Journal of Human Genetics (2010) 55, 195–206

www.nature.com/jhg

REVIEW

The pursuit of genome-wide association studies: where are we now?

Chee Seng Ku[1], En Yun Loy[1], Yudi Pawitan[2] and Kee Seng Chia[1,2]

It is now 5 years since the first genome-wide association studies (GWAS), published in 2005, identified a common risk allele with large effect size for age-related macular degeneration in a small sample set. Following this exciting finding, researchers have become optimistic about the prospect of the genome-wide association approach. However, most of the risk alleles identified in the subsequent GWAS for various complex diseases are common with small effect sizes (odds ratio <1.5). So far, more than 450 GWAS have been published and the associations of greater than 2000 single nucleotide polymorphisms (SNPs) or genetic loci were reported. The aim of this review paper is to give an overview of the evolving field of GWAS, discuss the progress that has been made by GWAS and some of the interesting findings, and summarize what we have learned over the past 5 years about the genetic basis of human complex diseases. This review will focus on GWAS of SNPs association for complex diseases but not studies of copy number variations.
Journal of Human Genetics (2010) **55,** 195–206; doi:10.1038/jhg.2010.19; published online 19 March 2010

Keywords: cancer; complex diseases; copy number variants; genome-wide association studies; indels; rare variants; resequencing; 1000 Genomes Project

INTRODUCTION

It is the fifth year of genome-wide association studies (GWAS) after the first study was published in 2005, which identified the association of complement factor H (CFH) with age-related macular degeneration (AMD).[1] The publication of this landmark study marked the start of a new era in the genetic studies of human complex diseases. It was the first GWAS that used a commercial genotyping array and interrogated ~100 000 single-nucleotide polymorphisms (SNPs) throughout the human genome. The success of finding a common risk allele with an effect size of 4.6 (per allele odds ratio (OR) or 7.4 for homozygous risk allele) in a small sample set of 96 cases and 50 controls has generated considerable excitement in the genetics community. The *P*-value of the strongest SNP association surpassed the genome-wide significance threshold after Bonferroni correction. Both the high frequency of the allele and the large effect size contributed to the highly significant association.

Before this discovery, some were skeptical about the genome wide association approach, whether it would be viable to identify novel risk alleles or genetic loci for human complex diseases. The AMD study gave firm assurance to the genetics community about the efficiency and feasibility of the GWAS approach to look for unknown disease-associated variants. This encouraging finding also raised the enthusiasm and confidence among researchers worldwide to conduct numerous GWAS to decipher the genetic basis of various complex traits, and finally led to the explosion of GWAS publications in 2007, which was labeled as The Year of GWAS (Figure 1). Unfortunately, the discovery of CFH was a low-hanging fruit, and the attempt to find additional common risk alleles with moderate to large effect sizes (OR >2.0) has not been fruitful. Instead, most of the disease-associated risk alleles conferred small effect sizes (OR <1.5).[2,3] The finding of not many large effect size variants is expected, as the purifying selection pressure will remove them from the populations or keep their population frequencies low. However, GWAS is not a powerful approach for studying rare or uncommon SNPs (regardless of their effect sizes), because they are poorly represented in the commercial whole genome genotyping arrays.[4] On the other hand, the finding of most of the risk alleles with small effect sizes is in concordance with the common-disease common-variant (CD/CV) hypothesis which formed the basis of GWAS and also steered the SNPs selection approaches towards common SNPs in genotyping arrays.[5] However, it is unexpected that the effect sizes of a considerable fraction of the disease-associated risk alleles are as small as OR <1.1–1.2.

GWAS BEFORE HAPMAP

The definition of GWAS varied among researchers, in fact no consensus criteria exist on defining a GWAS (for example, the minimum number of SNPs and samples that need to be genotyped and included in a study, the density and distribution or coverage of SNPs throughput the genome and the requirement of replication and validation steps) even after more than 450 studies have been published

[1]Department of Epidemiology and Public Health, Yong Loo Lin School of Medicine, Centre for Molecular Epidemiology, National University of Singapore, Singapore and [2]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden
Correspondence: CS Ku or Professor KS Chia, Department of Epidemiology and Public Health, Center for Molecular Epidemiology, National University of Singapore, 16 Medical Drive, Singapore 117597, Singapore.
E-mail: cmekcs@nus.edu.sg; or ephcks@nus.edu.sg
Received 7 December 2009; revised 9 February 2010; accepted 19 February 2010; published online 19 March 2010

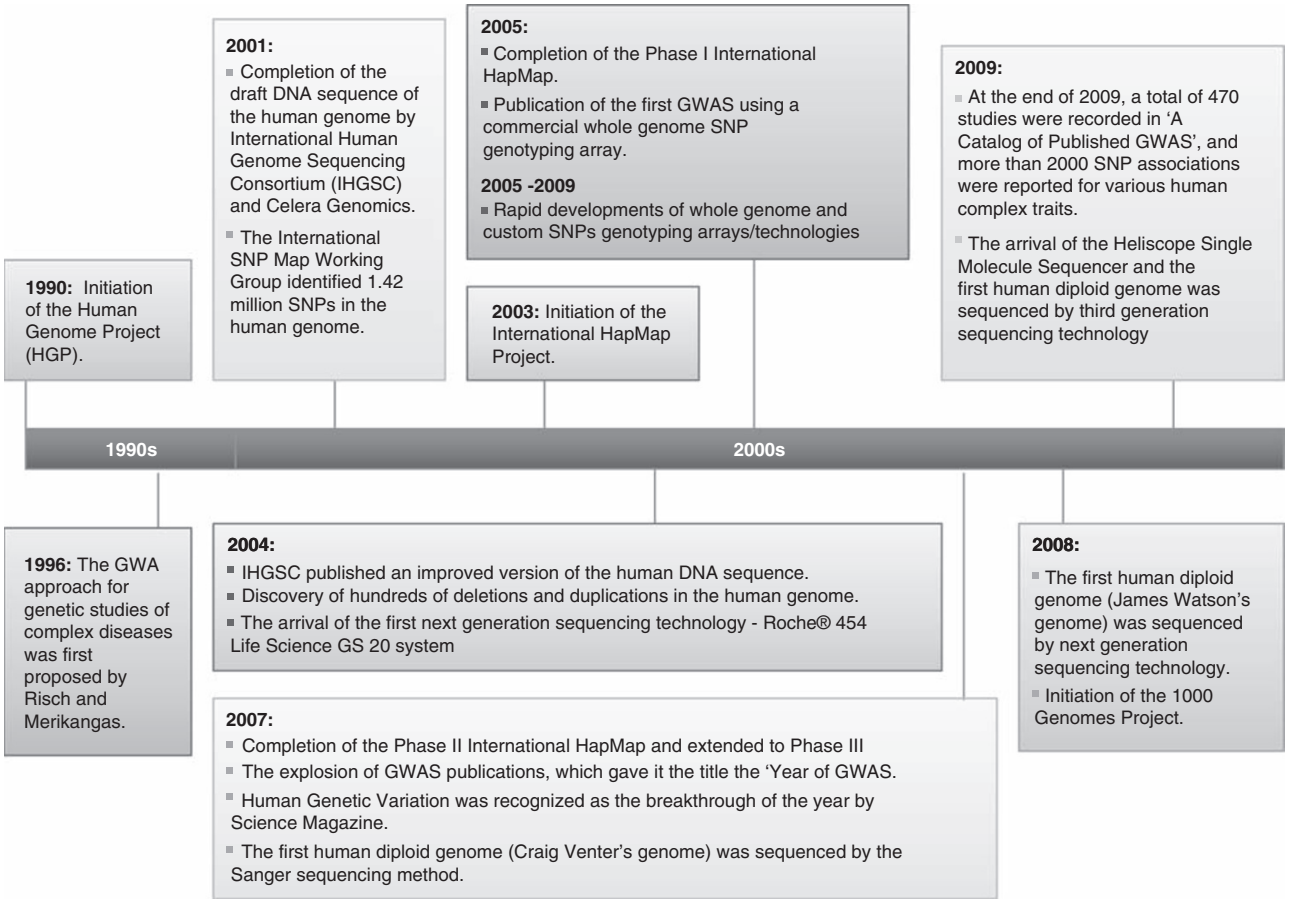

Figure 1 Major developments in GWAS and human genome sequencing.

since 2005. All the published GWAS have been cataloged in the National Human Genome Research Institute 'A Catalog of Published Genome-Wide Association Studies' (http://www.genome.gov/gwastudies/) and this resource includes only those studies that attempted to genotype at least 100 000 SNPs in the initial stage.[3] As a result, the AMD study by Klein *et al.*[1] was recorded as the first entry in the catalog and was generally being recognized as the first GWAS.

