/4013 (Kenyan)	1.69E-09	2.47 (1.84–3.31)	[108]

(continued)

Table 6.3 (continued)

Gene	Polymorphism	Study design	Cases/ Controls (Population)	P	Effect size (Odds ratio)	Ref.
Salmonella						
HLA-DRB1/-DQB1	rs7765379	Cases vs Controls	522/2011 (Vietnamese/ Nepalese)	2.29E-13	0.22 (0.15–0.34)	[112]
STAT4	No information	Cases vs Controls	323/3013 (Kenyan/ Malawian)	1.40E-09	7.2 (3.8–13.5)	[113]
Staphylococcus						
HLA-DRA/-DRB1	rs115231074	Cases vs Controls	4701/45,344 (American)	1.30E-10	1.22 (No 95% CI data)[a]	[116]
KAT2B	rs61440199	Intermittent nasal carriers vs non-carriers	97/620 (Mexican-American)	8.68E-09	8.68 (4.16–18.13)	[117]
Sepsis						
FER	rs4957796	Non-survivor vs Survivors	460/2078 (European)	5.60E-08	0.56 (0.45–0.69)	[119]
Dengue						
MICB	rs3132468	Cases vs Controls	3745/4952 (Vietnamese)	4.41E-11	1.34 (1.23–1.46)	[122]
PLCE1	rs37665524	Cases vs Controls	3745/4952 (Vietnamese)	3.08E-10	0.80 (0.75–0.86)	[122]
Herpes zoster						
HLA-B/HCP5	rs114045064	Cases vs Controls	2016/16,407 (European)	2.24E-08	0.78 (0.71–0.85)	[130]
HPV						
HLD-DQB1	rs9357152	HPV8 seropositive vs seronegative	1333/3414 (European)	1.20E-10	1.37 (1.24–1.50)	[134]
Leishmaniasis						
HLA-DRB1/–DQA1	rs9271858	Cases vs Controls	2287/2079 (Indian/ Brazilian)	2.76E-17	1.41 (1.30–1.52)	[140]

CI confidence interval
[a]Imputed data

which has been identified using a candidate gene approach [90]. After combining three replication studies from Han Chinese and minority groups in China, receptor-interacting serine/threonine kinase 2 (*RIPK2*), tumor necrosis factor superfamily member 15 (*TNFSF15*), nucleotide-binding oligomerization domain containing 2 (*NOD2*), and laccase domain containing 1/coiled-coil domain containing 122 (*LACC1/CCDC122*) were identified to be associated with leprosy [90]. Thereafter, expanded GWASs in the Chinese population further identified 12 loci including *IL23R* for leprosy risk with a small effect size (odds ratio <1.5) [91–93]. Another GWAS in Indian and Turkey confirmed the association of *HLA-DR-DQ* with leprosy and identified a novel association with toll-like receptor 1 (*TLR1*) [94]. A replication study in Brazilians confirmed the association of *NOD2* and *LACC1/CCDC122* with leprosy [95]. NOD2 recognizes bacterial molecules and triggers innate immune responses. Unique muramyl dipeptide in *M. leprae* was shown to be recognized by NOD2 [96]. Expression of *LACC1* was shown to be regulated by the peroxisome proliferator-activated receptor (PPAR) signaling pathway, which plays important anti-inflammatory roles [97]. In addition, the Zhang group focused on evidence that *NOD2*, *TNFSF15*, *IL23R*, and *LACC1/CCDC122* genes, which have been identified in GWASs of leprosy, have also been reported in GWASs of Crohn's disease and ulcerative colitis, which are autoinflammatory diseases. They evaluated the effect of other Crohn's disease risk genes and revealed the association of *IL18AP/IL18R1* and *IL12B* with leprosy [98]. Variants in *IL18AP/IL18R1* and *IL12B* genes showed opposing associations between leprosy and inflammatory bowel disease. Their results suggest shared or pleiotropic genetic susceptibility between infectious diseases and inflammatory diseases.

Meningococcus

Neisseria meningitidis (meningococcus) causes meningococcal diseases such as meningitis and septicemia, which are major causes of death in children of European descent. The sibling familial risk ratio for meningococcal disease is similar to that for polygenic diseases [99], suggesting the importance of genetic factors in meningococcal disease. A GWAS for meningococcal disease was conducted using 475 cases and 4703 controls in the United Kingdom and identified the association of SNPs at the locus between complement factor H (*CFH*) and CFH-related protein 3 (*CFHR3*) with meningococcal disease [100]. The association was replicated in Western European [100], South European [100], and Central European [101] cohorts. Association of *CFH*-related genes was consistent with *in vitro* evidence that *N. meningitides* evades complement-mediated killing through the binding of host CFH protein to meningococcal factor H-binding protein (fHbp) [102]. Altered risk to *N. meningitides* infection mediated by genetic variation in the *CFH*/*CFHR3* locus was subsequently attributed to differences in circulating levels of CFH protein and CFHR3 protein, which compete for binding to fHbp [103]. Moreover, higher

expression of CFHR3 than CFH was proposed to enhance protection against *N. meningitides* and protect hosts from disease onset [103].

Helicobacter pylori

H. pylori is a major cause of gastritis and gastric ulcers and is linked to the development of cancer. Although 90% of individuals are infected by *H. pylori* in developing countries [104], some individuals are never colonized, regardless of exposure. To identify genetic loci associated with anti-*H. pylori* serum IgG antibody titer, a GWAS compared 2623 cases and 7862 controls from three independent European cohorts. This GWAS identified associations of SNPs on the *TLR* locus and the Fc gamma receptor 2A/2B (*FCGR2A/FCG2B*) locus with anti-*H. pylori* serum IgG antibody titer and the identified SNPs were significantly correlated with mRNA levels of *TLR1* and *FCGR2A/FCGR2B* [105]. The association between the *TLR* locus and anti-*H. pylori* antibody levels was confirmed in an independent Finnish population [106]. TLRs are known to be essential for protective immunity against infection. *TLR1* forms a heterodimer with *TLR2* and recognizes triacylated lipopeptides released from the cell envelope of Gram-negative bacteria [107]. As *H. pylori* possesses lipid A, which can consist of triacylated lipopeptides, TLR-mediated differential prevalence of *H. pylori* antibodies seems biologically plausible.

Pneumococcus

Streptococcus pneumoniae (pneumococcus) causes lung, ear, brain, spinal cord, and bloodstream infections, which can lead to hearing loss, brain damage, and death in young children. Despite widespread exposure and asymptomatic carriage of this bacterium, only a proportion of individuals develop bacterial bloodstream infection (bacteremia). A GWAS consisted of 542 Kenyan children with culture-confirmed pneumococcal bacteremia and 4013 healthy controls identified a statistically significant association of a SNP in a long intergenic non-coding RNA (lincRNA) gene with pneumococcal infection [108]. LincRNAs are transcribed from non-coding DNA sequences between protein-coding genes and more than 8000 human lincRNAs have been reported. Additionally, expression of lincRNA is more tissue-specific than that of protein-coding genes. The associated lincRNA is expressed only in neutrophils, which is consistent with the fact that neutrophils are a major player in pneumococcal clearance [109, 110]. LincRNAs are key regulators of diverse cellular processes through the attachment to messenger RNA to block protein production [111]. To our knowledge, this GWAS is the first to propose that lincRNAs have a role in immunity by regulating host susceptibility to pathogen infections.

Salmonella

Salmonella causes intestinal tract infections through consumption of water or food contaminated with *Salmonella enterica* serovar Typhi, *Salmonella enterica* serovar Paratyphi, or non-typhoidal *Salmonella* (NTS) and leads to diarrhea, fever, vomiting, and abdominal cramps. The first GWAS for *Salmonella* infection was reported in 2014. This three-stage GWAS included 432 patients with clinical signs and symptoms of enteric fever with culture-confirmed *Salmonella* Typhi or *Salmonella* Paratyphi A and 2011 controls in Vietnam, followed by two independent datasets from Nepal and Vietnam. Although almost all cases (>99%) in Vietnam were infected by *Salmonella* Typhi, whereas 66.7% and 33.3% of cases in Nepal were colonized by *Salmonella* Typhi and *Salmonella* Paratyphi A, respectively, a variation in *HLA-DRB1* was associated with resistance to enteric fever both in Vietnam and Nepal [112]. The minor allele of the identified SNP conferred nearly fivefold greater resistance, indicating a substantial effect of *HLA* class II variation on susceptibility to enteric fever caused by *Salmonella* species.

A GWAS for NTS infection was later reported using 180 Kenyan cases and 2677 controls, followed by an replication analysis with 143 Malawian cases and 336 controls [113]. An intronic variant in the *STAT4* gene, which is a well-known cytokine production-related transcriptional factor, was identified and the risk allele for NTS infection was associated with lower *STAT4* gene expression [113]. This finding is consistent with the role of *STAT4* as a transcription activator, which is essential for mediating responses to IL-12 in lymphocytes and regulating T helper cell differentiation.

Staphylococcus

Staphylococcus aureus is present on the nose and skin in 30–50% of healthy individuals. Infection by *S. aureus* can cause a variety of diseases ranging from mild skin and soft tissue, eye (cornea), and bone infections to life-threatening bloodstream, lung, and heart infections. To identify genetic variants for the risk of *S. aureus* infection, initial GWASs compared 361 *S. aureus* bacteremia cases and 699 controls [114] and 309 *S. aureus* infected cases and 2925 uninfected controls [115], however, no statistically significant associations were detected. A larger scale GWAS including 4701 culture-confirmed *S. aureus* infected cases and 45,344 uninfected controls identified a significant association between the *HLA* class II (*HLA-DRA/HLA-DRB1*) region and *S. aureus* infection [116], suggesting that previous GWASs were underpowered to detect an effect at genome-wide significance. Moreover, a GWAS focusing on *S. aureus* nasal carriage identified a significant association of the intronic variant of lysine acetyltransferase 2B (*KAT2B*) gene with intermittent carriage of *S. aureus* [117]. Interestingly, this GWAS recruited only 97 intermittent carriers and 620 non-carriers. KAT2B, also known as P300/

CBP-associated factor, is involved in immune function. *KAT2B* expression in mice was reported to be affected by the nature of the infecting *S. aureus* strain [118]. These results support the importance of *KAT2B* in *S. aureus* infection. As *S. aureus* is known to have a complex infection mechanism, using a wide variety of virulence factors that interact with several host pathways, analyses focusing on the specific status of *S. aureus* may help facilitate identification of host genetic variants for this pathogen.

Sepsis

Sepsis is a complication caused by the body's overwhelming and life-threatening response to a pathogen infection and can lead to tissue damage, organ failure, and death. Sepsis is often associated with infections of the lungs (pneumonia), urinary tract, skin, and gut. Several types of Gram-positive species, e.g., *Staphylococcus* and *Streptococcus*, and Gram-negative species, e.g., *Escherichia* and *Neisseria*, are often observed in sepsis cases. The first GWAS evaluated 28-day survival from sepsis caused by pneumonia and recruited 460 non-survivors and 2078 survivors among European sepsis patients. Although the cases including both Gram-positive and -negative bacteria infected cases, this GWAS successfully identified an association of a SNP in the intronic region of the Fps/Fes related tyrosine kinase (*FER*) gene with reduced risk of death from sepsis [119]. The reduction in mortality associated with this SNP was substantial; the approximately 25% mortality rate observed in major allele homozygous patients was decreased to 15% in heterozygous carriers and 10% in minor allele homozygous carriers. When patients with sepsis due to pneumonia and intra-abdominal infection were combined, no significant associations were detected, suggesting the importance of tailoring homogeneous categories in sepsis. FER is known to have a role in the regulation of neutrophil chemotaxis and endothelial permeability [120, 121]. As neutrophil recruitment to the site of infection is essential in innate immune defense and changes in relevant signaling pathways can lead to failure of bacterial clearance or promotion of tissue damage, FER may be a potential mechanism affecting survival from sepsis.

Dengue

Dengue is an acute systemic viral infection caused by dengue virus. Dengue is a mosquito-borne infection and a wide variety of disease manifestations is seen from asymptomatic infection to severe and fatal hypovolemic shock, called dengue shock syndrome (DSS). In southern Vietnam, serological studies have estimated that 85% of the population is exposed to dengue virus infection by 15 years of age, and DSS is estimated to occur in <1% of exposed individuals. To reveal genetic risk factors for severe dengue, a GWAS comparing 2008 DSS cases among Vietnamese children

and 2018 controls were conducted, and two strong, independent associations were observed between DSS and the MHC class I polypeptide-related sequence B (*MICB*) locus on chromosome 6 near *HLA* class I and II loci and phospholipase C epsilon 1 (*PLCE1*) on chromosome 19 [122]. These associations were confirmed in replication studies in independent Vietnamese DSS cases [122] and Thai cases [123] or non-severe dengue cases without shock [124]. MICB is one of the stress-induced molecules expressed by virus-infected cells and activates the receptor NKG2D on NK cells [15, 125]. Activated NK cells induce killing of virus-infected cells through cytokine expression and cytolytic response [126]. The DSS risk allele was significantly associated with lower mRNA expression of *MICB* [123], which can lead to decreased killing by NK cells early in infection and increased viral burden for severe disease progression. Additionally, mutations in *PLCE1* have been shown to be associated with nephrotic syndrome [127], which leads to proteinuria that has been proposed as a potential predictor in determining the risk to develop severe dengue [128].

Herpes Zoster

Herpes zoster, also known as shingles, is caused by varicella zoster virus (VZV). VZV initially manifests as chicken pox. It can remain asymptomatic in nerve tissues for many years but later lead to a painful skin rash with blisters in a localized area. In the absence of vaccination, person who live to 85 years of age has a 50% risk of herpes zoster and 10–50% of them will develop chronic postherpetic neuralgia [129]. To identify the genetic risk for re-emergence of VZV, a GWAS was conducted including 2016 cases and 16,407 controls in the European ancestry group and identified protective variants in the HLA-B/HCP5 locus [130]. This locus has been associated with delayed development of AIDS as described above, suggesting a shared and critical role of the HLA-B/HCP5 locus in viral suppression.

Human Papillomavirus (HPV)

HPV is a DNA virus that infects mucosal or cutaneous epithelia through skin-to-skin contact and causes warts, squamous intraepithelial lesions, and anogenital and oropharyngeal cancers, such as cervical cancers. HPVs have great diversity in their genomes and more than 100 HPV types, which share nucleotide identity, have been reported [131]. A limited number of HPV types cause anogenital and oropharyngeal cancers, whereas other HPV types lead to non-melanoma skin cancer [132]. Antibodies against HPVs are considered as markers for HPV infections, however, not all infected persons show detectable levels of specific antibodies [133]. To understand the genetic basis of serological immune responses to HPV infections, a GWAS evaluated serology data on 13 HPV types in 4811 European subjects with

lung, head and neck, and kidney cancers. A significant association between HPV8 seropositivity and a SNP located in the *HLA* class II region was identified, and this association was subsequently confirmed in an independent set of 2344 Latin American patients with head and neck cancers [134]. These results provide a proof of concept that genetic variation plays a role in antibody reactivity to HPV infection.

Influenza

Influenza is an infectious disease caused by influenza virus that leads to a high fever, runny nose, sore throat, muscle pains, headache, and coughing. Influenza virus A (H1N1)pdm09 caused the first influenza pandemic of the twenty-first century in 2009 and avian influenza A(H7N9) caused a >30% case-fatality rate in 2013–2014 [135]. Although statistically significant associations between influenza virus infection and host genetic factors have not yet been identified due to limited sample sizes, GWASs for these pandemic influenza virus infections have been attempted. A GWAS on influenza A(H7N9) compared 102 A(H7N9) patients and 106 heavily-exposed healthy poultry workers and revealed a potential association with lectin galactoside-binding soluble 1 (*LGALS1*) gene variants, which regulates the expression of a beta-galactoside-binding protein, galectin 1 [136]. Moreover, another GWAS on influenza A(H1N1)pdm09 compared 162 cases with severe infection and 247 controls with mild infection and suggested an association of higher-expression variants of the transmembrane protease, serine 2 (*TMPRSS2*) gene with a risk of severe A(H1N1)pdm09 influenza infection [137]. Interestingly, this GWAS reported that the same risk variants increased susceptibility to human A(H7N9) influenza [137], and *TMPRSS2*-knockout mice were highly tolerant to the lethal challenge of A(H1N1)pdm09 and A(H7N9) viruses, demonstrating the essential role of TMPRSS2 during influenza virus infections [138].

Leishmaniasis

Leishmaniasis is an infectious disease caused by protozoan parasites, *Leishmania* species, which live in macrophages and are transmitted by sand flies. Most infected people remain asymptomatic throughout life, whereas some infected people develop cutaneous, mucocutaneous, and visceral leishmaniasis [139]. Visceral leishmaniasis can be fatal if not treated. A GWAS on visceral leishmaniasis compared 2287 cases and 2079 controls in Indian and Brazilian populations. Although leishmaniasis in India is caused by *L. donovani* and that in Brazil is caused by *L. infantum chagasi*, the combined analysis successfully identified significant associations between visceral leishmaniasis and *HLA* class II region polymorphisms [140]. This result indicated shared genetic risk factors for visceral leishmaniasis that cross the human

population and parasite species, emphasizing the biological importance of peptide presentation from infected macrophages and dendritic cells to CD4-positive T cells to drive immune responses to this pathogen.

6.5 Conclusions

During the last decade since the first GWAS report in infectious disease, we have observed the identification of various genetic factors associated with a variety of clinical manifestations. Especially, the importance of *HLA* genes has been confirmed in GWASs. Associations have been reported between *HLA* class I alleles and AIDS, HBV infection, and herpes zoster, and between *HLA* class II alleles and TB, HCV- and HBV-infection, HCV- and HBV-related diseases, leprosy, *Salmonella* infection, *Staphylococcus* infection, HPV infection, and leishmaniasis. Associations between non-HLA genes and infection have also been reported with biological plausibility. Especially, the identification of *IL28B* in HCV clearance is a striking example to illustrate the impact of GWASs in infectious disease [54–56, 58–60].

Simple case-control GWASs have, of course, identified significant associations between host genetic factors and disease. Some researchers have focused on heterogeneity of disease onset in infectious disease and identified genetic factors that showed more clear associations in selected patients. For example, Mahasirimongkol et al. focused on young age onset TB patients and identified a significant association of one SNP that did not reach significance before the consideration of disease onset [49]. Brown et al. focused on nasal infection of *S. aureus* and revealed a clear and significant association with *KAT2B* variant in a relatively smaller sample number than that used to identify genetic variants associated with *Staphylococcus* bloodstream infection [117]. As pathogens possess a wide variety of virulence factors that interact with several host pathways during infection, analyses focusing on a specific aspect of infection may reveal more clear associations with host genetics in infectious diseases.

Furthermore, specific associations between pathogen genome variants and host genetic susceptibility factors have been identified, as described above, for AIDS [31], TB [52, 53], HCV [59, 64], HBV [89], and HPV [134], whereas several simple case-control association studies without considering pathogen species have also demonstrated specific associations in the susceptibility to *Salmonella* species [112], *Leishmania* species [140], and sepsis survival [119]. This evidence suggests that heterogeneity in pathogen genomes can also be an important factor in host susceptibility to infectious diseases. Consideration of both host and pathogen factors in GWASs can provide critical clues to reveal detailed mechanisms involved in infectious diseases.

References

1. Frodsham AJ, Hill AV (2004) Genetics of infectious diseases. Hum Mol Genet 13(2):R187–R194
2. Blum JS, Wearsch PA et al (2013) Pathways of antigen processing. Annu Rev Immunol 31:443–473
3. Fellay J, Shianna KV et al (2007) A whole-genome association study of major determinants for host control of HIV-1. Science (New York, NY) 317(5840):944–947
4. Dalmasso C, Carpentier W et al (2008) Distinct genetic loci control plasma HIV-RNA and cellular HIV-DNA levels in HIV-1 infection: the ANRS genome wide association 01 study. PLoS One 3(12):e3907
5. Pelak K, Goldstein DB et al (2010) Host determinants of HIV-1 control in African Americans. J Infect Dis 201(8):1141–1149
6. Wei Z, Liu Y et al (2015) Genome-wide association studies of HIV-1 host control in ethnically diverse Chinese populations. Sci Rep 5:10879
7. McLaren PJ, Coulonges C et al (2015) Polymorphisms of large effect explain the majority of the host genetic contribution to variation of HIV-1 virus load. Proc Natl Acad Sci U S A 112(47):14658–14663
8. Apps R, Qi Y et al (2013) Influence of HLA-C expression level on HIV control. Science (New York, NY) 340(6128):87–91
9. Kulkarni S, Savan R et al (2011) Differential microRNA regulation of HLA-C expression and its association with HIV control. Nature 472(7344):495–498
10. Fellay J, Ge D et al (2009) Common genetic variation and the control of HIV-1 in humans. PLoS Genet 5(12):e1000791
11. Limou S, Le Clerc S et al (2009) Genomewide association study of an AIDS-nonprogression cohort emphasizes the role played by HLA genes (ANRS genomewide association study 02). J Infect Dis 199(3):419–426
12. Pereyra F, Jia X et al (2010) The major genetic determinants of HIV-1 control affect HLA class I peptide presentation. Science (New York, NY) 330(6010):1551–1557
13. Guergnon J, Dalmasso C et al (2012) Single-nucleotide polymorphism-defined class I and class III major histocompatibility complex genetic subregions contribute to natural long-term nonprogression in HIV infection. J Infect Dis 205(5):718–724
14. Le Clerc S, Delaneau O et al (2014) Evidence after imputation for a role of MICA variants in nonprogression and elite control of HIV type 1 infection. J Infect Dis 210(12):1946–1950
15. Bauer S, Groh V et al (1999) Activation of NK cells and T cells by NKG2D, a receptor for stress-inducible MICA. Science (New York, NY) 285(5428):727–729
16. Limou S, Coulonges C et al (2010) Multiple-cohort genetic association study reveals CXCR6 as a new chemokine receptor involved in long-term nonprogression to AIDS. J Infect Dis 202(6):908–915
17. Troyer JL, Nelson GW et al (2011) Genome-wide association study implicates PARD3B-based AIDS restriction. J Infect Dis 203(10):1491–1502
18. Herbeck JT, Gottlieb GS et al (2010) Multistage genomewide association study identifies a locus at 1q41 associated with rate of HIV-1 disease progression to clinical AIDS. J Infect Dis 201(4):618–626
19. Le Clerc S, Coulonges C et al (2011) Screening low-frequency SNPS from genome-wide association study reveals a new risk allele for progression to AIDS. J Acquir Immune Defic Syndr 56(3):279–284
20. Le Clerc S, Limou S et al (2009) Genomewide association study of a rapid progression cohort identifies new susceptibility alleles for AIDS (ANRS genomewide association study 03). J Infect Dis 200(8):1194–1201
21. Lingappa JR, Petrovski S et al (2011) Genomewide association study for determinants of HIV-1 acquisition and viral set point in HIV-1 serodiscordant couples with quantified virus exposure. PLoS One 6(12):e28632

22. Petrovski S, Fellay J et al (2011) Common human genetic variants and HIV-1 susceptibility: a genome-wide survey in a homogeneous African population. AIDS (London, England) 25(4):513–518
23. Johnson EO, Hancock DB et al (2015) Novel genetic locus implicated for HIV-1 acquisition with putative regulatory links to HIV replication and infectivity: a genome-wide association study. PLoS One 10(3):e0118149
24. Joubert BR, Lange EM et al (2010) A whole genome association study of mother-to-child transmission of HIV in Malawi. Genome Med 2(3):17
25. Luo M, Sainsbury J et al (2012) A genetic polymorphism of FREM1 is associated with resistance against HIV infection in the Pumwani sex worker cohort. J Virol 86(21):11899–11905
26. Limou S, Delaneau O et al (2012) Multicohort genomewide association study reveals a new signal of protection against HIV-1 acquisition. J Infect Dis 205(7):1155–1162
27. McLaren PJ, Coulonges C et al (2013) Association study of common genetic variants and HIV-1 acquisition in 6,300 infected cases and 7,200 controls. PLoS Pathog 9(7):e1003515
28. Euler Z, van Gils MJ et al (2013) Genome-wide association study on the development of cross-reactive neutralizing antibodies in HIV-1 infected individuals. PLoS One 8(1):e54684
29. Bol SM, Moerland PD et al (2011) Genome-wide association study identifies single nucleotide polymorphism in DYRK1A associated with replication of HIV-1 in monocyte-derived macrophages. PLoS One 6(2):e17190
30. van Manen D, Delaneau O et al (2011) Genome-wide association scan in HIV-1-infected individuals identifying variants influencing disease course. PLoS One 6(7):e22208
31. Bartha I, Carlson JM et al (2013) A genome-to-genome analysis of associations between human genetic variation, HIV-1 sequence diversity, and viral control. elife 2:e01123
32. Hill AV (1999) The immunogenetics of resistance to malaria. Proc Assoc Am Physicians 111(4):272–277
33. Mackinnon MJ, Mwangi TW et al (2005) Heritability of malaria in Africa. PLoS Med 2(12):e340
34. Jallow M, Teo YY et al (2009) Genome-wide and fine-resolution association analysis of malaria in West Africa. Nat Genet 41(6):657–665
35. Timmann C, Thye T et al (2012) Genome-wide association study indicates two novel resistance loci for severe malaria. Nature 489(7416):443–446
36. Band G, Rockett KA et al (2015) A novel locus of resistance to severe malaria in a region of ancient balancing selection. Nature 526(7572):253–257
37. Loscertales MP, Owens S et al (2007) ABO blood group phenotypes and *Plasmodium falciparum* malaria: unlocking a pivotal mechanism. Adv Parasitol 65:1–50
38. Reappraisal of known malaria resistance loci in a large multicenter study (2014) Nat Genet 46(11):1197–1204
39. Stauffer TP, Guerini D et al (1995) Tissue distribution of the four gene products of the plasma membrane Ca2+ pump. A study using specific antibodies. J Biol Chem 270(20):12184–12190
40. Louicharoen C, Patin E et al (2009) Positively selected G6PD-Mahidol mutation reduces *Plasmodium vivax* density in Southeast Asians. Science (New York, NY) 326(5959):1546–1549
41. World Health Organization Global tuberculosis report 2015
42. Thye T, Vannberg FO et al (2010) Genome-wide association analyses identifies a susceptibility locus for tuberculosis on chromosome 18q11.2. Nat Genet 42(9):739–741
43. Thye T, Owusu-Dabo E et al (2012) Common variants at 11p13 are associated with susceptibility to tuberculosis. Nat Genet 44(3):257–259
44. Curtis J, Luo Y et al (2015) Susceptibility to tuberculosis is associated with variants in the ASAP1 gene encoding a regulator of dendritic cell migration. Nat Genet 47(5):523–527
45. Sveinbjornsson G, Gudbjartsson DF et al (2016) HLA class II sequence variants influence tuberculosis risk in populations of European ancestry. Nat Genet 48(3):318–322
46. Chimusa ER, Zaitlen N et al (2014) Genome-wide association study of ancestry-specific TB risk in the South African Coloured population. Hum Mol Genet 23(3):796–809
47. Grant AV, Sabri A et al (2016) A genome-wide association study of pulmonary tuberculosis in Morocco. Hum Genet 135(3):299–307

48. Png E, Alisjahbana B et al (2012) A genome wide association study of pulmonary tuberculosis susceptibility in Indonesians. BMC Med Genet 13:5
49. Mahasirimongkol S, Yanai H et al (2012) Genome-wide association studies of tuberculosis in Asians identify distinct at-risk locus for young tuberculosis. J Hum Genet 57(6):363–367
50. Sobota RS, Stein CM et al (2016) A locus at 5q33.3 confers resistance to tuberculosis in highly susceptible individuals. Am J Hum Genet 98(3):514–524
51. Coscolla M, Gagneux S (2014) Consequences of genomic diversity in *Mycobacterium tuberculosis*. Semin Immunol 26(6):431–444
52. Omae Y, Toyo-Oka L, Tokunaga K, Yanai H, Nedsuwan S, Wattanapokayakit S, Satproedprai N, Inunchot W, Wichukchinda N, Mahasirimongkol S, Smittipat N, Palittapongarnpim P, Pasomsub E, Sawanpanyalert P, Mushiroda T, Kubo M (2017) Pathogen lineage-based genome-wide association study identified CD53 as susceptible locus in tuberculosis. J Hum Genet 62(12):1015–1022
53. Toyo-Oka L, Mahasirimongkol S, Yanai H, Mushiroda T, Wattanapokayakit S, Wichukchinda N, Yamada N, Smittipat N, Juthayothin T, Palittapongarnpim P, Nedsuwan S, Kantipong P, Takahashi A, Kubo M, Sawanpanyalert P, Tokunaga K (2017) Strain-based HLA association analysis identified *HLADRB1** 09:01 associated with modern strain tuberculosis. HLA 90(3):149–156. https://doi.org/10.1111/tan.13070
54. Ge D, Fellay J et al (2009) Genetic variation in IL28B predicts hepatitis C treatment-induced viral clearance. Nature 461(7262):399–401
55. Suppiah V, Moldovan M et al (2009) IL28B is associated with response to chronic hepatitis C interferon-alpha and ribavirin therapy. Nat Genet 41(10):1100–1104
56. Tanaka Y, Nishida N et al (2009) Genome-wide association of IL28B with response to pegylated interferon-alpha and ribavirin therapy for chronic hepatitis C. Nat Genet 41(10):1105–1109
57. McHutchison JG, Lawitz EJ et al (2009) Peginterferon alfa-2b or alfa-2a with ribavirin for treatment of hepatitis C infection. N Engl J Med 361(6):580–593
58. Thomas DL, Thio CL et al (2009) Genetic variation in IL28B and spontaneous clearance of hepatitis C virus. Nature 461(7265):798–801
59. Rauch A, Kutalik Z et al (2010) Genetic variation in IL28B is associated with chronic hepatitis C and treatment failure: a genome-wide association study. Gastroenterology 138(4):1338–1345, 1345.e1331–1337
60. Duggal P, Thio CL et al (2013) Genome-wide association study of spontaneous resolution of hepatitis C virus infection: data from multiple cohorts. Ann Intern Med 158(4):235–245
61. Kotenko SV, Gallagher G et al (2003) IFN-lambdas mediate antiviral protection through a distinct class II cytokine receptor complex. Nat Immunol 4(1):69–77
62. Sheppard P, Kindsvogel W et al (2003) IL-28, IL-29 and their class II cytokine receptor IL-28R. Nat Immunol 4(1):63–68
63. Miki D, Ochi H et al (2013) HLA-DQB1∗03 confers susceptibility to chronic hepatitis C in Japanese: a genome-wide association study. PLoS One 8(12):e84226
64. Hayashi K, Katano Y et al (2011) Association of interleukin 28B and mutations in the core and NS5A region of hepatitis C virus with response to peg-interferon and ribavirin therapy. Liver Int: Off J Int Assoc Study Liver 31(9):1359–1365
65. Kumar V, Kato N et al (2011) Genome-wide association study identifies a susceptibility locus for HCV-induced hepatocellular carcinoma. Nat Genet 43(5):455–458
66. Miki D, Ochi H et al (2011) Variation in the DEPDC5 locus is associated with progression to hepatocellular carcinoma in chronic hepatitis C virus carriers. Nat Genet 43(8):797–800
67. Matsuura K, Sawai H et al (2017) Genome-wide association atudy identifies TLL1 variant associated with development of hepatocellular carcinoma after eradication of hepatitis C virus infection. Gastroenterology 152:1383–1394
68. Patin E, Kutalik Z et al (2012) Genome-wide association study identifies variants associated with progression of liver fibrosis from HCV infection. Gastroenterology 143(5):1244–1252. e1241–1212

69. Urabe Y, Ochi H et al (2013) A genome-wide association study of HCV-induced liver cirrhosis in the Japanese population identifies novel susceptibility loci at the MHC region. J Hepatol 58(5):875–882
70. Zignego AL, Wojcik GL et al (2014) Genome-wide association study of hepatitis C virus- and cryoglobulin-related vasculitis. Genes Immun 15(7):500–505
71. Nagao Y, Nishida N et al (2017) Genome-wide association study identifies risk variants for lichen planus in patients with hepatitis C virus infection. Clin Gastroenterol Hepatol: Off Clin Pract J Am Gastroenterol Assoc 15:937–944.e5
72. Kamatani Y, Wattanapokayakit S et al (2009) A genome-wide association study identifies variants in the HLA-DP locus associated with chronic hepatitis B in Asians. Nat Genet 41(5):591–595
73. Mbarek H, Ochi H et al (2011) A genome-wide association study of chronic hepatitis B identified novel risk locus in a Japanese population. Hum Mol Genet 20(19):3884–3892
74. Nishida N, Sawai H et al (2012) Genome-wide association study confirming association of HLA-DP with protection against chronic hepatitis B and viral clearance in Japanese and Korean. PLoS One 7(6):e39175
75. Hu Z, Liu Y et al (2013) New loci associated with chronic hepatitis B virus infection in Han Chinese. Nat Genet 45(12):1499–1503
76. Chang SW, Fann CS et al (2014) A genome-wide association study on chronic HBV infection and its clinical progression in male Han-Taiwanese. PLoS One 9(6):e99724
77. Nishida N, Sawai H et al (2014) New susceptibility and resistance HLA-DP alleles to HBV-related diseases identified by a trans-ethnic association study in Asia. PLoS One 9(2):e86449
78. Wake CT (1986) Molecular biology of the HLA class I and class II genes. Mol Biol Med 3(1):1–11
79. Kim YJ, Kim HY et al (2013) A genome-wide association study identified new variants associated with the risk of chronic hepatitis B. Hum Mol Genet 22(20):4233–4238
80. Jiang DK, Ma XP et al (2015) Genetic variants in five novel loci including CFB and CD40 predispose to chronic hepatitis B. Hepatology (Baltimore, Md) 62(1):118–128
81. Li Y, Si L et al (2016) Genome-wide association study identifies 8p21.3 associated with persistent hepatitis B virus infection among Chinese. Nat Commun 7:11664
82. Zhang H, Zhai Y et al (2010) Genome-wide association study identifies 1p36.22 as a new susceptibility locus for hepatocellular carcinoma in chronic hepatitis B virus carriers. Nat Genet 42(9):755–758
83. Zhong R, Tian Y et al (2012) HBV-related hepatocellular carcinoma susceptibility gene KIF1B is not associated with development of chronic hepatitis B. PLoS One 7(2):e28839
84. Sawai H, Nishida N et al (2012) No association for Chinese HBV-related hepatocellular carcinoma susceptibility SNP in other East Asian populations. BMC Med Genet 13:47
85. Li S, Qian J et al (2012) GWAS identifies novel susceptibility loci on 6p21.32 and 21q21.3 for hepatocellular carcinoma in chronic hepatitis B virus carriers. PLoS Genet 8(7):e1002791
86. Jiang DK, Sun J et al (2013) Genetic variants in STAT4 and HLA-DQ genes confer risk of hepatitis B virus-related hepatocellular carcinoma. Nat Genet 45(1):72–75
87. Chan KY, Wong CM et al (2011) Genome-wide association study of hepatocellular carcinoma in Southern Chinese patients with chronic hepatitis B virus infection. PLoS One 6(12):e28798
88. Al-Qahtani A, Khalak HG et al (2013) Genome-wide association study of chronic hepatitis B virus infection reveals a novel candidate risk allele on 11q22.3. J Med Genet 50(11):725–732
89. Wen J, Song C et al (2015) Hepatitis B virus genotype, mutations, human leukocyte antigen polymorphisms and their interactions in hepatocellular carcinoma: a multi-centre case-control study. Sci Rep 5:16489
90. Zhang FR, Huang W et al (2009) Genomewide association study of leprosy. N Engl J Med 361(27):2609–2618
91. Zhang F, Liu H et al (2011) Identification of two new loci at IL23R and RAB32 that influence susceptibility to leprosy. Nat Genet 43(12):1247–1251
92. Liu H, Irwanto A et al (2015) Discovery of six new susceptibility loci and analysis of pleiotropic effects in leprosy. Nat Genet 47(3):267–271