Nevertheless, it is noteworthy that several 'early versions' of GWAS have already been performed and published before 2005. The first of them came from Ozaki *et al.*,[6] which was published in 2002. The study attempted to genotype ~93 000 gene-based SNPs in a case–control study for myocardial infarction (MI) and identified the association of LTA gene for the disease.[6] Two more studies were also being carried out in 2003 to identify the genetic risk factors for immunoglobulin A nephropathy and diabetic nephropathy where each study genotyped about 80 000 and 56 000 SNPs, respectively, and both studies identified a new candidate gene for each disease, namely, PIGR and SLC12A3.[7,8] Subsequently, an additional candidate gene (that is, ELMO1) was also identified for diabetic nephropathy in 2005 using the same study design, and likewise, the identification of TNFSF15 for Crohn's disease (CD), and CALM1 for osteoarthritis where all the studies have genotyped around 80 000 SNPs.[9–11]

The examples of these early days of GWAS were performed in Japanese populations by the same group of researchers. However, there are at least two major differences between these studies and the GWAS performed by Klein *et al.*[1] In contrast to the GWAS of AMD, which used a commercial genotyping array; that is, Affymetrix Human Mapping 100K Set, the genotyping of all of the early versions of GWAS was not performed by microarray technologies, but instead by applying high-throughput multiplex PCR-Invader assay methods.[12] However, the success rate of this genotyping method was much lower than that of genotyping arrays, that is, about 70% as reported by Ozaki *et al.*[6] in their GWAS of MI. Furthermore, these studies only focused on gene-based SNPs where they were selected randomly from a gene-based SNPs discovery study,[13,14] instead of random selection of SNPs covering genes and intergenic regions as in genotyping arrays. Although the first GWAS only genotyped less than 100 000 SNPs, this number of SNPs was considered large at that time in the absence of high-throughput genotyping technologies.

INTERNATIONAL HAPMAP PROJECT

One of the important developments and a significant milestone in the history of GWAS was the completion of the phase I and phase II International HapMap Project in 2005 and 2007, respectively.[15,16] The aim of the HapMap initiative is to validate the millions of SNPs that were identified during and after the completion of the Human Genome Project, and then to characterize their correlation or linkage disequilibrium (LD) patterns in populations of European, Asian and African ancestry. The two phases of the HapMap Project have resulted in the validation and characterization of the LD patterns of more than 4 million SNPs. The HapMap data have shown that the existence of LD significantly reduces the number of SNPs that needs to be genotyped in GWAS. This genome-wide indirect association approach is dependent on surrogate markers to locate disease or functional

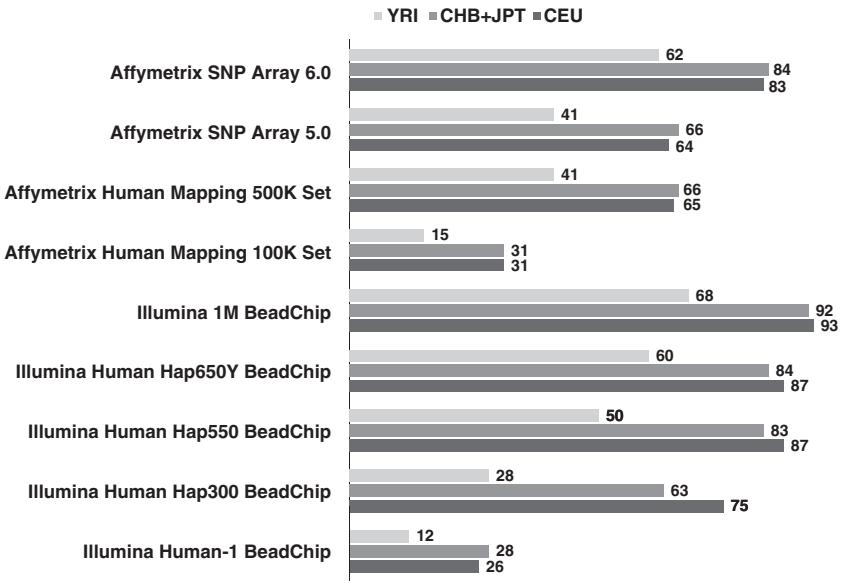

Figure 2 Genome coverage (%) at $r^2 > 0.8$ for several whole genome genotyping arrays in International HapMap populations.

variants through LD. Some published data have shown that only half to one million tagging SNPs are required to capture the information of most of the SNPs genotyped in the HapMap Project at $r^2 > 0.8$ (Figure 2).[17–19]

The HapMap database also provided a useful resource for developing whole genome SNPs genotyping arrays. The data are used to optimize genome coverage by taking advantage of the efficiency of the tagging SNP approach. Commercial whole genome SNPs genotyping arrays manufactured by Illumina (San Diego, CA, USA) and Affymetrix (Santa Clara, CA, USA) have been used in almost all the GWAS to genotype several hundred thousand to one million SNPs. Since the completion of the Phase I HapMap Project in 2005, a number of genotyping arrays have been designed and introduced to the market, and the newer arrays have significantly improved in genome coverage and have also expanded their application to detect copy number variants (CNVs), in addition to SNP genotyping.[4] Clearly, the completion of the International HapMap Project and the subsequent rapid developments of whole genome genotyping arrays have been the key components in making the GWAS a feasible approach.

GWAS AFTER HAPMAP—THE PROGRESS OVER THE PAST 5 YEARS

The first 2 years: 2005 and 2006

The publication of GWAS was slow in the initial 2 years; only 10 GWAS were published, but that was actually the germination period that led to the explosion of GWAS publications in the subsequent years. Some promising findings were achieved, for example, two new candidate genes were identified for AMD, the CFH gene for the dry-subtype and the HTRA1 gene was associated with the wet-subtype of AMD.[1,20] The associations of these two genes have been robustly replicated in subsequent studies.[21,22] The other significant finding was the identification of the interleukin (IL)23R gene for CD, which encodes a subunit of the receptor binding to IL23 (a proinflammatory cytokine). This made it a strong biologically plausible gene for CD, a chronic inflammatory bowel disease (IBD).[23] Likewise, the association of IL23 receptor (IL23R) has been unequivocally replicated.[24] One common feature of these studies is the careful ascertainment of the cases, for example, in the AMD GWAS, only cases with the presence of large drusen were recruited, and for the CD GWAS, only ileal cases were enrolled. This careful case selection minimized the misclassification effect due to possible phenotypic heterogeneity, and hence enriched the genetic component of the sub-phenotype. This step would enhance the detection power for genetic variants associated with the specific sub-phenotype if the phenotypic heterogeneity (diverse disease manifestations) were due to genetic heterogeneity (different sets of genetic factors for each subtype of the disease).