93. Wang Z, Sun Y et al (2016) A large-scale genome-wide association and meta-analysis identified four novel susceptibility loci for leprosy. Nat Commun 7:13760
94. Wong SH, Gochhait S et al (2010) Leprosy and the adaptation of human toll-like receptor 1. PLoS Pathog 6:e1000979
95. Sales-Marques C, Salomao H et al (2014) NOD2 and CCDC122-LACC1 genes are associated with leprosy susceptibility in Brazilians. Hum Genet 133(12):1525–1532
96. Schenk M, Mahapatra S et al (2016) Human NOD2 recognizes structurally unique muramyl dipeptides from Mycobacterium leprae. Infect Immun 84(9):2429–2438
97. Assadi G, Vesterlund L et al (2016) Functional analyses of the Crohn's disease risk gene LACC1. PLoS One 11(12):e0168276
98. Liu H, Irwanto A et al (2012) Identification of IL18RAP/IL18R1 and IL12B as leprosy risk genes demonstrates shared pathogenesis between inflammation and infectious diseases. Am J Hum Genet 91(5):935–941
99. Haralambous E, Weiss HA et al (2003) Sibling familial risk ratio of meningococcal disease in UK Caucasians. Epidemiol Infect 130(3):413–418
100. Davila S, Wright VJ et al (2010) Genome-wide association study identifies variants in the CFH region associated with host susceptibility to meningococcal disease. Nat Genet 42(9):772–776
101. Biebl A, Muendlein A et al (2015) Confirmation of host genetic determinants in the CFH region and susceptibility to meningococcal disease in a central European study sample. Pediatr Infect Dis J 34(10):1115–1117
102. Schneider MC, Prosser BE et al (2009) Neisseria meningitidis recruits factor H using protein mimicry of host carbohydrates. Nature 458(7240):890–893
103. Caesar JJ, Lavender H et al (2014) Competition between antagonistic complement factors for a single protein on N. meningitidis rules disease susceptibility. eLife 3
104. Bardhan PK (1997) Epidemiological features of helicobacter pylori infection in developing countries. Clin Infect Dis 25(5):973–978
105. Mayerle J, den Hoed CM et al (2013) Identification of genetic loci associated with helicobacter pylori serologic status. JAMA 309(18):1912–1920
106. Sung H, Camargo MC et al (2015) Association of 4p14 TLR locus with antibodies to helicobacter pylori. Genes Immun 16(8):567–570
107. Jin MS, Kim SE et al (2007) Crystal structure of the TLR1-TLR2 heterodimer induced by binding of a tri-acylated lipopeptide. Cell 130(6):1071–1082
108. Rautanen A, Pirinen M et al (2016) Polymorphism in a lincRNA associates with a doubled risk of pneumococcal bacteremia in Kenyan children. Am J Hum Genet 98(6):1092–1100
109. Gingles NA, Alexander JE et al (2001) Role of genetic resistance in invasive pneumococcal infection: identification and study of susceptibility and resistance in inbred mouse strains. Infect Immun 69(1):426–434
110. Brinkmann V, Reichard U et al (2004) Neutrophil extracellular traps kill bacteria. Science (New York, NY) 303(5663):1532–1535
111. Kung JT, Colognori D et al (2013) Long noncoding RNAs: past, present, and future. Genetics 193(3):651–669
112. Dunstan SJ, Hue NT et al (2014) Variation at HLA-DRB1 is associated with resistance to enteric fever. Nat Genet 46(12):1333–1336
113. Gilchrist JJ, Mills TC et al (2015) Genetic variants associated with non-typhoidal Salmonella bacteraemia in African children. Lancet (London, England) 385(Suppl 1):S13
114. Nelson CL, Pelak K et al (2014) A genome-wide association study of variants associated with acquisition of *Staphylococcus aureus* bacteremia in a healthcare setting. BMC Infect Dis 14:83
115. Ye Z, Vasco DA et al (2014) Genome wide association study of SNP-, gene-, and pathway-based approaches to identify genes influencing susceptibility to *Staphylococcus aureus* infections. Front Genet 5:125
116. DeLorenze GN, Nelson CL et al (2016) Polymorphisms in HLA class II genes are associated with susceptibility to *Staphylococcus aureus* infection in a white population. J Infect Dis 213(5):816–823

117. Brown EL, Below JE et al (2015) Genome-wide association study of *Staphylococcus aureus* carriage in a community-based sample of Mexican-Americans in Starr County, Texas. PLoS One 10(11):e0142130
118. Modak R, Das Mitra S et al (2014) Epigenetic response in mice mastitis: role of histone H3 acetylation and microRNA(s) in the regulation of host inflammatory gene expression during *Staphylococcus aureus* infection. Clin Epigenetics 6(1):12
119. Rautanen A, Mills TC et al (2015) Genome-wide association study of survival from sepsis due to pneumonia: an observational cohort study. Lancet Respir Med 3(1):53–60
120. Khajah M, Andonegui G et al (2013) Fer kinase limits neutrophil chemotaxis toward end target chemoattractants. J Immunol (Baltimore, Md: 1950) 190(5):2208–2216
121. Qi W, Ebbert KV et al (2005) Absence of Fer protein tyrosine kinase exacerbates endotoxin induced intestinal epithelial barrier dysfunction in vivo. Gut 54(8):1091–1097
122. Khor CC, Chau TN et al (2011) Genome-wide association study identifies susceptibility loci for dengue shock syndrome at MICB and PLCE1. Nat Genet 43(11):1139–1141
123. Dang TN, Naka I et al (2014) A replication study confirms the association of GWAS-identified SNPs at MICB and PLCE1 in Thai patients with dengue shock syndrome. BMC Med Genet 15:58
124. Whitehorn J, Chau TN et al (2013) Genetic variants of MICB and PLCE1 and associations with non-severe dengue. PLoS One 8(3):e59067
125. Jinushi M, Takehara T et al (2003) Critical role of MHC class I-related chain A and B expression on IFN-alpha-stimulated dendritic cells in NK cell activation: impairment in chronic hepatitis C virus infection. J Immunol (Baltimore, Md: 1950) 170(3):1249–1256
126. Trinchieri G (1995) Natural killer cells wear different hats: effector cells of innate resistance and regulatory cells of adaptive immunity and of hematopoiesis. Semin Immunol 7(2):83–88
127. Hinkes B, Wiggins RC et al (2006) Positional cloning uncovers mutations in PLCE1 responsible for a nephrotic syndrome variant that may be reversible. Nat Genet 38(12):1397–1405
128. Vasanwala FF, Puvanendran R et al (2011) Could peak proteinuria determine whether patient with dengue fever develop dengue hemorrhagic/dengue shock syndrome? – a prospective cohort study. BMC Infect Dis 11:212
129. Cohen JI (2013) Clinical practice: Herpes zoster. N Engl J Med 369(3):255–263
130. Crosslin DR, Carrell DS et al (2015) Genetic variation in the HLA region is associated with susceptibility to herpes zoster. Genes Immun 16(1):1–7
131. de Villiers EM, Fauquet C et al (2004) Classification of papillomaviruses. Virology 324(1):17–27
132. Bouvard V, Baan R et al (2009) A review of human carcinogens – Part B: biological agents. Lancet Oncol 10(4):321–322
133. Frazer IH (2009) Interaction of human papillomaviruses with the host immune system: a well evolved relationship. Virology 384(2):410–414
134. Chen D, McKay JD et al (2011) Genome-wide association study of HPV seropositivity. Hum Mol Genet 20(23):4714–4723
135. Widdowson MA, Iuliano AD et al (2014) Challenges to global pandemic mortality estimation. Lancet Infect Dis 14(8):670–672
136. Chen Y, Zhou J et al (2015) Functional variants regulating LGALS1 (Galectin 1) expression affect human susceptibility to influenza A(H7N9). Sci Rep 5:8517
137. Cheng Z, Zhou J et al (2015) Identification of TMPRSS2 as a susceptibility gene for severe 2009 pandemic A(H1N1) influenza and A(H7N9) influenza. J Infect Dis 212(8):1214–1221
138. Sakai K, Ami Y et al (2014) The host protease TMPRSS2 plays a major role in in vivo replication of emerging H7N9 and seasonal influenza viruses. J Virol 88(10):5608–5616
139. World Health Organization (2016) Leishmaniasis in high-burden countries: an epidemiological update based on data reported in 2014
140. Fakiola M, Strange A et al (2013) Common variants in the HLA-DRB1-HLA-DQA1 HLA class II region are associated with susceptibility to visceral leishmaniasis. Nat Genet 45(2):208–213

Chapter 7
Pharmacogenomics

Hitoshi Zembutsu

Abstract Pharmacogenomics is the field of study to discover the genetic factors which affect the response to drugs. The final goal of the pharmacogenomics is to identify clinically useful biomarkers for the drug efficacy or toxicity and to provide the most appropriate drugs to each individual based on the results of genetic test. Genome-wide association study (GWAS) has considered to be a powerful tool to identify novel genetic variations related to disease susceptibility as well as drug efficacy and toxicity. The results of GWAS could clarify the cause of the diseases or interindividual differences of drug response. The validation studies or meta-analysis for the results of GWASs are essential for clinical application of biomarkers identified in the GWASs. This chapter highlights the notable results of pharmacogenomic GWASs which have been published until today.

Keywords Genome-wide association study · Pharmacogenomics · Precision medicine · Adverse drug reactions (ADRs) · Drug efficacy · Toxicity

7.1 Introduction

Pharmacogenomics, which is a part of precision medicine, is the study to discover the role of genetic variations that affect drug-response phenotype such as responder or nonresponder to the drug or adverse drug reactions. Genetic variations including common and rare genetic variants in the genes encoding drug transporters or enzymes could explain a part of this interindividual difference in drug-response phenotype such as drug efficacy or drug toxicity. Through the candidate gene approach, genetic variation in *TPMT* gene has been identified as a well-known biomarker for the risk of 6-mercaptopurine-induced myelosuppression for the treatment of acute lymphoblastic leukemia, and the genetic variation in *UGT1A1* also has been reported to be a biomarker for camptothecin-induced neutropenia and diarrhea for the

H. Zembutsu (✉)
Cancer Precision Medicine Center, Japanese Foundation for Cancer Research, Tokyo, Japan
e-mail: hitoshi.zembutsu@jfcr.or.jp

T. Tsunoda et al. (eds.), *Genome-Wide Association Studies*,
https://doi.org/10.1007/978-981-13-8177-5_7

treatment of solid cancers [1, 2]. The drug labels of 6-mercaptopurine and camptothecin were revised by the Food and Drug Administration (FDA) in United States to describe that genotypes of *TPMT* and *UGT1A1* could be risk factors for toxicity, and they stated that the genotypes of the above genes could be useful to predict the risk of adverse drug reactions before initiation of the treatment [3]. The main approach of pharmacogenomics studies has been a candidate gene approach focusing on the genes involved in drug transport, metabolism, and so on. Moreover, the association of genes involved in immune-mediated responses such as human leukocyte antigens (HLA) and the drug response has also been reported by many research groups [4, 5].

Pharmacogenomics studies using candidate gene approach have been extremely successful in human genetics; however, since genome-wide association studies (GWAS) had been prevalent, genetic variations associated with drug response as well as susceptibility of common diseases have been successfully identified [6], and it has become one of the most powerful tools in pharmacogenomics study. Although the candidate gene approach mainly focuses on only the genes involved in pharmacokinetics or drug metabolism, GWAS could discover novel biomarker genes or genetic variations, and it could identify the novel mechanism which regulate the efficacy or toxicity of the drugs. Since President Obama announced the Precision Medicine Initiative in 2015, the integration of genetic and environmental factors has been thought to be of importance to classify the subpopulations of patients based on their susceptibility to diseases or their responses to the treatments [7]. In this chapter, we describe the current status of pharmacogenomics in precision medicine, indicating the promising biomarkers, which have been identified through GWAS, for the response to the drugs including the anticancer drugs and adverse drug reactions, suggesting the possibility of clinical application.

7.2 GWAS of Adverse Drug Reactions

The World Health Organization (WHO) defines adverse drug reactions (ADRs) as "any noxious, unintentional, and undesired effect of a drug, which occurs at doses used in humans for prophylaxis, diagnosis, or therapy" [8, 9]. ADRs are one of the major issues in drug treatment because they could interfere with continuous and effective drug treatment, and lead to unnecessary hospital admission and result in death. Basically, adverse drug reactions (ADRs) are classified into two categories: type A pharmacological and type B idiosyncratic [10]. The former represents an augmentation of the pharmacological actions of a drug and is dose-dependent and therefore readily reversible on decreasing the dose or withdrawing the drug administration. On the other hand, the latter is usually unrelated to the dose and unpredictable from the general pharmacological information such as dose of the drug.

Although GWAS to identify susceptible genes of common disease usually require thousands of cases and controls, the GWAS for pharmacogenomics could identify the loci which are associated with ADRs with genome-wide significant levels using relatively smaller sample sizes [11, 12]. Table 7.1 shows the 7 genome-wide association

Table 7.1 The results of GWASs for adverse drug reactions

Drug	ADR	The number of samples (case/control)	Ethnicity	Gene	SNP	*P*-value	Odds ratio(hazard ratio)	References
Paclitaxel	Neurotoxicity	Discovery: 280/575	Discovery: European	*FGD4*	rs10771973	2.6×10^{-6}	HR[a]: 1.57	13
		Replication: 82/189	Replication: European, African American					
Chemotherapy	Alopecia	Discovery: 303/880	Discovery: Japanese	*CACNB4*	rs3820706	8.1×10^{-9}	OR[a] = 3.71	16
			Replication: Japanese					
Gemcitabine	Neutropenia	Discovery: 51/28	Discovery: Japanese	*DAPK1*	rs11141915	1.27×10^{-6}	OR = 4.10	14
		Replication: 33/62		*PDE4B*	rs1901440	3.11×10^{-6}	OR = 34.00	
			Replication: Japanese		rs12046844	4.56×10^{-5}	OR = 4.13	
					rs11719165	5.98×10^{-5}	OR = 2.60	
Epirubicin	Neutropenia	Discovery: 67/203	Discovery: Japanese	*MCPH1*	rs2916733	2.27×10^{-9}	OR = 2.74	15
		Replication: 48/0	Replication: Japanese					
Nevirapine	Skin hypersensitivity	Discovery: 72/77	Discovery: Thailand	*CCHCR1*	rs1265112	1.20×10^{-8}	OR = 4.36	17
		Replication: 88/145	Replication: Thailand					
Carbamazepine	Skin hypersensitivity	Discovery: 53/882	Discovery: Japanese	*HLA*	*HLA-A*[a]*3101*	3.64×10^{-15}	OR = 10.8	18
		Replication: 61/376	Replication: Japanese					
Antituberculosis drugs	Liver injury	Discovery: 48/354	Discovery: Ethiopian	*FAM65B*	rs10946737	4.4×10^{-6}	OR = 3.4	19
		Replication: 27/217	Replication: Ethiopian					

[a]*OR,* odds ratio, *HR* hazard ratio

studies which reported the SNPs strongly associated with ADRs [13–19]. The median sample size in screening phase was 402 (range 79–1183), and a median number of cases and controls in screening phase were 67 (range 48–303) and 354 (range 28–882), respectively. The GWAS strategy needs validation phase using independent cohorts because some of the results from screening stage should be false positive. To prove that the significant association results from screening phase are true positive, these validation studies should be carried out using as many as independent cohorts. Here representative results of GWASs to identify the biomarker for ADRs are summarized.

Peripheral Neurotoxicity

Anticancer drug-induced peripheral neurotoxicity is one of the most severe and common adverse reactions especially in cytotoxic anticancer drug therapy [20]. It is known that treatment with anticancer drugs such as platinum drugs (cisplatin, oxaliplatin), taxanes (paclitaxel, docetaxel), vinca alkaloids could cause this toxicity, which is reversible in some kinds of anticancer drugs (taxanes and so on), but irreversible in the other drugs such as platinum drugs [21]. The variation of the phenotype might be partially due to the difference of mechanism of action in each drug. Although the damage to body of neuron in the ganglion and axonal toxicity through transport deficits has been suggested to be one of the mechanisms of action for peripheral neurotoxicity, the detail mechanisms had been unclear [22].

Although the interindividual genetic variation (common and rare variants) has been considered to be involved in peripheral neurotoxicity, genetic loci responsible for this toxicity have been unclear. Since GWAS has been proven as a powerful tool to identify genetic factors which make individuals susceptible to the ADRs, lots of researchers attempted the GWAS to identify a genetic variation(s) which regulate the susceptibility to neurotoxicity. One of the first GWAS of drug-induced neuropathy was reported by Baldwin in 2012 [13]. To identify genetic risk factors for the development of paclitaxel-induced neuropathy, they performed a genome-wide association study using 855 samples of European ancestry, and replication study using additional 154 European and 117 African American samples [13]. As shown in Table 7.1, they identified the single nucleotide polymorphism in *FGD4* which was associated with the risk of peripheral neurotoxicity in the screening cohort (rs10771973, P value of 2.6×10^{-6}, hazard ratio of 1.57). Moreover, in two independent replication cohorts, European and African American subjects successfully showed P value of 0.013, hazard ratio of 1.72 and P value of 6.7×10^{-3}, hazard ratio of 1.93, respectively. The other two loci including *EPHA5* (rs7349683) and *FZD3* (rs10771973) also showed possible association with the onset of paclitaxel-induced peripheral neurotoxicity [13].

Alopecia

Chemotherapy-induced alopecia is one of the most common ADRs which is experienced by thousands of cancer patients every year [23]. Although the treatment of some ADRs has been developed, the treatment for alopecia is still critical issue for the patients treated with anticancer drugs [24]. Chemotherapy-induced alopecia leads lower quality of life and a negative body image in patients with cancer, and it is psychologically difficult for women to manage. Alopecia is induced as one of ADRs of taxanes, alkylating agents, anthracyclines, which are commonly used for the treatment of many cancers. Patients might need to select less effective chemotherapy to avoid the above anticancer-induced alopecia. To identify molecular mechanisms of chemotherapy-induced alopecia and contribute to development of drugs for prevention or treatment of this toxicity, many researchers have reported the pathogenesis and the mechanism of action for this toxicity [25, 26].