On the contrary, the other two GWAS of complex diseases, one for obesity, and the other for Parkinson's disease (PD) have yielded some inconsistent results.[25,26] The identification of the INSIG2 gene for obesity has not been consistently replicated; many studies have been trying to verify the finding, but the association was only found in some but not all the studies.[27] Therefore, a well-conducted meta-analysis is required to resolve the conflicting results. In fact, a meta-analysis involving a large sample size of more than 70 000 individuals has been carried out recently to examine the INSIG2 association with obesity, and it found no evidence to support the association of the landmark SNP (rs7566605) with obesity.[28] Similarly, the SNPs that were identified for PD were not replicated in a large scale international study.[29]

The second study for PD was only in its first stage of analysis at that time and it did not find any significant result. Nevertheless, it was the first GWAS that made the genotype data of cases and controls publicly available.[30] In the following years, many GWAS also shared their genotype data together with the disease phenotype, for example, the Wellcome Trust Case Control Consortium (WTCCC) also made the GWAS data of seven common diseases publicly available to other researchers through application to the consortium.[31] Similarly, the Genetic Association Information Network, a public–private partnership that conducted a series of GWAS also shares its genotype data with other researchers; the data have been deposited in the database of Genotype and Phenotype.[32] Availability of these resources allows the research community to accelerate the pace of discovery of disease-associated variants. Consistent with this effort, private companies such as Illumina also set up the iControl database that contains the genotype data for control samples that are genotyped by various Illumina genotyping arrays. The data can be downloaded and

integrated into one's GWAS to save the genotyping cost on controls or to increase the statistical power by increasing the ratio of controls to cases, although several issues need to be taken care of such as population stratification and disease misclassification when using the controls. Even if the controls were genotyped by a different genotyping array from the cases, the controls can still be used, because the different sets of SNPs in cases and controls can be 'standardized' through imputation methods, followed by association analysis.[33]

The other four GWAS investigated other complex traits such as QT interval, memory performance, addiction and nicotine dependency. Of particular interest was the finding of NOS1AP association with QT interval variation, which is an important measure of cardiac repolarization.[34] The NOS1AP association was further substantiated by two later GWAS in a large sample set of nearly 30 000 samples for the initial and replication studies.[35,36] Besides confirming the previously known gene, several new loci were also identified. Interestingly, both studies also found KCNQ1 to be associated with the trait; this is because the gene was also associated with type-2 diabetes (T2D).[37,38] This finding further revealed the pleiotropic effect of gene or genetic locus influences on multiple traits. Prolongation of the QT interval increases the risk of ventricular arrhythmias and sudden cardiac death, and this index also predicts cardiovascular mortality among healthy individuals.

The other relevant example of genetic pleiotropic effect in this context is the 9p21 locus, which was also found to be associated with coronary artery disease (CAD), MI and T2D.[39–41] As far as the association of the 9p21 locus with CAD, MI and T2D (that is, possible genetic pleiotropy) is concerned, it is noteworthy that the SNPs for CAD/MI (rs10757278—G allele) and T2D (rs10811661—T allele) are not found in the same LD block and they are not correlated. The LD block where rs10757278 is located contains two candidate genes that is, CDKN2A and CDKN2B, but the other LD block for rs10811661 contains no known genes. Although the SNPs were found in two adjacent LD blocks, they are located close to each other where their distance is less than 10 kb, and CDKN2A and CDKN2B are the nearest genes to them. As the two SNPs are independent, it was unclear at that time whether the SNP that is associated with T2D is also associated with CAD/MI and vice versa. To further investigate this, Helgadottir *et al.*[42] examined the association of these two SNPs with T2D, CAD and four other arterial diseases, and found that rs10757278—G allele is associated with CAD, abdominal aortic aneurysm and intracranial aneurysm, but not with T2D. Similarly, the association of rs10811661—T allele with T2D does not apply to CAD and other arterial diseases that they have investigated.[42] Although this study confirmed the specificity of the SNP associations to CAD and T2D respectively, it is still unclear whether the two risk alleles residing in the same locus will eventually affect the same gene or functional element leading to a defect or alteration of a biological pathway that is responsible for the diseases. Therefore, the pleiotropic effect of 9p21 locus warrants further investigation. In summary, these studies seem to provide some preliminary evidence (statistical association) of shared genetic susceptibility loci (that is, 9p21 locus and KCNQ1) between cardiovascular and metabolic diseases, but further biological functional studies are needed to investigate whether some common etiological pathways exist for these two distinct but related groups of diseases.

The year of GWAS: 2007

The rapid publication of GWAS has led to the identification of a plethora of new SNPs or genetic loci for various complex diseases and traits. So far, more than 450 GWAS have been published since 2005, and associations of greater than 2000 SNPs or loci were reported, some of which are statistically robust associations, with others still requiring further replication studies for confirmation. Because of the rapid finding of new disease-associated SNPs and the concomitant discovery of thousands of CNVs in the human genome, Human Genetic Variation was named the Breakthrough of The Year in 2007, to recognize the prominent progress achieved in the research of human genetic variation.[43] These discoveries have remarkably transformed the field of disease genetics. This is in comparison with the pre-GWAS era where whole genome linkage mapping and candidate gene association studies were broadly applied and there was limited success achieved in the genetic studies of complex diseases at that time.

The landmark GWAS in 2007 was the WTCCC study, which examined the genetic basis of seven complex diseases. A total of about 17 000 samples, 2000 cases for each of the diseases and 3000 shared controls were genotyped for half a million SNPs.[31] Substantial success was achieved by the WTCCC GWAS, in which many novel genes and genetic loci were identified and later replicated in subsequent studies, for example, the finding of IRGM gene for CD and four new loci were also found to be associated with type-1 diabetes (T1D).[44,45] The IRGM was the second autophagy gene identified for CD after the ATG16L1,[46] which further supports the involvement of the autophagy pathway in the patho-physiology of CD. More importantly, the WTCCC study also addressed some crucial methodological and analytical issues in designing and conducting a GWAS.

The other studies to be noted are the Breast Cancer Association Consortium (BCAC) GWAS and the five GWAS of T2D. The BCAC study is the first published GWAS of a large-scale endeavor and collaboration from multiple countries to study the genetic basis of one disease. Researchers and study groups from more than 20 countries have been participating in the BCAC. Using a three-stage study design and a total sample size greater than 50 000, the BCAC GWAS identified several novel susceptibility loci for breast cancer, of which the locus containing the FGFR2 gene is of particular interest. This gene encodes a tyrosine kinase receptor and is found to be overexpressed in breast cancer.[47] At the same time, the FGFR2 association was also uncovered by another breast cancer GWAS.[48]

One of the well-studied diseases in 2007 was T2D; five novel genetic loci were identified (namely, SLC30A8, HHEX, CDKAL1, CDKN2A/2B and IGF2BP2) and robustly confirmed in a number of subsequent studies. These newly identified loci were not previously suspected for T2D.[49–53] Four smaller GWAS of T2D were also published in that year, but did not produce any significant findings. Before the GWAS era, only two genes (PPAGγ and KCNJ11) have been consistently associated with this metabolic disorder, one of which, the TCF7L2 gene, was identified through linkage analysis and fine mapping.[54] Besides breast cancer and T2D, several other diseases have also been studied intensively in that year; the prominent ones being CD, T1D, amyotrophic lateral sclerosis, coronary diseases, restless legs syndrome, prostate cancer and colorectal cancer.

The results of the colorectal cancer GWAS also deserves some attention, because these studies provide the first evidence showing that the 8q24 locus was associated with more than one cancer.[55,56] Both studies did not identify other SNP associations for colorectal cancer except the 8q24 locus, which was found earlier for prostate cancer.[57,58] In fact, subsequent studies have shown the association of 8q24 with multiple cancers. Intriguingly, the 8q24 locus was only associated with some, but not all the cancers that have been studied so far.[59] Therefore, it would be interesting to explore the reason behind it, to gain further insights on the similarities and differences in the pathology of different cancers.