Chung et al. first reported the GWAS of chemotherapy-induced alopecia using patients with breast cancer [16]. They used 303 breast cancer cases who developed grade 2 alopecia and 880 controls who did not show alopecia after chemotherapy, and carried out association study between them. rs3820706 in *CACNB4* (calcium channel voltage-dependent subunit beta 4) was identified as an SNP significantly associated with chemotherapy-induced alopecia with P value of 8.13×10^{-9} and odds ratio of 3.71 as shown in Table 7.1. CACNB4 is a member of a beta subunit family of the voltage-dependent calcium channel (VDCC) complex [27]. Calcium ion is reported to function as a messenger in some cellular signal transduction pathways including cell proliferation or apoptosis [28]. Chung et al. speculated that Ca^{2+} involved in the pathogenesis of alopecia as a potassium channel opener, minoxidil, is effective a subset of hair loss patients [16, 29]. They also established the scoring system for prediction of chemotherapy-induced alopecia and found that patients in the highest risk group showed 443 times higher risk of chemotherapy-induced alopecia than the lowest risk group [16].

Neutropenia

There is heterogeneity in the occurrence of the toxicity among patients who are treated with anticancer drugs. Neutropenia, which is one of the most common ADRs of anticancer drugs, could be dose-limiting toxicity, and could prevent the patients from receiving the effective anticancer treatment. Candidate gene approaches previously identified the association of genetic variants in *TPMT* with 6-mercaptopurine-induced myelosuppression in hematopoietic cancer treatment and the association of *UGT1A1* variants with camptothecin-induced neutropenia and diarrhea in cancer treatment, and genetic test of these genes have been recommended for the prediction of severe adverse reactions prior to use of the drugs by US Food and Drug Administration [1, 2, 30]. Accurate genotyping around a million genetic variants is

currently possible by using genome-wide SNP array system. Today, GWAS using clinical samples (normal cells) from patients treated with anticancer drugs could be a promising tool to identify novel genetic marker(s) for the risk of chemotherapy-induced neutropenia, and lots of GWAS of chemotherapy-induced neutropenia have been reported [14, 15, 31, 32].

Gemcitabine-Induced Neutropenia

Gemcitabine, which is a deoxycytidine analogue, is the anticancer drug of the treatment for various types of cancers including pancreatic and non-small-cell lung cancers [33, 34]. Hematological toxicities such as neutropenia and leukopenia are common ADRs of gemcitabine, and these toxicities often limit the effective gemcitabine treatment. The frequency of gemcitabine-induced severe leukopenia/neutropenia was reported to be 13–35% [35, 36]. Although candidate gene approaches to identify the genes associated with the toxicities of gemcitabine have been reported, any genetic variation is not yet used as a biomarker for the risk of gemcitabine-induced leukopenia in clinic [37]. As shown in Table 7.1, Kiyotani et al. conducted a genome-wide association study to identify a genetic variation associated with the risk of gemcitabine-induced leukopenia/neutropenia using 54 cases (grade 3 or more leukopenia/neutropenia) and 120 controls (without any toxicities) [14]. In the GWAS, four loci were identified as possibly associated region with gemcitabine-induced leukopenia/neutropenia (rs11141915 in *DAPK1*, $P_{\text{combined}} = 1.27 \times 10^{-6}$, odds ratio (OR) =4.10; rs1901440, $P_{\text{combined}} = 3.11 \times 10^{-6}$, OR = 34.00; rs12046844 in *PDE4B*, $P_{\text{combined}} = 4.56 \times 10^{-5}$, OR = 4.13; rs11719165, $P_{\text{combined}} = 5.98 \times 10^{-5}$, OR = 2.60) [14]. When they investigated the combined effects of the above four SNPs, significantly higher risks of gemcitabine-induced leukopenia/neutropenia were observed in the patients having 3 risk genotypes ($P = 4.13 \times 10^{-9}$, OR = 50.00) relative to patients with 0 or 1 risk genotype, suggesting the clinical usefulness of the scoring system [14].

Epirubicin-Induced Neutropenia

Epirubicin, an anthracycline cytotoxic agent, forms a complex with DNA by intercalation between base pairs in the nucleus of cell and have cytotoxic activity. Many types of cancers including breast cancer, ovarian cancer, and so on, were treated with this anticancer drug, and neutropenia, which could be dose-limiting toxicities, is one of the most common ADRs for the patients treated with epirubicin [38]. Its frequency is reported to be approximately 42% [38, 39]. As shown in Table 7.1, Srinivasan et al. reported genetic factors affecting the risk of epirubicin-induced leukopenia/neutropenia through the GWAS [15]. They used 270 patients including 67 cases (patients with grade 3 or more leukopenia/neutropenia) and 203 controls

(no toxicity), and further carried out replication study using 48 cases with grade 3 or more epirubicin-induced leukopenia/neutropenia. In their study, rs2916733 in microcephalin 1 showed significant association with epirubicin-induced leukopenia/neutropenia ($P_{combined} = 2.27 \times 10^{-9}$, OR = 2.74), suggesting that the above SNPs could be a genetic marker for the risk of epirubicin-induced neutropenia.

Chemotherapy-Induced Neutropenia

Majority of anticancer drugs especially cytotoxic agents can cause neutropenia/leukopenia as a dose-limiting or life-threatening toxicity. Therefore, clarification of mechanism of the interindividual difference in the risk of ADRs including neutropenia/leukopenia and establishment of prediction system for the risk of ADRs have considered to be important to provide safe and effective chemotherapy to the patients with cancer. It has been thought that there should be common and specific mechanism to cause neutropenia/leukopenia among anticancer drugs, which are the common ADRs after treatment with anticancer drugs [40]. To fully clarify the underlying mechanism and susceptible risk factors that cause neutropenia, Low et al. carried out GWAS using 13,122 cancer patients who had been treated with various drug regimens (cyclophosphamide, platinum, anthracycline, and antimetabolite, antimicrotubule drugs, and topoisomerase inhibitors monotherapy, or combination therapy of them) [32]. Although they could not identify genetic variants which achieve the genome-wide significant level through the GWAS, they showed that weighted genetic risk score (wGRS) analysis could be the possible prediction system for the risk of chemotherapy-induced neutropenia/leukopenia. This GWAS is one of the largest studies for the ADRs in patients treated with anticancer drugs [32].

Skin Hypersensitivity

Skin hypersensitivity is basically dose-independent, unpredictable, and sometimes life-threatening ADRs (type B ADRs) [41]. Most of the drugs have possibility to cause hypersensitivity syndrome. Drug-induced hypersensitivity syndrome (DIHS), which is also described as severe cutaneous adverse drug reactions (cADRs), is characterized by skin rash, fever, and systemic reactions such as hepatitis and so on [42, 43]. Moreover, Stevens-Johnson syndrome (SJS) and toxic epidermal necrolysis (TEN) are also severe hypersensitivities [44]. GWAS have proven to be useful tool for identification of genetic factors of many kinds of ADRs. Although the report for GWASs of skin hypersensitivity is limited, we introduce the representative results of GWASs for this toxicity.

Nevirapine-Induced Skin Hypersensitivity

Nevirapine is a potent nonnucleoside reverse transcriptase inhibitor and one of the first-line drugs for antiretroviral therapy to human immunodeficiency virus type 1 (HIV-1) infection. Nevirapine often causes cADRs with approximate incidence of 15–20% [45–47]. Mild to severe skin reactions including SJS and TEN could be induced by this drug [48]. Chantarangsu et al. reported the first GWAS to identify the genetic variations associated with nevirapine-induced rash using 72 HIV–infected Thai patients with nevirapine-induced rash and 77 controls (without nevirapine-induced toxicity), and as a replication cohort, 88 cases (cADRs positive) and 145 controls were used. The GWAS and replication studies showed that rs1265112 and rs746647 within *CCHCR1* were significantly associated with nevirapine-induced rash ($P_{\text{combined}} = 1.2 \times 10^{-8}$, OR = 4.36) as shown in Table 7.1. They suggested that a predictive model that includes genetic and clinical risk factors for nevirapine-associated rash could be a useful model in reducing the risk of rash induced by nevirapine in HIV-infected patients [17].

Carbamazepine-Induced Skin Hypersensitivity

Carbamazepine (CBZ) is one of the drugs for control of epilepsy [49]. This drug works by reducing abnormal electrical signal in the brain. Pharmacogenomics study of CBZ-induced cADR using Taiwanese population has shown that HLA-B∗1502 was associated with SJS/TEN induced by carbamazepine [50]. This result was confirmed by the studies using populations in the other Southern Asian countries [51, 52]. Although this positive association could be a prediction model of CBZ-induced cADRs in clinic in the above countries, the allelic frequencies of this loci in the other populations are <1% [53]. Hence, HLA-B∗1502 could not be a widely used genetic biomarker for the carbamazepine-induced cADRs in these populations. Ozeki et al. performed the GWAS using 53 cases (the CBZ-induced cADRs including SJS, TEN) and 882 controls in Japanese population [18]. They identified significantly associated SNP, rs1633021 ($P = 1.18 \times 10^{-13}$), which is located in the locus including HLA-A. They further genotyped the HLA-A alleles using 61 cases and 376 controls (no CBZ-induced ADRs) and observed that HLA-A∗3101 was present in 60.7% of the patients with CBZ-induced cADRs; however, it was present in 12.5% of the CBZ-tolerant controls (OR = 10.8, $P = 3.64 \times 10^{-15}$), suggesting that HLA-A alleles could be a useful biomarker for making a decision of individualized treatment of epilepsy [18].

Drug-Induced Liver Injury (DILI)

Drug-induced liver injury (DILI) is one of the common ADRs [54]. Many drugs can induce liver injuries through different mechanisms [55]. The annual incidence of DILI is reported to be about 13.9–24.0 in 100,000 patients [56]. Although there are

few pathognomonic findings in DILI, immune-mediated (known as allergic) and metabolism-mediated mechanisms for this ADR have been suggested [57]. Many association studies between HLA and ADRs were reported, and some HLA types have suggested to be a predictive marker for DILI [58–61]. However, GWAS of DILI is limited partially due to its low frequency of the incidence. Petros et al. reported the GWAS and replication study of antituberculosis drug-induced liver injury in 2016 [19]. To identify the antituberculosis drug-induced liver injury, the authors carried out GWAS using 646 Ethiopian patients receiving rifampicin-based short course antituberculosis therapy. In the first screening phase, they used 48 DILI cases and 354 controls (antituberculosis tolerant) [19]. They further perform replication study for the 50 SNPs showing lowest *P* values using an independent cohort consisting of 27 DILI cases and 217 controls. The top SNP showing lowest *P* value was rs10946737 ($P = 4.4 \times 10^{-6}$, OR = 3.4) in the intron of *FAM65B* in chromosome 6 in the combined analysis (Table 7.1). A cluster of SNPs, which was possibly associated with antituberculosis-induced liver injury, was also observed in the intron of ATP-/GTP-binding protein-like 4 (*AGBL4*) [19].

Moreover, Nicoletti et al. also reported the association of DILI by specific drugs or groups of drugs with HLA type and SNPs in other genes through GWAS [62]. They performed a GWAS using 862 patients with DILI and 10,588 population-matched controls. In the first screening cases, they used 137 cases from European and 274 cases from USA. They found that rs114577328 (A∗33:01 a HLA class I allele) and with rs72631567 on chromosome 2 were significantly associated with DILI (odds ratio of 2.7 and 2.0, *P* value of 2.4×10^{-8} and 9.7×10^{-9}, respectively) [62]. HLA-A∗33:01 was strongly associated with terbinafine-, fenofibrate-, and ticlopidine-induced liver toxicity [62]. They successfully validated the association between A∗33:01 terbinafine- and sertraline-induced liver toxicity. Furthermore, they showed the significant association between rs28521457 (within the *LRBA* gene) and hepatocellular DILI with *P* value of 4.8×10^{-9}, odds ratio of 2.1. The results from the above 2 GWASs of DILI are expected to be useful predictors for DILI in clinical setting.

7.3 GWAS of Drug Efficacy

Many candidate gene approaches had suggested associations between genetic variation and responses to drugs before GWAS has been common [63]. A candidate gene approach had been prevalently used to identify the predictive marker for drug efficacy because this approach could discover the causative genetic variant in well-known genes (drug transporter, metabolic enzyme, and so on) with lower cost for the experiment. Although the efficacy of the drugs considered to be regulated by the genetic background of cancer tissues (somatic mutation, gene expression patterns) and germline variation in host (human), many research groups reported the significant association between germline variation and drug efficacy.

GWAS of Tamoxifen Efficacy

Tamoxifen has been mainly used for the adjuvant therapy for patients with estrogen receptor (ER)-positive breast cancers. It is reported that five-year tamoxifen therapy could improve the risk of its relapse at least for 15 years, particularly for ER-positive invasive tumors in premenopausal women [64]. The results of ATLAS (adjuvant tamoxifen longer against the shorter) trial showed that the risk of recurrence during years 5–14 was more than 20% in the tamoxifen-treated patients in adjuvant setting [65]. The mechanisms underlying the efficacy of this drug in a subset of the patients are not fully clarified. Two representative metabolites of tamoxifen, 4-hydroxytamoxifen and endoxifen (4-hydroxy-N-desmethyltamoxifen), are known to be active therapeutic moieties [66, 67]. These two metabolites have greater affinity to ER and greater potency in inhibiting estrogen-dependent cell growth compared with a parent compound, tamoxifen [66–68]. Therefore, the differences in the formation of these active metabolites considered to affect the interindividual variability in efficacy of tamoxifen.

Cytochrome P450 2D6 (CYP2D6) is one of the well-known enzymes for the generation of the strong active metabolites of tamoxifen, "4-hydroxytamoxifen" and "endoxifen" [69]. As a candidate gene approach in tamoxifen pharmacogenomics, many studies indicated that decreased—or null-function—alleles of *CYP2D6* were associated with poor response to tamoxifen [70–73]. Moreover, results of *CYP2D6* genotype-guided dose-adjustment studies of tamoxifen proved that dose adjustment based on the genotype could realize the personalized tamoxifen therapy [74, 75]. There are several reports claiming the lack of association between *CYP2D6* genotypes and tamoxifen efficacy [76–79]; however, these studies have been criticized due to multiple issues which cause false-negative results, i.e., inappropriate patient population, inappropriate DNA sources, and incomplete genotyping analysis [80].

It is known that some of the patients with homozygous *CYP2D6* wild-type allele, who should have potent CYP2D6 activity, could recur after tamoxifen therapy. Moreover, some of the patients carrying variant alleles (*CYP2D6 Wt/V* or *V/V*), who should have intermediate or weak CYP2D6 activity, do not recur after tamoxifen therapy [81]. Although *CYP2D6* genotype could be associated with tamoxifen efficacy and promising predictive marker for the response to this drug, there should be also the other genetic factors which related to the response to tamoxifen treatment. The genes, such as Cytochrome P450 2C19 (*CYP2C19*), Cytochrome P450 3A5 (*CYP3A5*), sulfotransferase 1A1 (*SULT1A1*), UDP-glucuronosyltransferase 2B15 (*UGT2B15*) and ATP-binding cassette sub-family C member 2 (*ABCC2*), could be possible candidates related to response to tamoxifen therapy [72, 76, 81, 82]; however, associations of these candidate genes have not yet been sufficiently validated.

To fully understand and identify the genetic factors determining individual response to tamoxifen, Kiyotani et al. carried out and reported a genome-wide association study (GWAS) in 2012 [83]. They studied 462 Japanese patients with hormone receptor-positive, invasive breast cancer treated with tamoxifen in adjuvant setting. They observed significant associations with recurrence-free survival at 15 SNPs on 9 chromosomal loci (1p31, 1q41, 5q33, 7p11, 10q22, 12q13, 13q22, 18q12, and 19p13) that

satisfied a genome-wide significant threshold (log-rank $P = 2.87 \times 10^{-9} - 9.41 \times 10^{-8}$) in the GWAS stage. Of the above SNPs, rs10509373 in *C10orf11* gene on 10q22 showed significant association with clinical outcome in two independent replication studies (105 and 107 cases, respectively) and a combined analysis showed a strong association of this SNP with clinical outcome of breast cancer patients treated with tamoxifen (log-rank $P = 1.26 \times 10^{-10}$) [83]. Moreover, in a combined analysis of rs10509373 with *CYP2D6* and *ABCC2*, the number of risk alleles of these genes had cumulative effects on recurrence-free survival among 345 breast cancer patients treated with tamoxifen in adjuvant setting (log-rank $P = 2.28 \times 10^{-12}$), suggesting the clinical usefulness of this prediction system for the response to tamoxifen [83].

7.4 GWAS of Dose Adjustment

GWAS of Warfarin Dose Adjustment

Warfarin is one of the most commonly used anticoagulants for thromboembolic therapy [84]. The interindividual variability in its maintenance dose is known to be large [85, 86]. International normalized ratio (INR) is used to monitor the appropriate (effective but not toxic) dose, and it usually takes about 30–60 days to decide the appropriate maintenance dose by monitoring the INR in each patient. As a result of candidate gene approach for warfarin pharmacogenomics, genetic variations in the *CYP2C9* (cytochrome P450, family 2, subfamily C, polypeptide 9) and *VKORC1* (vitamin K epoxide reductase complex subunit 1) genes are considered to influence warfarin responsiveness because these gene products play essential roles in the pharmacokinetics and pharmacodynamics of warfarin [87–89]. However, it has been suggested that there are the other unknown factors to determine interindividual variability in warfarin dose [90].

Cha et al. carried out the GWAS of warfarin responsiveness and identified rs2108622 in cytochrome P450, family 4, subfamily F, and polypeptide 2 (*CYP4F2*) as a genetic determinant of warfarin responsiveness for Japanese [90]. They incorporate the genotypes of rs2108622 into a warfarin dosing algorithm that they previously had established considering age, body surface area, status of amiodarone coadministration, and genotypes of SNPs in the *CYP2C9* and *VKORC1* genes and found the improvement of the model's predictability to 43.4% [90, 91].

GWAS Between Mercaptopurine Dose and Its Toxicity

Thiopurines such as mercaptopurine (MP), thioguanine, and azathioprine are commonly used anticancer drugs for the treatment of hematopoietic cancer including acute lymphoblastic leukemia (ALL) [92–95]. A subset of patients is known to suffer from mercaptopurine-induced myelosuppression, which could prevent the patients

with ALL from receiving the effective treatment [96–98]. The lack of thiopurine methyltransferase (*TPMT*) resulting from genetic polymorphisms is known to increase the levels of active metabolites of thiopurines and the risk of thiopurine-induced myelosuppression [2]. However, interindividual variation in thiopurine-induced myelosuppression could not be explained by only genetic variations in *TPMT*, and many patients carrying *TPMT* wild type also suffer from myelosuppression [99].

To identify the genetic factors which could be associated with variability in MP tolerance, Yang reported the result of GWAS in two prospective clinical trials of childhood ALL with common chronic MP treatment regimens [100]. They used 657 and 371 patients in discovery GWAS and replication cohorts, respectively, and regarded MP dose intensity during maintenance therapy as a marker of the drug tolerance and toxicities [100]. They observed two significantly associated loci with MP dose intensity: rs1142345 in *TPMT* (Tyr240Cys, present in ∗3A and ∗3C variants; $P = 8.6 \times 10^{-9}$) and rs116855232 in *NUDT15* ($P = 8.8 \times 10^{-9}$). In this study, patients with TT genotype at rs116855232 showed significantly lower MP dose intensity (%), with an average dose intensity of 8.3%, compared with those with TC and CC genotypes, who tolerated 63% and 83.5% of the planned dose, respectively [100]. In the result of their study, of children homozygous for either *TPMT* or *NUDT15* variants or heterozygous for both, 100% required ≥50% MP dose reduction, compared with only 7.7% for other patients, suggesting that these two genetic polymorphisms could be useful predictors for the appropriate maintenance dose of MP [99].

7.5 Conclusion

This chapter summarized the current GWASs of pharmacogenomic, especially studies of adverse drug reactions, drug efficacy, and dose adjustment. Advances in genotyping technologies enabled us to perform GWAS with relatively lower cost than previous, and have accelerated identification of hundreds of candidate genetic markers for drug response. The results of GWASs of pharmacogenomic could identify useful biomarkers for drug response and provide novel insights into pharmacological mechanism which could explain the interindividual difference of drug response. To identify available biomarkers for drug efficacy and/or toxicity in clinic, it is clear that multicenter validation studies for pharmacogenomics are important and essential.

References

1. Innocenti F, Undevia SD, Iyer L, Chen PX, Das S, Kocherginsky M, Karrison T, Janisch L, Ramirez J, Rudin CM, Vokes EE, Ratain MJ (2004) Genetic variants in the UDP-glucuronosyltransferase 1A1 gene predict the risk of severe neutropenia of irinotecan. J Clin Oncol 22:1382–1388
2. Relling MV, Hancock ML, Rivera GK, Sandlund JT, Ribeiro RC, Krynetski EY, Pui CH, Evans WE (1999) Mercaptopurine therapy intolerance and heterozygosity at the thiopurine S-methyltransferase gene locus. J Natl Cancer Inst 91:2001–2008

3. McBride KL, Gilchrist GS, Smithson WA, Weinshilboum RM, Szumlanski CL (2000) Severe 6-thioguanine-induced marrow aplasia in a child with acute lymphoblastic leukemia and inherited thiopurine methyltransferase deficiency. J Pediatr Hematol Oncol 22:441–445
4. Gozalo C, Gerard L, Loiseau P, Morand-Joubert L, Peytavin G, Molina JM, Dellamonica P, Becquemont L, Aboulker JP, Launay O, Verstuyft C (2011) Pharmacogenetics of toxicity, plasma trough concentration and treatment outcome with nevirapine-containing regimen in anti-retroviral-naive HIV-infected adults: an exploratory study of the TRIANON ANRS 081 trial. Basic Clin Pharmacol Toxicol 109:513–520
5. de Boer YS, Kosinski AS, Urban TJ, Zhao Z, Long N, Chalasani N, Kleiner DE, Hoofnagle JH (2017) Features of autoimmune hepatitis in patients with drug-induced liver injury. Clin Gastroenterol Hepatol 15:103–112 e102
6. Low SK, Takahashi A, Mushiroda T, Kubo M (2014) Genome-wide association study: a useful tool to identify common genetic variants associated with drug toxicity and efficacy in cancer pharmacogenomics. Clin Cancer Res 20:2541–2552
7. Chambliss AB, Chan DW (2016) Precision medicine: from pharmacogenomics to pharmacoproteomics. Clin Proteomics 13:25
8. Ferner RE, Aronson JK (2006) Clarification of terminology in medication errors: definitions and classification. Drug Saf 29:1011–1022
9. Aronson JK, Ferner RE (2005) Clarification of terminology in drug safety. Drug Saf 28:851–870
10. Iasella CJ, Johnson HJ, Dunn MA (2017) Adverse drug reactions: Type A (Intrinsic) or Type B (Idiosyncratic). Clin Liver Dis 21:73–87
11. Nelson MR, Bacanu SA, Mosteller M, Li L, Bowman CE, Roses AD, Lai EH, Ehm MG (2009) Genome-wide approaches to identify pharmacogenetic contributions to adverse drug reactions. Pharmacogenomics J 9:23–33
12. Chan SL, Jin S, Loh M, Brunham LR (2015) Progress in understanding the genomic basis for adverse drug reactions: a comprehensive review and focus on the role of ethnicity. Pharmacogenomics 16:1161–1178
13. Baldwin RM, Owzar K, Zembutsu H, Chhibber A, Kubo M, Jiang C, Watson D, Eclov RJ, Mefford J, McLeod HL, Friedman PN, Hudis CA, Winer EP, Jorgenson EM, Witte JS, Shulman LN, Nakamura Y, Ratain MJ, Kroetz DL (2012) A genome-wide association study identifies novel loci for paclitaxel-induced sensory peripheral neuropathy in CALGB 40101. Clin Cancer Res 18:5099–5109
14. Kiyotani K, Uno S, Mushiroda T, Takahashi A, Kubo M, Mitsuhata N, Ina S, Kihara C, Kimura Y, Yamaue H, Hirata K, Nakamura Y, Zembutsu H (2012) A genome-wide association study identifies four genetic markers for hematological toxicities in cancer patients receiving gemcitabine therapy. Pharmacogenet Genomics 22:229–235
15. Srinivasan Y, Sasa M, Honda J, Takahashi A, Uno S, Kamatani N, Kubo M, Nakamura Y, Zembutsu H (2011) Genome-wide association study of epirubicin-induced leukopenia in Japanese patients. Pharmacogenet Genomics 21:552–558
16. Chung S, Low SK, Zembutsu H, Takahashi A, Kubo M, Sasa M, Nakamura Y (2013) A genome-wide association study of chemotherapy-induced alopecia in breast cancer patients. Breast Cancer Res 15:R81
17. Chantarangsu S, Mushiroda T, Mahasirimongkol S, Kiertiburanakul S, Sungkanuparph S, Manosuthi W, Tantisiriwat W, Charoenyingwattana A, Sura T, Takahashi A, Kubo M, Kamatani N, Chantratita W, Nakamura Y (2011) Genome-wide association study identifies variations in 6p21.3 associated with nevirapine-induced rash. Clin Infect Dis 53:341–348
18. Ozeki T, Mushiroda T, Yowang A, Takahashi A, Kubo M, Shirakata Y, Ikezawa Z, Iijima M, Shiohara T, Hashimoto K, Kamatani N, Nakamura Y (2011) Genome-wide association study identifies HLA-A∗3101 allele as a genetic risk factor for carbamazepine-induced cutaneous adverse drug reactions in Japanese population. Hum Mol Genet 20:1034–1041
19. Petros Z, Lee MM, Takahashi A, Zhang Y, Yimer G, Habtewold A, Amogne W, Aderaye G, Schuppe-Koistinen I, Mushiroda T, Makonnen E, Kubo M, Aklillu E (2016) Genome-wide association and replication study of anti-tuberculosis drugs-induced liver toxicity. BMC Genomics 17:755

20. Travis LB, Fossa SD, Sesso HD, Frisina RD, Herrmann DN, Beard CJ, Feldman DR, Pagliaro LC, Miller RC, Vaughn DJ, Einhorn LH, Cox NJ, Dolan ME (2014) Chemotherapy-induced peripheral neurotoxicity and ototoxicity: new paradigms for translational genomics. J Natl Cancer Inst 106
21. Mykletun A, Dahl AA, Haaland CF, Bremnes R, Dahl O, Klepp O, Wist E, Fossa SD (2005) Side effects and cancer-related stress determine quality of life in long-term survivors of testicular cancer. J Clin Oncol 23:3061–3068
22. Argyriou AA, Bruna J, Marmiroli P, Cavaletti G (2012) Chemotherapy-induced peripheral neurotoxicity (CIPN): an update. Crit Rev Oncol Hematol 82:51–77
23. Saini VK, Sewal RK, Ahmad Y, Medhi B (2015) Prospective observational study of adverse drug reactions of anticancer drugs used in cancer treatment in a tertiary care hospital. Indian J Pharm Sci 77:687–693
24. Hesketh PJ, Batchelor D, Golant M, Lyman GH, Rhodes N, Yardley D (2004) Chemotherapy-induced alopecia: psychosocial impact and therapeutic approaches. Support Care Cancer 12:543–549
25. Brosh R, Sarig R, Natan EB, Molchadsky A, Madar S, Bornstein C, Buganim Y, Shapira T, Goldfinger N, Paus R, Rotter V (2010) p53-dependent transcriptional regulation of EDA2R and its involvement in chemotherapy-induced hair loss. FEBS Lett 584:2473–2477
26. Muller-Rover S, Rossiter H, Paus R, Handjiski B, Peters EM, Murphy JE, Mecklenburg L, Kupper TS (2000) Overexpression of Bcl-2 protects from ultraviolet B-induced apoptosis but promotes hair follicle regression and chemotherapy-induced alopecia. Am J Pathol 156:1395–1405
27. Tadmouri A, Kiyonaka S, Barbado M, Rousset M, Fablet K, Sawamura S, Bahembera E, Pernet-Gallay K, Arnoult C, Miki T, Sadoul K, Gory-Faure S, Lambrecht C, Lesage F, Akiyama S, Khochbin S, Baulande S, Janssens V, Andrieux A, Dolmetsch R, Ronjat M, Mori Y, De Waard M (2012) Cacnb4 directly couples electrical activity to gene expression, a process defective in juvenile epilepsy. EMBO J 31:3730–3744
28. Yang H, Zhang Q, He J, Lu W (2010) Regulation of calcium signaling in lung cancer. J Thorac Dis 2:52–56
29. Price VH (1999) Treatment of hair loss. N Engl J Med 341:964–973
30. Iyer L, Das S, Janisch L, Wen M, Ramirez J, Karrison T, Fleming GF, Vokes EE, Schilsky RL, Ratain MJ (2002) UGT1A1∗28 polymorphism as a determinant of irinotecan disposition and toxicity. Pharmacogenomics J 2:43–47
31. Han JY, Shin ES, Lee YS, Ghang HY, Kim SY, Hwang JA, Kim JY, Lee JS (2013) A genome-wide association study for irinotecan-related severe toxicities in patients with advanced non-small-cell lung cancer. Pharmacogenomics J 13:417–422
32. Low SK, Chung S, Takahashi A, Zembutsu H, Mushiroda T, Kubo M, Nakamura Y (2013) Genome-wide association study of chemotherapeutic agent-induced severe neutropenia/leucopenia for patients in Biobank Japan. Cancer Sci 104:1074–1082
33. Cortes-Funes H, Martin C, Abratt R, Lund B (1997) Safety profile of gemcitabine, a novel anticancer agent, in non-small cell lung cancer. Anti-Cancer Drugs 8:582–587
34. Heinemann V, Wilke H, Mergenthaler HG, Clemens M, Konig H, Illiger HJ, Arning M, Schalhorn A, Possinger K, Fink U (2000) Gemcitabine and cisplatin in the treatment of advanced or metastatic pancreatic cancer. Ann Oncol 11:1399–1403
35. Tanaka T, Ikeda M, Okusaka T, Ueno H, Morizane C, Hagihara A, Iwasa S, Kojima Y (2008) Prognostic factors in japanese patients with advanced pancreatic cancer treated with single-agent gemcitabine as first-line therapy. Jpn J Clin Oncol 38:755–761
36. Lee JO, Kim DY, Lim JH, Seo MD, Yi HG, Oh DY, Im SA, Kim TY, Bang YJ (2009) Palliative chemotherapy for patients with recurrent hepatocellular carcinoma after liver transplantation. J Gastroenterol Hepatol 24:800–805
37. Sugiyama E, Kaniwa N, Kim SR, Kikura-Hanajiri R, Hasegawa R, Maekawa K, Saito Y, Ozawa S, Sawada J, Kamatani N, Furuse J, Ishii H, Yoshida T, Ueno H, Okusaka T, Saijo N

(2007) Pharmacokinetics of gemcitabine in Japanese cancer patients: the impact of a cytidine deaminase polymorphism. J Clin Oncol 25:32–42
38. Therasse P, Mauriac L, Welnicka-Jaskiewicz M, Bruning P, Cufer T, Bonnefoi H, Tomiak E, Pritchard KI, Hamilton A, Piccart MJ (2003) Final results of a randomized phase III trial comparing cyclophosphamide, epirubicin, and fluorouracil with a dose-intensified epirubicin and cyclophosphamide + filgrastim as neoadjuvant treatment in locally advanced breast cancer: an EORTC-NCIC-SAKK multicenter study. J Clin Oncol 21:843–850
39. Case DC Jr, Gams R, Ervin TJ, Boyd MA, Oldham FB (1987) Phase I-II trial of high-dose epirubicin in patients with lymphoma. Cancer Res 47:6393–6396
40. Crawford J, Dale DC, Lyman GH (2004) Chemotherapy-induced neutropenia: risks, consequences, and new directions for its management. Cancer 100:228–237
41. Uetrecht J, Naisbitt DJ (2013) Idiosyncratic adverse drug reactions: current concepts. Pharmacol Rev 65:779–808
42. Hashimoto K, Yasukawa M, Tohyama M (2003) Human herpesvirus 6 and drug allergy. Curr Opin Allergy Clin Immunol 3:255–260
43. Shiohara T, Kano Y (2007) A complex interaction between drug allergy and viral infection. Clin Rev Allergy Immunol 33:124–133
44. Harr T, French LE (2012) Stevens-Johnson syndrome and toxic epidermal necrolysis. Chem Immunol Allergy 97:149–166
45. Pollard RB, Robinson P, Dransfield K (1998) Safety profile of nevirapine, a nonnucleoside reverse transcriptase inhibitor for the treatment of human immunodeficiency virus infection. Clin Ther 20:1071–1092
46. Carr A, Cooper DA (2000) Adverse effects of antiretroviral therapy. Lancet 356:1423–1430
47. de Maat MM, ter Heine R, Mulder JW, Meenhorst PL, Mairuhu AT, van Gorp EC, Huitema AD, Beijnen JH (2003) Incidence and risk factors for nevirapine-associated rash. Eur J Clin Pharmacol 59:457–462
48. Metry DW, Lahart CJ, Farmer KL, Hebert AA (2001) Stevens-Johnson syndrome caused by the antiretroviral drug nevirapine. J Am Acad Dermatol 44:354–357
49. Marson AG, Williamson PR, Hutton JL, Clough HE, Chadwick DW (2000) Carbamazepine versus valproate monotherapy for epilepsy. Cochrane Database Syst Rev:CD001030
50. Chung WH, Hung SI, Hong HS, Hsih MS, Yang LC, Ho HC, Wu JY, Chen YT (2004) Medical genetics: a marker for Stevens-Johnson syndrome. Nature 428:486
51. Lonjou C, Thomas L, Borot N, Ledger N, de Toma C, LeLouet H, Graf E, Schumacher M, Hovnanian A, Mockenhaupt M, Roujeau JC (2006) A marker for Stevens-Johnson syndrome ...: ethnicity matters. Pharmacogenomics J 6:265–268
52. Man CB, Kwan P, Baum L, Yu E, Lau KM, Cheng AS, Ng MH (2007) Association between HLA-B∗1502 allele and antiepileptic drug-induced cutaneous reactions in Han Chinese. Epilepsia 48:1015–1018
53. Tangamornsuksan W, Chaiyakunapruk N, Somkrua R, Lohitnavy M, Tassaneeyakul W (2013) Relationship between the HLA-B∗1502 allele and carbamazepine-induced Stevens-Johnson syndrome and toxic epidermal necrolysis: a systematic review and meta-analysis. JAMA Dermatol 149:1025–1032
54. Ocete-Hita E, Salmeron-Fernandez MJ, Urrutia-Maldonado E, de Rueda PM, Salmeron-Ruiz M, Martinez-Padilla MC, Ruiz-Extremera A (2016) Analysis of Immunogenetic factors in idiosyncratic drug-induced liver injury in the paediatric population. J Pediatr Gastroenterol Nutr
55. Verma S, Kaplowitz N (2009) Diagnosis, management and prevention of drug-induced liver injury. Gut 58:1555–1564
56. Suk KT, Kim DJ (2012) Drug-induced liver injury: present and future. Clin Mol Hepatol 18:249–257
57. Mushiroda T, Yanai H, Yoshiyama T, Sasaki Y, Okumura M, Ogata H, Tokunaga K (2016) Development of a prediction system for anti-tuberculosis drug-induced liver injury in Japanese patients. Hum Genome Var 3:16014