Meanwhile, the GWAS of exfoliation glaucoma should be noted as well, as this study uncovered another common risk allele of large effect size for complex diseases. Two non-synonymous SNPs in exon 1 of LOXL1 gene were found to increase the susceptibility risk of the eye disease. One of the alleles increased the risk by 20-folds.[60] This is by far the largest effect size identified by the GWAS for complex diseases. As GWAS was evolving, researchers have also begun recognizing the importance of replication to corroborate the associations found by GWAS. Replication has been accepted as the gold standard to distinguish between false-positives and genuine associations and it has become mandatory to validate the results from GWAS before declaring a genuine association. This imperative has been further underscored by the publication of guidelines to conduct proper replication studies by NCI-NHGRI Working Group on Replication in Association Studies in 2007.[61]

Two GWAS in the same year also demonstrated the utility of expression quantitative trait loci (eQTL) data for disease gene mapping. For example, the asthma GWAS identified a number of SNPs in a strong LD region, which spanned >200 kb on chromosome 17q23. The LD region contains 19 genes, but as none of the genes has an apparent biological role or link to asthma, it is unclear which gene is likely to be the disease causative gene. However, through the eQTL experiment, they found that the associated SNPs have significant effects on the expression levels of ORMDL3 from the cluster of genes in that region. This work illustrates the use of eQTL to help in discerning the potential disease functional gene from others in a region with strong LD.[62] For the other GWAS, Libioulle *et al.*[63] found that multiple SNPs on chromosome 5p13.1 were strongly associated with CD, even though the region is located within a 1.2 Mb gene desert and the nearest annotated gene PTGER4 is about 270 kb away from the association signals. Although the SNPs were consistently associated with the disease, their functional effect is not easy to infer, because these SNPs could either have an effect on the nearest gene or on other genes that are located further away. Using the same approach, they integrated the GWAS results with eQTL data and found that the associated SNPs influenced expression levels of PTGER4.[63] Taken together; these two studies have shown the application and feasibility of integrating GWAS results with eQTL data in disease gene mapping. This approach is feasible for regions with multiple genes in strong LD as well as for regions where no gene has previously been annotated.

The recent 2 years: 2008 and 2009

Over the past 3 years, there was increased popularity in the belief that the risk alleles that remain to be identified might have smaller effect sizes (OR <1.2), therefore despite thousands of samples, the current GWAS are still underpowered to identify them. Such SNPs are believed to exist in a considerable number in the genetic architecture of complex diseases, and the identification of common large effect alleles in CFH and LOXL1 were just the low-hanging fruits. In fact, it has been shown in a power calculation of the T2D GWAS meta-analysis, although with a considerable large sample set of ~10 000 (cases and controls), that the statistical power to detect risk alleles with OR of 1.1–1.2 was relatively low, and no power exists for those with OR <1.1 for varying allele frequencies.[64] To boost the statistical power to detect such loci, further increases in sample size seems to be the immediate next step to proceed. As a result, several meta-analysis studies combining the existing data sets of GWAS have been performed since 2008.

Different genotyping arrays by Illumina and Affymetrix were used in the GWAS, and due to different SNP selection methods between the arrays, this resulted in little overlapping of the array content, or the content in one array is only a subset of the SNPs in another array (for example, Illumina HumanHap300 and HumanHap 550). Therefore, imputation methods are very useful in combining the SNPs data sets generated from different genotyping arrays used in different GWAS in a meta-analysis. It is important to avoid unnecessary disposal of the non-overlapping SNPs among the GWAS. Genotype imputation is a method to infer the genotypes of missing or ungenotyped SNPs; that is, the SNPs that failed to be genotyped or did not genotype directly. Imputation is performed by using SNPs data generated from the genotyping array and LD information, inferring against a reference data set (International HapMap data). The discussion of meta-analysis and imputation methods is beyond the scope of this review paper. However, the guidelines and practical details of imputation-based meta-analysis of GWAS,[65] as well as the methodological and analytical issues of GWAS meta-analysis have been outlined and discussed in several papers.[66,67] The timely emergence of imputation methods together with the reference database have been the major driving force for conducting GWAS meta-analysis.

The first GWAS meta-analysis conducted by the Diabetes Genetics Replication and Meta-analysis (DIAGRAM) Consortium combined the data sets from three studies for further analysis.[68] Approximately 2.2 million SNPs were analyzed in about 10 000 European individuals in stage 1 of the study; these SNPs were either genotyped directly or inferred *in silico* using imputation methods. Following stage 1 analysis, replication testing of promising SNPs was carried out in an even larger sample set, with the sample size of the whole study approaching 100 000. This exercise yielded six additional loci that showed association with T2D at the genome-wide significance threshold. As expected, the effect sizes for all the SNPs that were identified from this meta-analysis study were small (OR <1.2). After this first study, several GWAS meta-analysis have also been performed for various diseases such as CD,[69] T1D,[70,71] multiple sclerosis,[72] rheumatoid arthritis,[73] colorectal cancer[74] and lung cancer,[75] among other diseases.

Significant progress has also been achieved in these 2 years for several major cancers. For prostate cancer, in addition to the well-established 8q24 locus, many more novel genetic loci were identified.[76–79] One striking observation is the association of TCF2 (or HNF1B) and JAZF1 with this male cancer, because both genes were also found to be associated with T2D. More interestingly, the variant in TCF2 increased the risk of prostate cancer but protective against T2D.[80] This is one of the first examples from GWAS showing a genetic variant having opposing actions on two different diseases. SNPs in 9p21 locus were also found to be associated with various cancers, in addition to cardiovascular diseases and T2D.[81–83] So far, most of the GWAS did a fast-track replication by selecting the top few or top tens of SNPs with the most significant *P*-values in stage 1 and proceeding to replicate those SNPs in stage 2 or stage 3 with larger sample sizes. However, other SNPs down the significance list are just as likely to be genuine, thus more SNPs need to be selected for replication in larger sample sets. The success of this approach to identify new prostate cancer susceptibility loci has been demonstrated by Eeles *et al.*[84] In the study, they conducted a more extensive follow-up of SNPs showing evidence of association in stage 1 of their previous GWAS. A total of ~43 000 SNPs was genotyped for replication in stage 2 and stage 3 with a sample size of about 7600 and 31 000, respectively, and this attempt successfully identified seven new loci for prostate cancer.

Similarly, considerable advances have also been made for colorectal, lung and breast cancer. One additional gene for colorectal cancer was confirmed other than the 8q24 locus, the SMAD7 gene is involved in TGF-β and Wnt signaling, and abnormalities in these pathways are well established in the pathogenesis of this cancer.[85–87] For lung

cancer, the first convincing genetic locus also emerged, 15q25, which contained several genes encoding for nicotinic acetylcholine receptors.[88–90] The combination of three GWAS further identified additional susceptibility loci for this common cancer,[75] where one of the loci (5p15.33) was also identified by another GWAS.[91] The 5p15.33 locus contains two genes, TERT and CLPTM1L, and interestingly the sequence variants at this locus were found to be associated with several cancers.[92] Although these studies firmly identified the 15q25 locus for lung cancer, the underlying associations were different between the studies. Two studies found that the 15q25 association was primarily with lung cancer and not with smoking, but the other study concluded that the link with lung cancer is mediated through smoking and nicotine dependence. Several reasons could be responsible for the disagreement among the studies.[93]

Likewise, for breast cancer, several new susceptibility loci were identified.[94,95] At the same time, the genetics of the reproductive lifespan of women, namely, the timing of menarche and menopause has also been intensively interrogated by five GWAS. The findings of these studies are interesting in the light of breast cancer and other female reproductive-associated cancers or diseases. The age at menarche and menopause influences the risk of several diseases; for example, the earlier age of menarche and/or later age of menopause increases the risk of breast, ovarian and endometrial cancers. Three studies focused on the age at menarche,[96–98] one GWAS studied the age at natural menopause,[99] and the other GWAS investigated both the traits.[100] Several loci were identified for menarche, but all the four studies definitely found an association in the 6p21 locus that contained LIN28B (a height-related gene) and two GWAS detected another locus 9q31.2. In addition, two convincing loci were also identified for menopause, 19q13.42 (BRSK1) and 20p12.3 (MCM8). Owing to the close relationship between the timing of menarche and menopause with several female cancers and other diseases, this finding also raises some intriguing questions; whether the SNPs are also directly associated with the risk of these cancers. Therefore, it would be interesting to test the associations in those diseases that could possibly provide some new insights into the disease patho-physiology related to hormone exposures.

Two GWAS of testicular germ cell cancer are noteworthy to be highlighted here; these studies together identified three novel genetic loci. Two loci were identified by both studies (KITLG and SPRY4); whereas a third locus (BAK1) was only found in one study. The finding of these GWAS is remarkable, because in comparison with other cancers studied by GWAS, several risk alleles in one of the loci that contained a strong biological plausible candidate gene for testicular cancer, which is KILTG gene, have effect sizes larger than 2.5 (per allele OR) and the risk alleles are very common (frequencies of risk alleles in controls are reported to be about 80%). The KILTG gene encodes the ligand for the membrane receptor tyrosine kinase (c-KIT), and both the somatic and germline mutations in the KIT/KILTG pathway have been well-linked to testicular cancer.[101,102] The results are notable because it is rare to find common risk alleles (in the context of SNPs) with ORs larger than 2, not only in cancers but also in other complex diseases.

Some interesting findings were also obtained for several autoimmune and chronic inflammatory diseases; the notable ones being multiple sclerosis, rheumatoid arthritis, IBD and systemic lupus erythematosus.[103–108] For example, significant advances have been achieved in identifying the genetic risk factors for IBD; that is, CD and ulcerative colitis (UC). Before the GWAS era, CARD15/NOD2 is the only candidate gene consistently associated with CD. However, the first breakthrough success came from the identification of two new candidate genes for CD, namely, TNFSF15 and IL23R in 2005 and 2006, respectively.[10,23] Subsequently, the finding of two autophagy-related genes, which are ATG16L1 and IRGM, further implicated a new and unexpected pathway underlying the pathophysiology of CD.[44,46] The combined analysis of three GWAS further identified 21 novel genetic susceptibility loci for CD, and in total, more than 30 loci have been identified and shown to have robust associations with the disease.

Significant discoveries have also been made in the GWAS of UC. Novel susceptibility loci have been identified for UC in Japanese population[109] but they were not detected in the other three GWAS conducted in Europeans.[110–112] Attempts to dissect the genetic basis of early-onset or pediatric-onset IBD have also yielded some success, where several previously unreported regions for IBD have been identified.[113–114] In addition, the GWAS also found associations of many loci previously implicated for adult-onset CD and UC with early-onset IBD, thus for the first time suggesting a close relationship in the patho-physiology between early- and adult-onset IBD.[114] In addition to many newly identified genes and loci, evidence showing shared or overlapping susceptibility loci among the autoimmune diseases has also been growing. For example, 6q23 locus was associated with both rheumatoid arthritis and systemic lupus erythematosus, whereas CD40 locus was identified for rheumatoid arthritis and multiple sclerosis. There are also some overlapping loci between CD and UC. For each of the aforementioned diseases, tens of new genetic loci or genes were found and several immuno-pathogenesis pathways have also been highlighted from the GWAS findings. Indeed, significant progress has been achieved in the genetic studies of these immune-related diseases.[103–108]

Although PD is one of the first few complex diseases interrogated by GWAS in 2005 and 2006, no significant discoveries about it have been made. Nevertheless, results from two recent GWAS published at the end of 2009 have shed some new light on the genetic risk factors for this complex disease where a number of novel loci or genes have been identified; for example, PARK16, BST1 and MAPT.[115,116] In addition to the new discoveries, both studies also found strong associations of common SNPs in the SNCA and LRRK2 loci with sporadic PD, where these loci were previously implicated in the familial form of the disease. These results further support the hypothesis that genes harboring rare dominant mutations for the familial or Mendelian form of a disease can also contain common genetic variants associated with the sporadic or complex form of the disease. The finding of a common SNP in TCF2 associated with T2D is another example supporting the notion.[117] Thus, examining common SNPs in the genes implicated for Mendelian diseases has opened up a new avenue for finding additional genetic risk variants for complex diseases.

These GWAS were carried out in Japanese and European populations and their data were shared and exchanged for replication and validation purposes before publication. One additional advantage of sharing the data and conducting the studies in different populations is that risk alleles that are shared and that are unique to distinct populations can be identified. For example, the Japanese GWAS uncovered two novel loci for PD, namely, PARK16, and BST1, but only the association of PARK16 was replicated in the Europeans. Similarly, the association of MAPT was only found in the European GWAS.[115,116] Although these comparisons have revealed some preliminary notable differences in the associations of genetic risk factors for PD between European and Japanese populations, inadequate statistical powers due to low allele frequencies could also lead to non-replication in either population. Therefore, further investigation is warranted to determine whether it is a genuine population

difference or an issue of statistical power. The finding of differences in association with risk alleles in different populations is not limited to PD, as the finding of association of KCNQ1 with T2D in the Japanese but not in the European population illustrates the same scenario.[37,38] The success of finding novel Parkinson's risk alleles is at least attributed to the well-powered GWAS where both the studies have genotyped a much larger sample size of several thousands in the initial screening and have also achieved robust replication, compared with the earlier GWAS of PD.

Besides PD, some significant progression has also been attained for other common neurological diseases, namely, Alzheimer's disease and schizophrenia. Although more than 10 GWAS of Alzheimer's disease have been carried out, most of them were limited by small sample sizes with inadequate or no replication. As a result, no additional genetic variants or loci with robust associations have been identified besides confirming the well-known gene for Alzheimer's disease, which is APOE.[118–120] The sample size limitation and replication issue have been overcome by two GWAS carried out by Harold *et al.*[121] and Lambert *et al.*,[122] in which a total of more than 14 000 samples were included in the initial and replication stages.[121,122] These GWAS provided firm statistical evidence for the association of three new loci for Alzheimer's disease. One locus (CLU) was identified by both the studies and the other two loci were only detected in one of the studies, which are PICALM[121] and CR1[122] but with suggestive evidence of association from the other study. The consensus gene identified by both studies is an excellent biological functional candidate, where it is also found in cerebrospinal fluid and amyloid plaques. On the other hand, for schizophrenia, the association with SNPs located in the major histocompatibility complex region was first demonstrated consistently by three GWAS implicating an immune component in the patho-physiology of the disease.[123–125] Several studies of genome-wide associations of structural variations and CNVs with schizophrenia have been quite successful and have also yielded some interesting and insightful findings about the genetic architecture of this complex disease,[126] but they are not reviewed here.

WHAT WE HAVE LEARNED

GWAS have progressed rapidly in the past 3 years and have generated a substantial amount of new information for various human complex diseases and traits. In this section, we will discuss what we have learned so far from the results of GWAS.

Disease-associated SNPs

GWAS is a comprehensive and biologically agnostic approach in searching for unknown disease-associated variants, and as demonstrated in more than 450 GWAS, this method has performed well in identifying novel genetic loci for many human complex diseases. Most of the identified genes or genetic loci were not previously thought to be associated with the diseases. More importantly, the findings have already started providing new insights into the biological pathways of several complex diseases even when most of the disease causative variants remain to be discerned from the correlated markers in the regions. For example, the three novel genes that are linked to CD, IL23R, ATG16L1 and IRGM, have highlighted the importance of the IL23R and autophagy pathways underlying the pathophysiology of this chronic inflammatory disease. Several additional genes or loci have also been found such as 5p13.1 (PTGER4), PHOX2B, FAM92B and NCF4, among others. Similar to many other complex diseases, limited success has been achieved in the past for CD.[24,127]

The majority of the risk alleles are common (allele frequency $>5\%$) and confer small effect sizes (OR <1.5). However, this observation may not really reflect the true allelic frequency spectrum of complex diseases. This is because for any given sample size, common SNPs are easier to be detected in association studies due to their higher statistical power than the rarer SNPs. In addition, the lower frequency SNPs (allele frequency $<5\%$) are not well-covered either directly or indirectly through LD by the markers in Illumina and Affymetrix genotyping arrays. As a result, they remain unexplored for disease association. The design of GWAS and SNPs selection approach in genotyping arrays have been largely influenced by the CD/CV model.[5] Common alleles with large effect sizes are scarce; so far, only a handful of them have been discovered by GWAS for the following diseases: AMD (CFH), exfoliation glaucoma (LOXL1), CD (IL23R) and testicular germ cell tumor (KIITG).