58. Nicoletti P, Werk AN, Sawle A, Shen Y, Urban TJ, Coulthard SA, Bjornsson ES, Cascorbi I, Floratos A, Stammschulte T, Gundert-Remy U, Nelson MR, Aithal GP, Daly AK (2016) HLA-DRB1∗16: 01-DQB1∗05: 02 is a novel genetic risk factor for flupirtine-induced liver injury. Pharmacogenet Genomics 26:218–224
59. Chen R, Zhang Y, Tang S, Lv X, Wu S, Sun F, Xia Y, Zhan SY (2015) The association between HLA-DQB1 polymorphism and antituberculosis drug-induced liver injury: a case-control study. J Clin Pharm Ther 40:110–115
60. Stephens C, Lopez-Nevot MA, Ruiz-Cabello F, Ulzurrun E, Soriano G, Romero-Gomez M, Moreno-Casares A, Lucena MI, Andrade RJ (2013) HLA alleles influence the clinical signature of amoxicillin-clavulanate hepatotoxicity. PLoS One 8:e68111
61. Daly AK, Donaldson PT, Bhatnagar P, Shen Y, Pe'er I, Floratos A, Daly MJ, Goldstein DB, John S, Nelson MR, Graham J, Park BK, Dillon JF, Bernal W, Cordell HJ, Pirmohamed M, Aithal GP, Day CP (2009) HLA-B∗5701 genotype is a major determinant of drug-induced liver injury due to flucloxacillin. Nat Genet 41:816–819
62. Nicoletti P, Aithal GP, Bjornsson ES, Andrade RJ, Sawle A, Arrese M, Barnhart HX, Bondon-Guitton E, Hayashi PH, Bessone F, Carvajal A, Cascorbi I, Cirulli ET, Chalasani N, Conforti A, Coulthard SA, Daly MJ, Day CP, Dillon JF, Fontana RJ, Grove JI, Hallberg P, Hernandez N, Ibanez L, Kullak-Ublick GA, Laitinen T, Larrey D, Lucena MI, Maitland-van der Zee AH, Martin JH, Molokhia M, Pirmohamed M, Powell EE, Qin S, Serrano J, Stephens C, Stolz A, Wadelius M, Watkins PB, Floratos A, Shen Y, Nelson MR, Urban TJ, Daly AK (2016) Association of Liver Injury From Specific Drugs, or Groups of Drugs, With Polymorphisms in HLA and Other Genes in a Genome-wide Association Study. Gastroenterology 152:1078
63. Harper AR, Topol EJ (2012) Pharmacogenomics in clinical practice and drug development. Nat Biotechnol 30:1249
64. Davies C, Godwin J, Gray R, Clarke M, Cutter D, Darby S, McGale P, Pan HC, Taylor C, Wang YC, Dowsett M, Ingle J, Peto R (2011) Relevance of breast cancer hormone receptors and other factors to the efficacy of adjuvant tamoxifen: patient-level meta-analysis of randomised trials. Lancet 378:771–784
65. Davies C, Pan H, Godwin J, Gray R, Arriagada R, Raina V, Abraham M, Medeiros Alencar VH, Badran A, Bonfill X, Bradbury J, Clarke M, Collins R, Davis SR, Delmestri A, Forbes JF, Haddad P, Hou MF, Inbar M, Khaled H, Kielanowska J, Kwan WH, Mathew BS, Mittra I, Muller B, Nicolucci A, Peralta O, Pernas F, Petruzelka L, Pienkowski T, Radhika R, Rajan B, Rubach MT, Tort S, Urrutia G, Valentini M, Wang Y, Peto R (2013) Long-term effects of continuing adjuvant tamoxifen to 10 years versus stopping at 5 years after diagnosis of oestrogen receptor-positive breast cancer: ATLAS, a randomised trial. Lancet 381:805–816
66. Borgna JL, Rochefort H (1981) Hydroxylated metabolites of tamoxifen are formed in vivo and bound to estrogen receptor in target tissues. J Biol Chem 256:859–868
67. Lien EA, Solheim E, Lea OA, Lundgren S, Kvinnsland S, Ueland PM (1989) Distribution of 4-hydroxy-N-desmethyltamoxifen and other tamoxifen metabolites in human biological fluids during tamoxifen treatment. Cancer Res 49:2175–2183
68. Johnson MD, Zuo H, Lee KH, Trebley JP, Rae JM, Weatherman RV, Desta Z, Flockhart DA, Skaar TC (2004) Pharmacological characterization of 4-hydroxy-N-desmethyl tamoxifen, a novel active metabolite of tamoxifen. Breast Cancer Res Treat 85:151–159
69. Desta Z, Ward BA, Soukhova NV, Flockhart DA (2004) Comprehensive evaluation of tamoxifen sequential biotransformation by the human cytochrome P450 system in vitro: prominent roles for CYP3A and CYP2D6. J Pharmacol Exp Ther 310:1062–1075
70. Goetz MP, Knox SK, Suman VJ, Rae JM, Safgren SL, Ames MM, Visscher DW, Reynolds C, Couch FJ, Lingle WL, Weinshilboum RM, Fritcher EG, Nibbe AM, Desta Z, Nguyen A, Flockhart DA, Perez EA, Ingle JN (2007) The impact of cytochrome P450 2D6 metabolism in women receiving adjuvant tamoxifen. Breast Cancer Res Treat 101:113–121

71. Schroth W, Goetz MP, Hamann U, Fasching PA, Schmidt M, Winter S, Fritz P, Simon W, Suman VJ, Ames MM, Safgren SL, Kuffel MJ, Ulmer HU, Bolander J, Strick R, Beckmann MW, Koelbl H, Weinshilboum RM, Ingle JN, Eichelbaum M, Schwab M, Brauch H (2009) Association between CYP2D6 polymorphisms and outcomes among women with early stage breast cancer treated with tamoxifen. JAMA 302:1429–1436
72. Schroth W, Antoniadou L, Fritz P, Schwab M, Muerdter T, Zanger UM, Simon W, Eichelbaum M, Brauch H (2007) Breast cancer treatment outcome with adjuvant tamoxifen relative to patient CYP2D6 and CYP2C19 genotypes. J Clin Oncol 25:5187–5193
73. Ramon y Cajal T, Altes A, Pare L, del Rio E, Alonso C, Barnadas A, Baiget M (2010) Impact of CYP2D6 polymorphisms in tamoxifen adjuvant breast cancer treatment. Breast Cancer Res Treat 119:33–38
74. Kiyotani K, Mushiroda T, Imamura CK, Tanigawara Y, Hosono N, Kubo M, Sasa M, Nakamura Y, Zembutsu H (2012) Dose-adjustment study of tamoxifen based on CYP2D6 genotypes in Japanese breast cancer patients. Breast Cancer Res Treat 131:137–145
75. Irvin WJ Jr, Walko CM, Weck KE, Ibrahim JG, Chiu WK, Dees EC, Moore SG, Olajide OA, Graham ML, Canale ST, Raab RE, Corso SW, Peppercorn JM, Anderson SM, Friedman KJ, Ogburn ET, Desta Z, Flockhart DA, McLeod HL, Evans JP, Carey LA (2011) Genotype-guided tamoxifen dosing increases active metabolite exposure in women with reduced CYP2D6 metabolism: a multicenter study. J Clin Oncol 29:3232–3239
76. Wegman P, Elingarami S, Carstensen J, Stal O, Nordenskjold B, Wingren S (2007) Genetic variants of CYP3A5, CYP2D6, SULT1A1, UGT2B15 and tamoxifen response in postmenopausal patients with breast cancer. Breast Cancer Res 9:R7
77. Wegman P, Vainikka L, Stal O, Nordenskjold B, Skoog L, Rutqvist LE, Wingren S (2005) Genotype of metabolic enzymes and the benefit of tamoxifen in postmenopausal breast cancer patients. Breast Cancer Res 7:R284–R290
78. Abraham JE, Maranian MJ, Driver KE, Platte R, Kalmyrzaev B, Baynes C, Luccarini C, Shah M, Ingle S, Greenberg D, Earl HM, Dunning AM, Pharoah PD, Caldas C (2010) CYP2D6 gene variants: association with breast cancer specific survival in a cohort of breast cancer patients from the United Kingdom treated with adjuvant tamoxifen. Breast Cancer Res 12:R64
79. Regan MM, Leyland-Jones B, Bouzyk M, Pagani O, Tang W, Kammler R, Dell'orto P, Biasi MO, Thurlimann B, Lyng MB, Ditzel HJ, Neven P, Debled M, Maibach R, Price KN, Gelber RD, Coates AS, Goldhirsch A, Rae JM, Viale G (2012) CYP2D6 genotype and tamoxifen response in postmenopausal women with endocrine-responsive breast cancer: the breast international group 1-98 trial. J Natl Cancer Inst 104:441–451
80. Kiyotani K, Mushiroda T, Zembutsu H, Nakamura Y (2013) Important and critical scientific aspects in pharmacogenomics analysis: lessons from controversial results of tamoxifen and CYP2D6 studies. J Hum Genet 58:327–333
81. Kiyotani K, Mushiroda T, Imamura CK, Hosono N, Tsunoda T, Kubo M, Tanigawara Y, Flockhart DA, Desta Z, Skaar TC, Aki F, Hirata K, Takatsuka Y, Okazaki M, Ohsumi S, Yamakawa T, Sasa M, Nakamura Y, Zembutsu H (2010) Significant effect of polymorphisms in CYP2D6 and ABCC2 on clinical outcomes of adjuvant tamoxifen therapy for breast cancer patients. J Clin Oncol 28:1287–1293
82. Gjerde J, Geisler J, Lundgren S, Ekse D, Varhaug JE, Mellgren G, Steen VM, Lien EA (2010) Associations between tamoxifen, estrogens, and FSH serum levels during steady state tamoxifen treatment of postmenopausal women with breast cancer. BMC Cancer 10:313
83. Kiyotani K, Mushiroda T, Tsunoda T, Morizono T, Hosono N, Kubo M, Tanigawara Y, Imamura CK, Flockhart DA, Aki F, Hirata K, Takatsuka Y, Okazaki M, Ohsumi S, Yamakawa T, Sasa M, Nakamura Y, Zembutsu H (2012) A genome-wide association study identifies locus at 10q22 associated with clinical outcomes of adjuvant tamoxifen therapy for breast cancer patients in Japanese. Hum Mol Genet 21:1665–1672

84. Johnson JA (2008) Warfarin: an old drug but still interesting. Pharmacotherapy 28:1081–1083
85. Lesko LJ (2008) The critical path of warfarin dosing: finding an optimal dosing strategy using pharmacogenetics. Clin Pharmacol Ther 84:301–303
86. Siguret V, Pautas E, Gouin-Thibault I (2008) Warfarin therapy: influence of pharmacogenetic and environmental factors on the anticoagulant response to warfarin. Vitam Horm 78:247–264
87. Takahashi H, Wilkinson GR, Nutescu EA, Morita T, Ritchie MD, Scordo MG, Pengo V, Barban M, Padrini R, Ieiri I, Otsubo K, Kashima T, Kimura S, Kijima S, Echizen H (2006) Different contributions of polymorphisms in VKORC1 and CYP2C9 to intra- and inter-population differences in maintenance dose of warfarin in Japanese, Caucasians and African-Americans. Pharmacogenet Genomics 16:101–110
88. Sconce EA, Khan TI, Wynne HA, Avery P, Monkhouse L, King BP, Wood P, Kesteven P, Daly AK, Kamali F (2005) The impact of CYP2C9 and VKORC1 genetic polymorphism and patient characteristics upon warfarin dose requirements: proposal for a new dosing regimen. Blood 106:2329–2333
89. Wadelius M, Chen LY, Downes K, Ghori J, Hunt S, Eriksson N, Wallerman O, Melhus H, Wadelius C, Bentley D, Deloukas P (2005) Common VKORC1 and GGCX polymorphisms associated with warfarin dose. Pharmacogenomics J 5:262–270
90. Cha PC, Mushiroda T, Takahashi A, Kubo M, Minami S, Kamatani N, Nakamura Y (2010) Genome-wide association study identifies genetic determinants of warfarin responsiveness for Japanese. Hum Mol Genet 19:4735–4744
91. Mushiroda T, Ohnishi Y, Saito S, Takahashi A, Kikuchi Y, Saito S, Shimomura H, Wanibuchi Y, Suzuki T, Kamatani N, Nakamura Y (2006) Association of VKORC1 and CYP2C9 polymorphisms with warfarin dose requirements in Japanese patients. J Hum Genet 51:249–253
92. Karran P, Attard N (2008) Thiopurines in current medical practice: molecular mechanisms and contributions to therapy-related cancer. Nat Rev Cancer 8:24–36
93. Pui CH, Carroll WL, Meshinchi S, Arceci RJ (2011) Biology, risk stratification, and therapy of pediatric acute leukemias: an update. J Clin Oncol 29:551–565
94. Vora A, Goulden N, Wade R, Mitchell C, Hancock J, Hough R, Rowntree C, Richards S (2013) Treatment reduction for children and young adults with low-risk acute lymphoblastic leukaemia defined by minimal residual disease (UKALL 2003): a randomised controlled trial. Lancet Oncol 14:199–209
95. Moriyama T, Nishii R, Perez-Andreu V, Yang W, Klussmann FA, Zhao X, Lin TN, Hoshitsuki K, Nersting J, Kihira K, Hofmann U, Komada Y, Kato M, McCorkle R, Li L, Koh K, Najera CR, Kham SK, Isobe T, Chen Z, Chiew EK, Bhojwani D, Jeffries C, Lu Y, Schwab M, Inaba H, Pui CH, Relling MV, Manabe A, Hori H, Schmiegelow K, Yeoh AE, Evans WE, Yang JJ (2016) NUDT15 polymorphisms alter thiopurine metabolism and hematopoietic toxicity. Nat Genet 48:367–373
96. Teml A, Schaeffeler E, Herrlinger KR, Klotz U, Schwab M (2007) Thiopurine treatment in inflammatory bowel disease: clinical pharmacology and implication of pharmacogenetically guided dosing. Clin Pharmacokinet 46:187–208
97. Candy S, Wright J, Gerber M, Adams G, Gerig M, Goodman R (1995) A controlled double blind study of azathioprine in the management of Crohn's disease. Gut 37:674–678
98. Fraser AG, Orchard TR, Jewell DP (2002) The efficacy of azathioprine for the treatment of inflammatory bowel disease: a 30 year review. Gut 50:485–489
99. Yang SK, Hong M, Baek J, Choi H, Zhao W, Jung Y, Haritunians T, Ye BD, Kim KJ, Park SH, Park SK, Yang DH, Dubinsky M, Lee I, McGovern DP, Liu J, Song K (2014) A common missense variant in NUDT15 confers susceptibility to thiopurine-induced leukopenia. Nat Genet 46:1017–1020
100. Yang JJ, Landier W, Yang W, Liu C, Hageman L, Cheng C, Pei D, Chen Y, Crews KR, Kornegay N, Wong FL, Evans WE, Pui CH, Bhatia S, Relling MV (2015) Inherited NUDT15 variant is a genetic determinant of mercaptopurine intolerance in children with acute lymphoblastic leukemia. J Clin Oncol 33:1235–1242

Chapter 8
The Future of and Beyond GWAS

Tatsuhiko Tsunoda

Abstract Although GWAS technologies themselves have become mature, there are still many issues to be solved. One such issue is the missing heritability problem. It is still unknown whether it is sufficient to base the genetic architecture, which is required when attempting to fully explain the heritability, on common markers, or if rare markers, markers other than SNVs, or interactions between the markers must be considered. This may depend on the specific disease types and traits. Simulation methods to estimate the heritability with hypothetical markers have found that the top few thousand markers may explain much of the heritability. However, because of the statistical power issue, whether this is valid will be unclear until the sample size is sufficiently large. Therefore, international meta-analyses to increase power have become popular. Another direction to advance GWAS is to consider molecules other than the genome, which is expected to approach the mechanism of disease with the GWAS results: genomic annotation with omic data, integrated association analysis with multiomics and transomics, in particular expression quantitative loci (eQTL), will be harnessed with GWAS data to focus on disease related genes and markers, and to identify correlation and even causality of the relationships between molecules and diseases. These must be based on different networks of cell types interacting with the environment. Disease phenotype itself could also be considered. These have a complex relationship with each other and cannot be categorized clearly. Rather, such relationships may be used effectively for GWAS and further analyses. Methodological advancement will be needed to solve these complex relationships and dynamics. GWAS applications include drug target discovery and precision medicine – personalized medicine and prevention. To properly achieve these, we need new mathematical methodologies. It is expected that data sharing and

T. Tsunoda (✉)
Tsunoda Laboratory (Medical Science Mathematics), Department of Biological Sciences, Graduate School of Science, The University of Tokyo, Tokyo, Japan

Department of Medical Science Mathematics, Medical Research Institute, Tokyo Medical and Dental University, Tokyo, Japan

Laboratory for Medical Science Mathematics, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan
e-mail: tsunoda@bs.s.u-tokyo.ac.jp

T. Tsunoda et al. (eds.), *Genome-Wide Association Studies*,
https://doi.org/10.1007/978-981-13-8177-5_8

utilization of molecular databases will be promoted, and a next generation of mathematical models and methods based on AI will be developed.

Keywords Missing heritability · Rare variants · Interaction · Structural variation · Omic analysis · eQTL · Disease phenotypes · Precision medicine · Artificial intelligence

8.1 Current Issues to Be Solved

Missing Heritability Problem and Common vs. Rare Variants

GWAS researchers have been facing the missing heritability problem – the accumulation of GWAS results cannot explain much of the observed disease heritability. They are considering both its cause and solutions, however, results have not yet been promising. It is frequently discussed whether heritability could be finally well captured through GWAS common SNPs, or will it be necessary to look at other factors, in particular rare SNVs.

Although it is under discussion how much GWAS common SNPs can explain heritability, several simulations suggest that, if potential ones are included, they will cover 50–100%. For example, simulations with polygenic models have shown that using all common variants, irrespective of significance, is able to explain almost half of type II diabetes (T2D) heritability. Soon, thousands of these smaller effect associations will likely be identified, in addition to those already identified [1]. Similarly in height, additional common SNPs are likely to explain more of the missing heritability than can now be explained: common SNPs and low frequency/rare causal variants will both explain 50% of the heritability [2, 3]. Using a simulation with the Approximate Bayesian Computation method, Stahl et al. estimated that about 65–100% of heritability of RA, celiac disease, MI/CAD, and T2D can be explained by thousands of GWAS common SNPs [4]. Recently, a more accurate method, which considers minor allele frequencies of and linkage disequilibrium between SNPs, for estimating disease SNPs' heritability was proposed [5]. Reevaluating the GWAS results using this method resulted in most heritability being explained, much more than before. There is an excellent review of SNP-based heritability methods and interpretations [6].