Only a small number of the risk alleles are non-synonymous SNPs in exons, which could alter the protein structure and function; for example, the non-synonymous SNPs in IL23R and ATG16L1 for CD, and SLC30A8 for T2D.[23,49] These SNPs are likely to be the functional ones, but they could also be the surrogate markers tagging for disease functional variants located outside the coding regions. In fact, most of the SNPs are located in either intron, intergenic or gene desert regions. These SNPs could still be functional because their locations may coincide with some regulatory elements, either already known or yet to be characterized, such as enhancers, insulators, transcription factor binding sites and sequences encoding for microRNAs. The association of the SNP rs6983267 in the 8q24 locus with colorectal and prostate cancer has been a mystery since its discovery, because the risk allele is located in a gene desert, which is >300 kb away from the nearest annotated gene (MYC gene). Fortunately, the mystery has been recently unveiled by two studies showing that the region containing the risk allele is a transcriptional enhancer that interacts with the MYC proto-oncogene.[128–129] This has opened up a new area of research into disease pathogenesis and also provided evidence supporting the biological functional roles of genetic variants in non-coding regions.

However, because of the indirect study design of GWAS, and the SNPs selection being guided by LD information and not functionality, it is more likely that the GWAS identified SNPs are only the surrogate markers tagging for functional variants. Unlike non-synonymous SNPs, the biological effects of these SNPs are ambiguous and not directly clear, although most of them are suspected to be involved in modulating gene or transcript expression levels. For example, the SNPs associated with asthma and CD altered the expression levels of ORMDL3 and PTGER4, respectively. Therefore, mapping of the GWAS-identified SNPs to eQTL data is an appealing approach to first identify the alleles that have regulatory roles on transcript expression levels; this could help in narrowing down a set of genes to elucidate the possible biological pathways of the disease.[130]

Heritability and disease risk prediction

Owing to the small effect sizes, collectively the SNPs only explained a small portion of the total inherited risk or heritability for any one disease. For example, all the SNPs that have been identified for T2D cumulatively only account for $\sim 5\%$ of the inherited risk, and for CD that is about 10%, although a relatively large number of confirmed SNPs or loci (about 30) was identified for CD compared with other diseases.[131,132] Furthermore, the risk alleles have also been shown to have limited predictive value for individual disease risk, for example, for breast cancer and prostate cancer.[133,134] The issues of unexplained or missing heritability and poor disease risk prediction have been getting considerable attention from the genetics community, leading to the skepticism of the promise of the GWAS approach to fully

decipher the genetic basis of complex diseases. These issues have been discussed and debated in several perspective papers.[135–137]

There are several caveats to the missing heritability and disease risk prediction. As GWAS is an indirect approach, the identified SNPs are less likely to be the disease causative variants, but are instead in strong LD with them. As a result, the effect sizes could be underestimated and further identification of the disease variants might explain more of the heritability. Moreover, the identified SNPs only represent a small fraction of the total number of genetic variants for any one disease because most of the inherited risk is not yet explained. It is almost certain that additional disease variants would be identified in the near future, and when more variants are identified, the prediction power will be improved. Further adding to the complexity are gene–gene and gene–environment interactions, which remains largely unexplored at the genome-wide scale; therefore, how much of the inherited risk could be attributed to these factors are still unclear.

Statistical power and meta-analysis

It has been shown that most of the individual GWAS have limited statistical power to detect common alleles with small effect sizes (OR <1.2). As a result, a number of GWAS meta-analyses have been carried out to boost the power by combining the data sets and have successfully identified additional loci for various diseases. Therefore, more meta-analyses are expected to be performed in the coming years. For example, the DIAGRAM Consortium identified an additional six loci for T2D by combining three GWAS data sets.[68] Likewise, an additional of about 20 loci were identified for CD and T1D from each meta-analysis.[69,71] By and large, the additional SNPs identified from these GWAS meta-analyses have had equivalent or smaller effect sizes, though with some exceptions; for example, the risk allele found in the region containing LRRK2 and MUC19 for CD has a larger OR of $\sim$1.5 and the frequency in controls was reported as 0.017.[69] The meta-analysis of multiple sclerosis also found an uncommon risk allele (a frequency of 0.02) with an OR of 1.6.[72] These findings have shown that not only would meta-analysis enhance the statistical power to identify common SNPs with smaller effect sizes, but rarer variants with stronger effect sizes could also be detected if they are included in the analysis. Thus, it is clear that by increasing the sample sizes, through conducting larger studies or combining the existing GWAS data sets in meta-analyses, many more SNPs or loci can be found.

Pleiotropy and shared disease loci

Evidence of shared genetic susceptibility among different diseases has been increasing. For example, CD and UC are two clinically distinct subtypes of IBD with some shared genes or genetic loci.[127] This suggests that some common causal pathways could underlie the diseases. Therefore, it would be an appealing approach to combine the GWAS data sets of CD and UC into a meta-analysis; this attempt should enhance the statistical power to detect the shared disease loci. This approach is not restricted to these two diseases, but should be applicable for other diseases as well, such as combining the GWAS data set of several immune-related diseases, as the evidence of shared genetics and pathogenesis for autoimmune and inflammatory diseases have been well established.[138] This approach should also be workable for the GWAS of cancers, as studies have shown that the 8q24 locus was linked to several cancers[59] and the sequence variants at the TERT-CLPTM1L locus were also found to be associated with a number of cancers.[92] SNPs in PSCA gene were also found to be associated with diffuse-type gastric cancer and urinary bladder cancer.[139,140] Likewise for neuropsychiatric diseases, the evidence of shared genetic liability for schizophrenia and bipolar disorder has also been increasing.[125] Thus, joint analysis of the GWAS data sets of the related diseases could potentially identify more genetic loci for those diseases.

Extending GWAS to different populations

Most of the GWAS have been undertaken in European populations for various diseases and traits, and not many GWAS have been carried out in other populations. One intriguing question is whether more SNPs would be uncovered if GWAS were to be performed in Asian or African populations for the same diseases? The answer is apparent from the discovery of the new T2D gene (KCNQ1) by two GWAS conducted in Japanese population, and the association was also replicated in other Asian and European populations.[37,38] These studies have underscored the importance and value of extending GWAS to different populations. Interestingly, the KCNQ1 gene was not unveiled by the previous European T2D GWAS and even by meta-analysis. It was likely because of a marked difference in allelic frequency, which resulted in lower statistical power to detect the association in European populations. In fact, one study has shown a wide variation in allele frequencies across different populations for the SNPs identified by GWAS for several complex diseases and traits.[141] A new risk allele for breast cancer was also identified by a GWAS carried out in a Chinese population and this was replicated in women of European ancestry. As in the T2D case, the risk allele has gone undetected by several European GWAS of breast cancer.[95] Results from a GWAS of systemic lupus erythematosus also support the presence of genetic heterogeneity of systemic lupus erythematosus susceptibility between Han Chinese and European populations.[142]

Two GWAS of psoriasis further support this notion of extending the GWAS in non-European populations. Besides the two well-known psoriasis loci (the major histocompatibility complex region and IL12B gene), which were found by both studies, Zhang *et al.*[143] found a new locus within the LCE gene cluster on 1q21 in a Chinese population that was not detected by the GWAS in European individuals, and several new psoriasis loci were found by the European GWAS that were not seen by the other study. Nevertheless, one has to keep in mind that the different results could also be attributed to some differences in the definition of cases and controls in each study. The findings of these studies implicated different pathways; one highlighted the involvement of IL23 and nuclear factor-κB pathways, whereas the other study indicated the importance of epidermal differentiation process in the patho-physiology of psoriasis.[143,144] In any case, it is of merit to conduct GWAS in different populations for the same disease to examine the extent of population genetic heterogeneity.