In the analyses of real data, although many common SNPs have been identified as disease associated markers, finding associations between rare SNVs and other markers is much less common; their contribution is still unknown. In T2D, there has not been any strong evidence that rare variants are associated with the disease [1]. For example, an exome sequencing study with 2000 Danes (1000 cases and 1000 controls) could not find any low frequency/rare variants associated with T2D. Therefore, at the very least, there is no support for an extreme model in which T2D can be explained by low frequency non-synonymous variants of large effects. In 2014, a genome sequencing study of 2630 Icelandic T2D patients revealed three

new T2D associated variants with low allele frequencies, however, all were within an already reported T2D or T2D-related loci. Also, using a candidate gene approach, various frequency alleles associated to T2D were found within genes that had been previously suggested to have common variants associated with T2D. However, the contribution of these variants was small. These results support the model that low frequency variation and coding region variation contribute, at most, only limitedly to T2D and common variants are the dominant genetic factors of T2D. A recent T2D GWAS, which explored low frequency/rare variants using datasets of 2657 whole genome sequencing (WGS), 12,940 whole exome sequencing (WES), and 111,548 GWAS with imputation, found associations to only already known GWAS loci, and the hypothesis that low frequency variants dominate was not supported [7]. T2D loci, particularly common ones, will continue to increase with GWAS using more than a hundred thousand samples, and much larger sample sets will be necessary to find rare variants associated with T2D [1]. In the latest study of height, 83 height-related coding variants (nonsynonymous/splice-site variants with 0.1%<MAF<4.8%) were found using the 241,453 SNV exome chip data (83% with MAF <5%) of 711,428 individuals [8]. Three additional loci were identified by applying a gene-based test method to the dataset. These results suggest that it is worth examining low frequency (0.5–5%) variants using an imputation technique, which is more efficient than resequencing, even when rare variants are targeted [9]. To accelerate such studies, much more accurate imputation, reference panels with many more markers, WGS, and larger sample sets, are necessary [1, 10]. In another recent study, a large-scale genome analysis of schizophrenia (an exome study with 4133 cases and 9274 controls, a de novo mutation study with 1077 trios, a CNV study with 6882 cases and 11,255 controls) was conducted [11]. As a result, it was found that rare damaging variants contribute to the disease, and that there is genetic overlap with neurodevelopmental disorders. Another expectation is that rare variants associated with disease could be captured by distal common variants that have common haplotypes with the rare variants under linkage disequilibrium, which is called synthetic association [12]. However, as far as the loci of autoimmune disease are concerned, synthetic association is unlikely, and the influence of rare coding variants on heritability appears to be small [13]. These findings support the model that accumulation of weak effects by common variants can cause disease. Recently, expanding on such GWAS, the UK10K consortium has been looking into the influence of rare and low frequency variants to various traits (lipid, adiponectin, etc.) and found that they are extremely small [14].

These results show that even with large sample sets, detection of rare/low-frequency variants that are associated with diseases/traits is challenging. However, it should be considered that these studies evaluated just association of single markers with disease using the case-control design [15]. Many types of methodologies, e.g. SKAT, that account for accumulation of variants within gene have been proposed. Still, not many significant results have been reported, and what amount the accumulation of rare variants significantly contributes to disease is unknown [16]. An Alzheimer disease study reported that gene-based tests had much better performance than single-variant analysis [17]. Conversely, in a blood lipid study, single variants showed much stronger association compared to the gene-level tests [18].

Interaction and Haplotypic Effect

One factor being considered that may explain the remaining heritability is interactions between GWAS loci/alleles. However, the power of interaction detection is, by most methods, insufficient. The number of marker combinations is huge, e.g. $10^6 \times 10^6/2 = 5 \times 10^{11}$, when we want to comprehensively identify the statistical interactions that may show synergistic effects to disease risk by combinations. When we apply Bonferroni correction, one of the most standard methods for multiple comparison, the significance level becomes very stringent: alpha = $0.05/(10^6 \times 10^6/2) = 10^{-13}$. At such a stringent significance level, huge sample sizes are required and current datasets are too small to give sufficient power. In addition, there is the issue that biological interactions are not exactly the same as statistical interactions, and need biological interpretation after their detection [19]. Although one may think haplotypic effects would be better to consider, they are much harder to detect because an additional degree of freedom is statistically necessary when we analyze with logistic regression [20]. Furthermore, it is very difficult to determine the regions (units) to define the haplotypes for markers to be compared between cases and controls. Another type of interaction would be the non-additive (dominance) effect between diploid alleles within each marker. Precise investigation into HLA markers have revealed the existence of interactions between different HLA alleles and between different amino acid products for disease susceptibility [21]. However, this kind of analysis is limited to datasets with large sample size, common alleles, and HLA alleles within each locus. Due to these restrictions, comprehensive interaction analysis has not progressed as much as we initially expected. Rather, recently, the risk of autism has been found to be additive, indicating that different genes and pathways contribute independently to it [22]. Also, a method of using SNP data to partition and estimate the proportion of phenotypic variance contributed by additive and dominance genetic variation at all SNPs was developed and applied to 79 quantitative traits in 6715 unrelated European Americans. This result suggests that the dominance variation contributes little to the missing heritability [23].

Copy Number Variation, Structural Variation, and Other Markers

Copy number variation (CNV) and structural variation (SV) have not been fully explored for disease. They can be strong candidates as disease markers because they may directly affect overlapping genes: changing gene expression and/or gene function. The relationship between CNV/SV and disease has been shown in psychiatric diseases. Recently, a schizophrenia study using GWAS chips with about 200,000 cases and controls identified more than eight CNV loci and four pathways related to the disease [24]. However, CNV/SV studies have so been less successful

than SNP GWAS because of a lack of databases, difficulty in their definition as markers, and their low frequency spectrum. Also, the differences of genomic DNA structures may indirectly influence transcription regulation of proxy and/or distal genes through three-dimensional alterations of chromatin conformation. Retrotransposons, which can jump into and sometimes move in the human genome, may have a similar effect. For example, non-coding RNA transcribed from Alu sequences might be a cause of diseases. Retrotransposons can affect germline as well as somatic genomes, and it has recently been suggested that C4 transposable element may be involved in schizophrenia occurrence [25, 26]. To promote this kind of analyses, we need richer genomic annotations, including three-dimensional structure of DNA, for example.

Attempts to Enlarge Study Size

Irrespective of common and rare variants, or the type of marker, current studies are small yet and do not have the power to find disease etiologies fully explaining the observed heritability. Researchers have given much effort to enlarge study sample sizes. Currently, the main approach is international meta-analyses, i.e. collecting and combining existing GWAS summaries worldwide accompanied with imputation techniques to use different GWAS chips simultaneously and to explore low-frequency SNVs. Another approach would be borrowing samples from other projects for making large control sets. For example, Exome Aggregation Consortium (ExAC) have collected various kinds of exome data, which would be used as control for association studies comparing variant accumulations between cases and controls mostly with exome sequencing. Also, although sample collections in GWAS have been conducted after designing research plans, future research will fully utilize electric medical records (EMR) and electric health records (EHR) available from patients in hospitals and medical institutions, simultaneously asking the patients to provide blood, etc. with written informed consent for collecting omic, particularly genomic, profiles. For example, recently, the EHR of 100,000 people from a GERA cohort, one of Kaiser RPGEH, were used for a blood pressure study [27]. They looked at association between blood pressure data obtained in a time-series and genomic variation, and identified 75 loci (of which 39 were novel). The results were validated with ICBP and UK BioBank cohorts. In addition, by combining the three studies, 241 additional loci were revealed as candidates. This study shows the advantages of multiple institutional EHR-based genomic cohorts. That is, the sample size is large, and the averaged data measured many times over a long period can be obtained. This study also showed that non-strictly controlled EHR can be used for medical studies. Such large-scale genome cohort studies are now occurring at a national level. The US government established the Precision Medicine Initiative: a prospective cohort to collect genomic sequence, clinical, and lifestyle data from one million people [28]. Data collection includes the human genome, cell-free DNA,

proteome, metabolome, biochemical data, as well as personal activity records such as social networking and procurement records of OTC drugs, which reflects the situation that citizens have becoming more conscious of their health. UK Biobank and INTERVAL are other examples reflecting these [29–31]. In addition, private companies have recently promoted this area very strategically. One example of a private enterprise's entry into academia disease research is 23andMe. By collecting clients' genomic data, which were originally sequenced for commercial interpretation of their genomic variation, and reusing it for GWAS, they have identified new disease related genes.

Although data sharing has recently been done to expand GWAS, privacy protection is a major issue [32]. For this, summary statistics analysis would be the simplest and easiest way, and depending on the aim, several categories exist: (a) single-variant association tests with meta-analysis, conditional analysis, and imputation using summary statistics, (b) gene-based association tests by accumulating signals across multiple rare variants or utilizing transcriptome data, (c) fine-mapping causal variants with the help of functional annotation and/or trans-ethnic data, (d) polygenic predictions of disease risk, and (e) joint analysis of multiple traits. Recently, advanced methods to analyze GWAS data with summary statistics has been proposed as an efficient method for conducting meta-analysis internationally [33]. Development of such methodologies will be much more important in the future.

Extreme Phenotype and Population Specificity

Looking at disease phenotype more deeply would be one of the methodologies for making GWAS more efficient. For example, limiting samples to only those with extreme phenotypes, e.g. severe and/or early onset[12)], would achieve almost the same results with much lower cost compared to studies with all samples [15]. Another aspect is that disease is often heterogeneous. Deep phenotyping of patients may help capturing the heterogeneity, which can be used for stratification of study samples [1]. Researchers try to find variants that increase disease risks for more harmful phenotypes [15]. For such phenotype strata, rare/low frequency and strong effect variants are expected to be found with WGS. From this study, it was suggested that functional-variant annotation with deeply phenotyped individuals would be useful for finding disease etiologies. Another issue related to populations is that we can use the fact that allele frequencies of variants are different depending on population groups. For each variant, if populations that have a greater advantage for its statistical detection power are selected by considering the difference of its allele frequencies across populations, we will have a much greater chance to detect new disease etiologies. Conducting GWAS across different populations will be of great help.

8.2 Trend and Future of GWAS

Omic Annotation

Although GWAS have reported many disease-associated loci, it is more challenging to identify the functional (causative) variants that directly influence disease. They might be variants located near or even distal to the landmark SNPs. Such analyses have not had great success thus far, and have only just begun.

When GWAS were first introduced, people expected that associated variations would be missense ones that change protein coding amino acids. However, the majority were not missense, and not under linkage disequilibrium with missense variation. That is, they are mostly related to gene expression. Therefore, researchers have developed methodologies of prioritizing variation that could affect gene expression among many markers under linkage disequilibrium at each associated locus.

The simplest method is genome annotation. Transcriptomic and transcription regulation sequence analyses have greatly helped in identification of potential causal variants and gene products. However, there are many examples that cannot be explained by proximal gene expression control, and even many variants within gene desert regions have been reported in association studies. Association between these regions and disease is thought to be explained with variants that regulate distal genes and/or variants that are not be captured in current databases because they are stemming from population-, tissue-, and cell-type specific gene expression regulation, interaction between regions, modification of chromatin structure, or noncoding RNA. An attempt was made to estimate the detailed function of the noncoding regions. Soon, catalogues of long non-coding RNA from various kinds of cells will be constructed, using CRISPRi-based genome-scale techniques for example, and used for fine-mapping of causal variants/genes and their interpretation [34, 35]. In addition, the amazing recent progress of sequencing technologies have enabled profiling of various molecules in human cells, which have been changing the interpretation of GWAS results from experimental functional analyses to integrated information analyses with trait information from different cell/tissue types. There are many cell traits related to the expression and modification of genes that can help interpret the GWAS results. Among them, and one of the most important traits, is the epigenome. The mouse ENCODE epigenome map was the first constructed, based on experiments with various mouse cells/tissues [36]. In addition, by mapping known GWAS loci to the map, it was found that they tend to align with enhancer elements. Building on this result, genome annotations will enable fine-mapping of causal variants, narrowing down the broad regions of many GWAS loci under linkage disequilibrium. Furthermore, the ENCODE and Epigenomics Roadmap projects have created reference maps of transcriptional regulatory regions, such as promoters and enhancer regions for various types of human tissues [1]. An epigenome analysis showed that more than 80% of genomic regions influence gene regulation or chromatin structures in at least one cell type. Therefore, it would be

reasonable to prioritize histone modification, transcription factor binding, open chromatin, and chromatin state in many different cells and tissues as the most important regulating regions. For example, using such an approach, transcription regulatory region clusters that are closely associated with pancreatic functions have been identified. Also, cell types most relevant to each disease were estimated. Recently, Mumbach et al. attempted to create high-resolution contact maps of active enhancers and disease target genes using the H3K27ac HiChIP method [37]. From now on, the epigenome map will be extended to account for the differences between individuals and populations.

In addition to these omics analyzes, many bioinformatic analyses have also been attempted: use of evolutionary conserved sequences obtained from the sequence alignment of multiple species [1], prediction of three dimensional structural changes by polymorphisms and SVs on DNA and prediction of their influence on disease, and evaluation of protein structure changes by polymorphisms in coding regions. Recently, fine-mapping to one nucleotide resolution of GWAS loci was done for inflammatory bowel disease using high density genotyping [38]. Most were found to be protein coding and TF binding site changes, and/or within tissue specific epigenome marks (especially enriched in immune cells). It can be expected that this field will be greatly advanced by genome editing technology in the future. In addition to interpreting GWAS results, such genomic annotations have been used to make GWAS much more efficient [39–41]. Through differential weighting by genome annotation (functionality), we can adjust Bonferroni correction and increase the power of GWAS [42]. Researchers have started statistical genetic analyses that integrate disease risk variants and epigenetic modification mechanisms.

Linkage Between Markers and Genes by Using QTL Analysis

One of techniques to find functional variants among many is to look at quantitative trait loci (QTL), particularly expression QTL (eQTL) that represent variants associated with gene expression. To identify eQTL, gene expression profiles and genomic variation are first obtained from cells, e.g. lymphoblastoid cell-lines and tissues, of many people. Next, correlation statistics are calculated across all or cis pairs of genomic variants and gene expression levels to exhaustively explore variation that could influence gene expression in the human genome. The first genome-wide eQTL set was identified using lymphoblastoid cell lines established in the international HapMap project [43]. This eQTL dataset is used for identifying disease-associated genes linked with functional SNPs, i.e. eQTL, discriminating them from other markers under linkage disequilibrium with the GWAS landmark SNPs. Recently, the GTEx consortium identified thousands of eQTL by simultaneously analyzing RNA-seq and genotype data of 43 type tissues obtained from each of 175 individuals. Subsequently, many studies have defined other types of QTL: chromatin accessibility QTL with DNaseI-seq, transcription factor binding sites and histone modification QTL with ChIP-seq, methylation QTL, and splicing QTL (sQTL)

[44]. Together, these results show that many variations in the human genome influence gene regulation. These QTL have been identified with lymphoblastoid cell lines from blood, peripheral blood cells, or tissues from donations (e.g. GTEx project), and differentiated cells from iPS have been used for determining eQTLs, meQTLs (methylation QTLs), and caQTLs (chromatin accessibility QTLs). In addition, 1960 individuals' WGS were recently examined for association with 644 blood metabolites, and 113 variants affecting 17 genes (mQTL) were found [45]. Interestingly, most of these were heterozygous rare variants.

Integrating GWAS and eQTL

Although it has been typical to look at eQTL for each GWAS result, some methodologies aim to analyze eQTL and GWAS in an integrated manner. One of the most striking methodologies is to do transcriptome-wide association studies (TWAS), which look at association between phenotypes and gene expression. In TWAS, gene expression is imputed (predicted) from the genotypes of the samples and eQTL reference panel of matched tissue types, obtained from projects like GTEx consortium [46]. In near future, owing to advanced profiling technologies such as next-generation sequencer, personal multi-omic data, in addition to genome data, will be obtained from many individuals and integrated for identifying functional/causal variants from GWAS results and for clarifying pathways that cause disease.

Association Studies with Omics Markers Besides SNVs

Although several studies have started to directly look at omic markers other than SNVs for disease association, they are a variety of difficulties. For example, research on epigenetic markers have the issue of whether or not the cells from the samples match well with the target disease. A recent study identifying epigenetic markers associated with BMI and adiposity used blood materials based on the hypothesis that epigenetic states in blood should correlate well with BMI and adiposity mechanisms [47]. Although some CNVs and SVs have already been used for GWAS markers, those marker sets are not yet exhaustive and many markers are too difficult to detect with current GWAS chip technologies [24]. Using WGS, it may possible to conduct genome wide association analyses with balanced rearrangements, small CNV, and STR. Catalogues, particularly population specific ones, of CNV and SV regions will likely be necessary for identifying those that are related to disease. One solution would be using long single molecule mapping in addition to the current short-read mapping techniques for next-generation sequencer data. Although it is gradually progressing at the laboratory level, it should be cataloged in a large project.

Molecular Network Analysis and Identifying Specific Cell Types for Disease

Once we obtain sufficient genes/loci associated with disease, we next expect a total analysis of disease mechanisms with the gene lists. Several methodologies have been developed to test whether or not the associated loci/markers are significantly accumulated in cell-specific, gene expression regulatory regions, molecules, or molecular networks. As a result, molecular networks, in particular, cell types deterministic for disease incidence, will be revealed. Particularly, constructing dynamic, time-dependent, and context-dependent network models for incidence and progression of common diseases will be one of most important issues. In near future, such network analyses will help making strategies for therapy targets, and drug repurposing and discovery [48].

Relationship with Environments

Although GWAS researchers have mostly been identifying genetic etiologies separate from environmental effects, it would be necessary to explore and interpret genetic etiologies while simultaneously considering environmental factors. For example, the influence of gene-environment interaction on BMI was recently investigated, and it was found that age-genotype and smoking-genotype interactions contributed to 8.1% and 4.0% of BMI variation, respectively [49]. As one of the more complicated examples, excess of weight does not always lead to diabetes; adipose tissues increase insulin resistance, and abdominal fat increases diabetes risk greater than hip and thigh fat [15]. Exercise not only allows one to control weight but also increases energy consumption of glucose and insulin sensitivity of the cells. Understanding their genetic etiologies, we will be able to clarify mechanisms behind such complicated and heterogeneous phenotypes of diseases. In addition, genetically small effects might be hindered through potential interactions with environmental factors that vary across loci. Consideration of environmental perturbation or drugs to gene transcription responses, particularly RNA processing, may increase the power of etiology detection.