Human genetic variation

Although GWAS is progressing, our understanding of human genetic variation is also improving. In addition to the several million SNPs, there is an abundance of other types of genetic variation. So far, seven human diploid genomes have been fully sequenced, and some important insights have been gained from these resequencing studies.[145–151] The most prominent finding from these studies is that, besides SNPs, other genetic variants are also abundant in the human genome. These studies found that in addition to the 3–4 million SNPs, several hundred-thousand of short indels (for example, sizes defined as 3 and 16 bp or less in the Bentley *et al.*[147] and Wang *et al.*[148] study, respectively) are also present in each individual human genome. A large number of novel SNPs were also reported in each study. Furthermore, a few thousand of CNVs and structural variants were also found. Although the numbers reported in these resequencing

studies are nowhere close to the total number of 'non-SNP' genetic variants present in the human genome, the 1000 Genomes Project should provide a comprehensive map of all the genetic variants for a wide spectrum of frequencies ranging from rare to common upon its completion.[152] The availability of these resources and maps will certainly drive the technological development of new microarrays or methods to capture the non-SNP variants in the near future, bringing another revolution to the genetic studies of complex diseases. The non-SNP genetic variants refer to short indels, tandem repeat polymorphisms, CNVs and structural variations (that is, genetic variants other than SNPs).

It is probably unlikely that the number of non-SNP variants will reach several millions similar to the SNPs, but the total number of nucleotides encompassed by the non-SNP variants has so far exceeded that of the SNPs.[153] Given the abundance of these non-SNP variants, their total nucleotide composition, and their functional impact on gene expression levels,[154–156] they could potentially account for some or even a substantial portion of the heritability of complex diseases. Evidence has shown that some of the non-SNP variants (for example, short indels and CNVs) could be tagged by SNPs through LD.[157–159] This suggests that a fraction of non-SNP variants could have already been interrogated indirectly in the GWAS. This was well illustrated in the discovery of the 20-kb deletion located immediately upstream of the IRGM gene for CD. The deletion was in perfect LD with the most strongly associated SNP that was initially identified through GWAS using a commercial genotyping array.[160] In fact, a recent study also confirmed that most of the common CNVs are well tagged by SNPs, suggesting that the existing GWAS have indirectly interrogated the association of common CNVs with complex diseases relatively well.[161] Nevertheless, as the non-SNP variants that have been discovered so far are only a fraction of the total number, there are still many uncertainties remaining about the LD. Therefore, when a more complete map or catalog of non-SNP variants is available in the future, more studies will be needed to further interrogate the LD relationship between them and SNPs. This will help to determine and discern the fraction of non-SNP variants that need to be assayed directly. Again, the completion of the 1000 Genomes Project should facilitate the studies in this area.

Non-SNP variants and complex diseases

Although the roles of non-SNP variants in disease susceptibility remain largely unexplored, associations of CNVs with complex diseases such as autoimmune disorders, HIV infection, schizophrenia, and autism have already been established from both candidate gene and genome-wide approaches.[162–166] In addition, one study has shown the correlations of about 30 SNPs (that found to be associated with various traits by GWAS) with CNVs at $r^2 > 0.5$, this provides preliminary evidence of the associations and possible roles of CNVs in human complex traits.[161] The amount of evidence is expected to increase in the near future, when we have a better understanding of the characteristics of non-SNP variants and a more comprehensive map of them constructed upon the completion of the 1000 Genomes Project, and when more efficient and accurate methods are available to detect the non-SNP variants for disease-association studies. It is crucial to remember that the current GWAS using the commercial genotyping arrays cover only a portion of the total genetic variants, thus a substantial false-negative rate or a significant portion of missing heritability could be attributed to incomplete interrogation of all the genetic variants for disease association.[132,167] For future studies, the focus should be directed on studying other genetic variants that have not yet been interrogated by the GWAS, although it is highly dependent on the development of the technologies and methods of detection and analysis.

Rare variants

It is also obvious from the results of the GWAS that the common SNPs are unable to account for the total inherited risk of a complex disease. But it is not clear at this stage how much of the heritability can be attributed to uncommon SNPs or rare mutations (frequency <1–5%). Uncommon SNPs are not well covered by the commercial genotyping arrays, as a result they have not been intensively studied for disease association. Fortunately, the current genotyping arrays seem to work fine for detecting rare CNVs for diseases.[164,168] The evidence linking complex diseases and traits to multiple rare variants has been growing; for example, for schizophrenia,[164,168] high-density lipoprotein cholesterol level and T1D.[169–171] This suggests that rare variants (both SNPs and non-SNP) should not be ignored in the future studies. Sequencing approaches will improve their detection, and consequently offer a better understanding of the genetic architecture of complex diseases. Although waiting for whole genome sequencing to be feasible (as currently the cost is still prohibitively expensive to sequence a large sample set and there are still many technical problems and challenges associated with the sequencing technologies), one can start with a targeted resequencing approach of the regions identified by GWAS and linkage studies, exome resequencing and resequencing of biologically plausible candidate genes.[171,172] The advances in sequencing technologies will enable researchers to study a wider spectrum of genetic variants compared with genotyping methods.

Genetic architecture of complex diseases

The genetic architecture of complex diseases still remains elusive; it is unclear how much each type of genetic variant contributes to the total inherited risk and what is the relative proportion of rare versus common causal variants. If non-SNP variants or rare SNPs/mutations constitute most of the genetic component of complex diseases, then the GWAS using the current genotyping arrays would likely overlook them, because they are not covered directly by the genotyping arrays and how much they can be tagged through LD by the markers on the arrays still needs to be determined.

CONCLUSION

Five years ago, researchers were uncertain whether the genome-wide association approach would be workable, the answer is apparent after more than 450 GWAS have been carried out, which have identified an enormous number of novel disease-associated SNPs. Currently, the question is whether the GWAS approach would be able to identify all the genetic variants for complex diseases. The answer is clearly no if GWAS continue to study disease association using the current commercial genotyping arrays, which mainly targets common SNPs. It is obvious from the results of GWAS that common SNPs only constitute a small portion of the total inherited risk of complex diseases. This is not to say that the GWAS approach would not be able to identify most of the disease genetic variants *per se*, but that the success of GWAS is highly dependent on the amount of genetic variants interrogated in the study, because the locations of the disease variants are unknown. The current GWAS that focuses primarily on common SNPs only interrogates a subset of the total genetic variants. Therefore, casting a wider net is important, the disease variants have to be directly studied or the markers that are in strong LD with them must be included in order to allow detection of the disease variants.

However, the ability to interrogate more genetic variants is reliant on the availability of technologies and methods to assay them (both SNP and non-SNP genetic variants). The completion of the 1000 Genomes Project should accelerate the research and development in these areas.