Revisiting Disease Phenotypes and Traits

Disease phenotypes and traits can not be independently defined – they could be related with each other, and also could be correlated well through genetic variations. We have to consider multi-functionality and the pleiotropy of genes, e.g. genes may be affecting BMI, WHR, fasting glucose, or fasting insulin levels simultaneously. It

is necessary to analyze mutually correlated diseases and traits. Analytical methods for evaluating the relationship of genetic factors among such traits have also been developed in recent years using the results obtained from GWAS. Genetic correlation, which calculates similarity measures of genetic background between traits with whole genome information, is one such method. For example, it has been shown that neuropsychiatric diseases, like schizophrenia and bipolar disease, have very similar genetic backgrounds, which suggests that common mechanisms might be involved in their incidence. In addition, even for subtypes that have been distinguished by differences in clinical findings, such as ulcerative colitis and Crohn's disease, and have been categorized as different diseases, have been suggested to share a common background. Another aspect is phenotyping for GWAS: we should proceed further deep phenotyping of groups that have specific variants for stratified GWAS [1]. In addition, in near future, complication, dynamic, time series, condition-dependent analyses will become challenges [48]. One of interesting approaches is to use insurance claims to investigate into correlation between familial shared genetic/environment backgrounds and common diseases [49]. There is a possibility that the definitions and concepts of diseases will be reviewed based on such genetic findings in the future [50–52].

Genetic correlation can be statistically evaluated with GCTA [3] and LD score regression [53], for example. Furthermore, several new methodologies have been proposed. One method is to find associated genes common to similar diseases by using common controls [54]. Another method uses Mixture Gaussians to investigate whether genetic differences are found between given two subgroups using all SNPs [55].

8.3 GWAS Applications and the Future

GWAS Applications

GWAS, which have been exploring genetic etiologies in basic research, will soon face more practical issues: adequate interpretation and social applications. One of most expected applications will be precision medicine, i.e. proposing the optimum therapy for patients (more precisely, strata according to patients' profiles). Another will be genomic drug discovery that searches for new targets though GWAS [48]. On the basis of molecular evidence, we will be able to achieve drug repurposing, the application of drugs to other diseases than the original target disease. Prediction accuracy of clinical trials, such as predicting the main action and side effects prior to administration will be improved [56]. Lastly, the prevention of disease will be targeted: establishment of preemptive medicine with exploration of preventive/protective factors against disease and risk prediction of disease incidence is expected.

Methodologies for GWAS Applications

To establish above GWAS applications, we will need more advanced methodologies than currently established. Integration of various databases is expected to build paths from related genes to applications, improving detectability and efficiency. We will need well organized methodologies for integrating heterogeneous databases across genetic sequence structure, gene expression and protein level change, epigenetic modifications, gene function change in model animals, high-dimensional network information, disease epidemiology, drug discovery databases, and clinical information. In addition, we have entered the era of analyzing omic data of patients'/individuals' tissue cells obtained from surgery, biopsy, and organ donation as well as blood. Multi-omic and trans-omic analyses of data from epigenome, transcriptome, proteome, metabolome etc., in addition to the genome, are expected to lead to a much better understanding of disease mechanisms. Furthermore, various types of next-generation sequencing methodologies and new assays will accelerate analyses of disease biomarkers and variants at a single cell level. Metagenomes of intestinal bacteria interacting with humans will also be one of the important omics as it greatly affects disease and drug response via our immune-systems. Another future issue is to develop methods of utilizing clinical information, including electric medical records, diagnostic tests, biochemical tests, drug treatments and their effects/side-effects, as well as electric health records.

PheWAS (phenome-wide association study), which have already been done for association studies with wide-phenotype sets and genotype data, could be a comprehensive analysis method of association between clinical information stored in the electronic medical records and genetic variation, for example. It is expected to be used for prediction of drug response, i.e. efficacy and side effects, in the human body before administration. The analysis of medical big data, including genomic information, has progressed, and the arrival of an era where big data are applied to improve prediction accuracy of clinical trials is soon to come. As an example of improving prediction accuracy of clinical trials, it may be considered with exome studies for looking at variants/mutations to prioritize drugs passing through phase III [56]. These will be done more comprehensively and systematically, and development of new methods will be necessary.

In addition to the omics and clinical data, there will also be an advancement of methodologies for comprehensively acquiring biomedical data using biological sensors and monitoring devices (bioimaging, PET, MRI, mobiles) and monitoring environment both comprehensively and time-dependently to completely describe the states of our bodies [57, 58]. Novel methods for collecting these data from large groups will be also necessary. Disease, population, retrospective, and prospective cohorts will become larger and larger. For example, BioVU already has Electronic Medical Records of more than 230,000 people and has started genotyping their DNA. They also explore methods for drug discovery and repurposing by combining expression data with the genomic data. The relationship between the Mendelian/orphan disease genes and common diseases has also been analyzed. The DiscovEHR

study is another big collaboration between the Regeneron Genetics Center and Geisinger Health System, which aims to combine high-throughput DNA sequencing technology with longitudinal electronic health records for discovery of genetic variation important for human disease and therapeutic response [59, 60].

The US government has launched the Precision Medicine Initiative: a prospective cohort to collect genomic sequence, clinical and lifestyle data of one million people [28]. In addition to the human genome, biochemical data of cell-free DNA, proteome, metabolome, and personal activity records, such as social networking and purchasing records of OTC drugs, are also subject to data collection, in reflection of the people's awareness of public health. One of the objectives of such comprehensive monitoring is to find prophylactic or alleviating factors by deeply investigating individuals who do not develop disease while having mutations related to Mendelian genetic diseases or familial tumors and based on evidence, and finally to establish preventive medicine. Such ideas of sharing genomic and clinical data are now planned on a global scale. The Global Alliance for Genomics and Health (GA4GH), which aims to accelerate research and clinical application by promoting global sharing of genomic and clinical data, was launched in 2013, in over 40 countries with more than 420 universities and companies participating, and are making a standard format for data sharing and a policy of ethics and regulation [61].

Now, in order to further advance precision medicine, it is necessary to predict the onset and progress of disease. For this, mathematical models of disease prediction will be necessary. In the situation of "N << p" (new NP problem), it will be more useful to take an approach to explaining and predicting phenomena using quantitative mathematical models rather than to clarify the function and contribution of each genomic variation. The simplest, and currently the most powerful, approach is using genome-wide polygenic score [62]. In actuality, the prediction of the risk of trafficking and disease development using a linear mixture model or Bayesian model, considering the variance and covariance matrix of genotype data as a kernel matrix, is shown to be highly accurate [63]. Moving forward, theoretical mathematical formulas will be examined and the addition of genetic factors (rare variants, intergenic interactions, gene-environment interactions, epigenome modifications, etc.) other than common variants to models will be in progress. In the future, it will be required to consider the complexity and temporal progression of diseases and their application to therapeutic target strategies based on the prediction of drug responses, network analysis that considers diseases as intermolecular relationships, and methodologies for time series analysis of diseases. Inference of causality will also be one of the important issues to solve the mechanisms and to predict the progress more accurately [33, 64, 65].

Furthermore, one of the most anticipated and promising medical techniques is artificial intelligence (AI). AI is being developed to realize human intelligence by a computer. It has three main functions: accumulation of external knowledge, learning from this, and reasoning about new cases. Recently, AI has had a resurgence in popularity due to the excellent progress of machine learning techniques based on deep learning and image big data, however, deep learning is not everything. Extraction of meaningful information from huge data, including non-structural data

such as electronic medical records and literature, and construction of meaningful reasoning of medical treatments largely depends on not only deep learning but also natural language processing, hypothetical reasoning, implementation for social value, and so on. Data sources in the medical field, MEDLINE, i.e. medical literature abstracts, NCCN guidelines, clinical trial reports, etc. can be used. Instead of a questionnaire, descriptions of a patient's symptoms and profiles including genetic information are input. As the outputs corresponding to the questionnaire, options such as suitable treatment methods, tests, and clinical trials, for examples, will be proposed. Jun Wang et al. recently launched ICarbonX, the center of an alliance to develop AI to revolutionize health care [66]. "The iCarbonX alliance will scour biological molecules from various tissues to provide a more accurate and actionable picture of someone's health" [66]. "The end result will be an unwieldy set of data from various sources, which is why Wang and a team at iCarbonX are developing algorithms to understand how these variables correlate with healthy or diseased states. The Meum app enables users to enter their meals and activity levels, as well as any physiological or vital-sign data, and gives advice on what to eat, when to sleep and how active they should be" [66]. With AI, it will be possible to accumulate enormous medical and biological knowledge in the past far exceeding the limits of humans, to make it efficient, to learn, and to infer therapies. In clinical practices, AI will be able to even propose candidates of new therapies that are not in conventional protocols. However, how to make such inferences correctly depends on how the AI is implemented. It should be kept in mind that AI is not a universal machine, it uses solutions that are developed by humans, and we use datasets that are rather small when we consider the algorithms for the inference. However, the data are getting far larger than originally thought because they are updated every day. Unexpected results may occur.

In the future, the key to success will be determining how to make the clinician, mathematicians, information scientists, and national policy makers work in collaboration to construct the whole system with AI, for interpreting and utilizing GWAS results.

References

1. Flannick J, Florez JC (2016) Type 2 diabetes: genetic data sharing to advance complex disease research. Nat Rev Genet 17:535–549
2. Visscher PM, Yang J, Goddard ME (2010) A commentary on 'common SNPs explain a large proportion of the heritability for human height' by Yang et al. (2010). Twin Res Hum Genet 13:517–524
3. Yang L et al (2011) GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet 88:76–82
4. Stahl EA et al (2012) Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. Nat Genet 44:483–489
5. Speed D et al (2017) Reevaluation of SNP heritability in complex human traits. Nat Genet 49:986–992

6. Yang J, Zeng J, Goddard ME, Wray NR, Visscher PM (2017) Concepts, estimation and interpretation of SNP-based heritability. Nat Genet 49:1304–1310
7. Fuchsberger C et al (2016) The genetic architecture of type 2 diabetes. Nature 536:41–47
8. Marouli E et al (2017) Rare and low-frequency coding variants alter human adult height. Nature 542:186–190
9. Surakka I et al (2015) The impact of low-frequency and rare variants on lipid levels. Nat Genet 47:589–597
10. Zheng HF et al (2015) Whole-genome sequencing identifies EN1 as a determinant of bone density and fracture. Nature 526:112–117
11. Singh T et al (2017) The contribution of rare variants to risk of schizophrenia in individuals with and without intellectual disability. Nat Genet 49:1167–1173
12. Dickson SP et al (2010) Rare variants create synthetic genome-wide associations. PLoS Biol 8:e1000294
13. Hunt KA et al (2013) Negligible impact of rare autoimmune-locus coding-region variants on missing heritability. Nature 498:232–235
14. UK10K Consortium et al (2015) The UK10K project identifies rare variants in health and disease. Nature 526:82–90
15. Rich SS (2016) Diabetes: still a geneticist's nightmare. Nature 536:37–38
16. Lee S et al (2014) Rare-variant association analysis: study designs and statistical tests. Am J Hum Genet 95:5–23
17. Cruchaga C et al (2014) Rare coding variants in the phospholipase D3 gene confer risk for Alzheimer's disease. Nature 505:550–554
18. Liu DJ et al (2014) Meta-analysis of gene-level tests for rare variant association. Nat Genet 46:200–204
19. Ahlbom A, Alfredsson L (2005) Interaction: a word with two meanings creates confusion. Eur J Epidemiol 20:563–564
20. Cordell HJ, Clayton DG (2002) A unified stepwise regression procedure for evaluating the relative effects of polymorphisms within a gene using case/control or family data: application to HLA in type 1 diabetes. Am J Hum Genet 70:124–141
21. Lenz TL et al (2015) Widespread non-additive and interaction effects within HLA loci modulate the risk of autoimmune diseases. Nat Genet 47:1085–1090
22. Weiner DJ et al (2017) Polygenic transmission disequilibrium confirms that common and rare variation act additively to create risk for autism spectrum disorders. Nat Genet 49:978–985
23. Zhu Z et al (2015) Dominance genetic variation contributes little to the missing heritability for human complex traits. Am J Hum Genet 96:377–385
24. CNV and Schizophrenia Working Groups of the Psychiatric Genomics Consortium (2017) Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects. Nat Genet 49:27–35
25. Sekar A et al (2016) Schizophrenia risk from complex variation of complement component 4. Nature 530:177–183
26. Chuong EB, Elde NC, Feschotte C (2017) Regulatory activities of transposable elements: from conflicts to benefits. Nat Rev Genet 18:71–86
27. Hoffmann TJ et al (2017) Genome-wide association analyses using electronic health records identify new loci influencing blood pressure variation. Nat Genet 49:54–64
28. https://obamawhitehouse.archives.gov/node/333101
29. http://www.ukbiobank.ac.uk
30. Moore C, Sambrook J, Walker M, Tolkien Z, Kaptoge S, Allen D, Mehenny S, Mant J, Di Angelantonio E, Thompson SG et al (2014) The INTERVAL trial to determine whether intervals between blood donations can be safely and acceptably decreased to optimise blood supply: study protocol for a randomised controlled trial. Trials 363:15
31. Astle WJ et al (2016) The allelic landscape of human blood cell trait variation and links to common complex disease. Cell 167:1415–1429
32. Pasaniuc B, Price AL (2017) Dissecting the genetics of complex traits using summary association statistics. Nat Rev Genet 18:117–127

33. Visscher PM et al (2017) Am J Hum Genet 101:5–22
34. Iyer MK et al (2015) The landscape of long noncoding RNAs in the human transcriptome. Nat Genet 47:199–208
35. Liu SJ et al (2017) CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells. Science 355:aah7111
36. Yue F et al (2014) A comparative encyclopedia of DNA elements in the mouse genome. Nature 515:355–364
37. Mumbach MR et al (2017) Enhancer connectome in primary human cells identifies target genes of disease-associated DNA elements. Nat Genet 49:1602–1612
38. Huang H et al (2017) Fine-mapping inflammatory bowel disease loci to single-variant resolution. Nature 547:173–178
39. Yang J et al (2017) A scalable Bayesian method for integrating functional information in genome-wide association studies. Am J Hum Genet 101:404–416
40. He Z, Xu B, Lee S, Ionita-Laza I (2017) Unified sequence-based association tests allowing for multiple functional annotations and meta-analysis of noncoding variation in metabochip data. Am J Hum Genet 101:340–352
41. Trynka G et al (2015) Disentangling the effects of colocalizing genomic annotations to functionally prioritize non-coding variants within complex-trait loci. Am J Hum Genet 97:139–152
42. Sveinbjornsson G et al (2016) Weighting sequence variants based on their annotation increases power of whole-genome association studies. Nat Genet 48:314–317
43. Stranger BE et al (2007) Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science 315:848–853
44. Zhang X et al (2015) Identification of common genetic variants controlling transcript isoform variation in human whole blood. Nat Genet 47:345–352
45. Long T et al (2017) Whole-genome sequencing identifies common-to-rare variants associated with human blood metabolites. Nat Genet 49:568–578
46. Ishigaki K et al (2017) Polygenic burdens on cell-specific pathways underlie the risk of rheumatoid arthritis. Nat Genet 49:1120–1125
47. Wahl S et al (2017) Epigenome-wide association study of body mass index, and the adverse outcomes of adiposity. Nature 541:81–86
48. Hu JX, Thomas CE, Brunak S (2016) Network biology concepts in complex disease comorbidities. Nat Rev Genet 17:615–629
49. Robinson MR et al (2017) Genotype-covariate interaction effects and the heritability of adult body mass index. Nat Genet 49:1174
50. Wang K, Gaitsch H, Poon H, Cox NJ, Rzhetsky A (2017) Classification of common human diseases derived from shared genetic and environmental determinants. Nat Genet 49:1319–1325
51. Cox NJ (2017) Reaching for the next branch on the biobank tree of knowledge. Nat Genet 49:1295–1296
52. Cortes A et al (2017) Bayesian analysis of genetic association across tree-structured routine healthcare data in the UK Biobank. Nat Genet 49:1311–1318
53. Bulik-Sullivan BK et al (2015) LD score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat Genet 47:291–295
54. Fortune MD et al (2015) Statistical colocalization of genetic risk variants for related autoimmune diseases in the context of common controls. Nat Genet 47:839–846
55. Liley J, Todd JA, Wallace C (2017) A method for identifying genetic heterogeneity within phenotypically defined disease subgroups. Nat Genet 49:310–316
56. Dewey FE et al (2016) Inactivating variants in ANGPTL4 and risk of coronary artery disease. N Engl J Med 374:1123–1133
57. Yom-Tov E (2016) Crowdsourced Health. The MIT Press, Cambridge, MA
58. Neff G, Nafus D (2016) Self-tracking. The MIT Press, Cambridge, MA
59. Dewey FE et al (2016) Distribution and clinical impact of functional variants in 50,726 whole-exome sequences from the DiscovEHR Study. Science 354(6319). https://doi.org/10.1126/science.aaf6814

60. Abul-Husn NS et al. Genetic identification of familial hypercholesterolemia within a single U.S. health care system. Science 354(6319). https://doi.org/10.1126/science.aaf7000
61. https://www.ga4gh.org
62. Khera AV et al (2018) Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nat Genet 50:1219–1224
63. Zhou X, Carbonetto P, Stephens M (2013) Polygenic modeling with bayesian sparse linear mixed models. PLoS Genet 9:e1003264
64. Bowden J, Davey Smith G, Burgess S (2015) Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. Int J Epidemiol 44:512–525
65. Burgess S, Butterworth A, Thompson SG (2013) Mendelian randomization analysis with multiple genetic variants using summarized data. Genet Epidemiol 37:658–665
66. Cyranoski D (2017) Chinese AI company plans to mine health data faster than rivals. Nature 541:141–142