1 Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J. Y., Sackler, R. S., Haynes, C. *et al.* Complement factor H polymorphism in age-related macular degeneration. *Science* **308,** 385–389 (2005).
2 Altshuler, D., Daly, M. J. & Lander, E. S. Genetic mapping in human disease. *Science* **322,** 881–888 (2008).
3 Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl Acad. Sci. USA* **106,** 9362–9367 (2009).
4 Ragoussis, J. Genotyping technologies for genetic research. *Annu. Rev. Genomics Hum. Genet.* **10,** 117–133 (2009).
5 Schork, N. J., Murray, S. S., Frazer, K. A. & Topol, E. J. Common vs rare allele hypotheses for complex diseases. *Curr. Opin. Genet. Dev.* **19,** 212–219(2009).
6 Ozaki, K., Ohnishi, Y., Iida, A., Sekine, A., Yamada, R., Tsunoda, T. *et al.* Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction. *Nat. Genet.* **32,** 650–654 (2002).
7 Obara, W., Iida, A., Suzuki, Y., Tanaka, T., Akiyama, F., Maeda, S. *et al.* Association of single-nucleotide polymorphisms in the polymeric immunoglobulin receptor gene with immunoglobulin A nephropathy (IgAN) in Japanese patients. *J. Hum. Genet.* **48,** 293–299 (2003).
8 Tanaka, N., Babazono, T., Saito, S., Sekine, A., Tsunoda, T., Haneda, M. *et al.* 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 (2003).
9 Shimazaki, A., Kawamura, Y., Kanazawa, A., Sekine, A., Saito, S., Tsunoda, T. *et al.* Genetic variations in the gene encoding ELMO1 are associated with susceptibility to diabetic nephropathy. *Diabetes* **54,** 1171–1178 (2005).
10 Yamazaki, K., McGovern, D., Ragoussis, J., Paolucci, M., Butler, H., Jewell, D. *et al.* Single nucleotide polymorphisms in TNFSF15 confer susceptibility to Crohn's disease. *Hum. Mol. Genet.* **14,** 3499–3506 (2005).
11 Mototani, H., Mabuchi, A., Saito, S., Fujioka, M., Iida, A., Takatori, Y. *et al.* A functional single nucleotide polymorphism in the core promoter region of CALM1 is associated with hip osteoarthritis in Japanese. *Hum. Mol. Genet.* **14,** 1009–1017 (2005).
12 Ohnishi, Y., Tanaka, T., Ozaki, K., Yamada, R., Suzuki, H. & Nakamura, Y. A high-throughput SNP typing system for genome-wide association studies. *J. Hum. Genet.* **46,** 471–477 (2001).
13 Haga, H., Yamada, R., Ohnishi, Y., Nakamura, Y. & Tanaka, T. Gene-based SNP discovery as part of the Japanese Millennium Genome Project: identification of 190 562 genetic variations in the human genome. Single-nucleotide polymorphism. *J. Hum. Genet.* **47,** 605–610 (2002).
14 Hirakawa, M., Tanaka, T., Hashimoto, Y., Kuroda, M., Takagi, T. & Nakamura, Y. JSNP: a database of common gene variations in the Japanese population. *Nucleic Acids Res.* **30,** 158–162 (2002).
15 International HapMap Consortium. A haplotype map of the human genome. *Nature* **437,** 1299–1320 (2005).
16 International HapMap ConsortiumFrazer, K. A. Ballinger, D. G. Cox, D. R. Hinds, D. A., Stuve, L. L. *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449,** 851–861 (2007).
17 Barrett, J. C. & Cardon, L. R. Evaluating coverage of genome-wide association studies. *Nat. Genet.* **38,** 659–662 (2006).
18 Eberle, M. A., Ng, P. C., Kuhn, K., Zhou, L., Peiffer, D. A., Galver, L. *et al.* Power to detect risk alleles using genome-wide tag SNP panels. *PLoS Genet.* **3,** 1827–1837 (2007).
19 Li, M., Li, C. & Guan, W. Evaluation of coverage variation of SNP chips for genome-wide association studies. *Eur. J. Hum. Genet.* **16,** 635–643 (2008).
20 Dewan, A., Liu, M., Hartman, S., Zhang, S. S., Liu, D. T., Zhao, C. *et al.* HTRA1 promoter polymorphism in wet age-related macular degeneration. *Science* **314,** 989–992 (2006).
21 Thakkinstian, A., Han, P., McEvoy, M., Smith, W., Hoh, J., Magnusson, K. *et al.* Systematic review and meta-analysis of the association between complement factor H Y402H polymorphisms and age-related macular degeneration. *Hum. Mol. Genet.* **15,** 2784–2790 (2006).
22 Chen, W., Xu, W., Tao, Q., Liu, J., Li, X., Gan, X. *et al.* Meta-analysis of the association of the HTRA1 polymorphisms with the risk of age-related macular degeneration. *Exp. Eye Res.* **89,** 292–300 (2009).
23 Duerr, R. H., Taylor, K. D., Brant, S. R., Rioux, J. D., Silverberg, M. S., Daly, M. J. *et al.* A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. *Science* **314,** 1461–1463 (2006).
24 Mathew, C. G. New links to the pathogenesis of Crohn disease provided by genome-wide association scans. *Nat. Rev. Genet.* **9,** 9–14 (2008).
25 Herbert, A., Gerry, N. P., McQueen, M. B., Heid, I. M., Pfeufer, A., Illig, T. *et al.* A common genetic variant is associated with adult and childhood obesity. *Science* **312,** 279–283 (2006).
26 Maraganore, D. M., de Andrade, M., Lesnick, T. G., Strain, K. J., Farrer, M. J., Rocca, W. A. *et al.* High resolution whole-genome association study of Parkinson disease. *Am. J. Hum. Genet.* **77,** 685–693 (2005).
27 Lyon, H. N., Emilsson, V., Hinney, A., Heid, I. M., Lasky-Su, J., Zhu, X. *et al.* The association of a SNP upstream of INSIG2 with body mass index is reproduced in several but not all cohorts. *PLoS Genet.* **3,** e61 (2007).
28 Heid, I. M., Huth, C., Loos, R. J., Kronenberg, F., Adamkova, V., Anand, S. S. *et al.* Meta-analysis of the INSIG2 association with obesity including 74 345 individuals: does heterogeneity of estimates relate to study design? *PLoS Genet.* **5,** e1000694 (2009).
29 Elbaz, A., Nelson, L. M., Payami, H., Ioannidis, J. P., Fiske, B. K., Annesi, G. *et al.* Lack of replication of thirteen single-nucleotide polymorphisms implicated in Parkinson's disease: a large-scale international study. *Lancet Neurol.* **5,** 917–923 (2006).
30 Fung, H. C., Scholz, S., Matarin, M., Simón-Sánchez, J., Hernandez, D., Britton, A. *et al.* Genome-wide genotyping in Parkinson's disease and neurologically normal controls: first stage analysis and public release of data. *Lancet Neurol.* **5,** 911–916 (2006).
31 Wellcome Trust Case Control Consortium. Genome-wide association study of 14 000 cases of seven common diseases and 3000 shared controls. *Nature* **447,** 661–678 (2007).
32 GAIN Collaborative Research Group. New models of collaboration in genome-wide association studies: the Genetic Association Information Network. *Nat. Genet.* **39,** 1045–1051 (2007).
33 Li, Y., Willer, C., Sanna, S. & Abecasis, G. Genotype imputation. *Annu. Rev. Genomics Hum. Genet.* **10,** 387–406 (2009).
34 Arking, D. E., Pfeufer, A., Post, W., Kao, W. H., Newton-Cheh, C., Ikeda, M. *et al.* A common genetic variant in the NOS1 regulator NOS1AP modulates cardiac repolarization. *Nat. Genet.* **38,** 644–651 (2006).
35 Newton-Cheh, C., Eijgelsheim, M., Rice, K. M., de Bakker, P. I., Yin, X., Estrada, K. *et al.* Common variants at ten loci influence QT interval duration in the QTGEN Study. *Nat. Genet.* **41,** 399–406 (2009).
36 Pfeufer, A., Sanna, S., Arking, D. E., Müller, M., Gateva, V., Fuchsberger, C. *et al.* Common variants at ten loci modulate the QT interval duration in the QTSCD Study. *Nat. Genet.* **41,** 407–414 (2009).
37 Unoki, H., Takahashi, A., Kawaguchi, T., Hara, K., Horikoshi, M., Andersen, G. *et al.* SNPs in KCNQ1 are associated with susceptibility to type 2 diabetes in East Asian and European populations. *Nat. Genet.* **40,** 1098–1102 (2008).
38 Yasuda, K., Miyake, K., Horikawa, Y., Hara, K., Osawa, H., Furuta, H. *et al.* Variants in KCNQ1 are associated with susceptibility to type 2 diabetes mellitus. *Nat. Genet.* **40,** 1092–1097 (2008).
39 McPherson, R., Pertsemlidis, A., Kavaslar, N., Stewart, A., Roberts, R., Cox, D. R. *et al.* A common allele on chromosome 9 associated with coronary heart disease. *Science* **316,** 1488–1491 (2007).
40 Helgadottir, A., Thorleifsson, G., Manolescu, A., Gretarsdottir, S., Blondal, T., Jonasdottir, A. *et al.* A common variant on chromosome 9p21 affects the risk of myocardial infarction. *Science* **316,** 1491–1493 (2007).
41 Zeggini, E., Weedon, M. N., Lindgren, C. M., Frayling, T. M., Elliott, K. S., Lango, H. *et al.* Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science* **316,** 1336–1341 (2007).
42 Helgadottir, A., Thorleifsson, G., Magnusson, K. P., Grétarsdottir, S., Steinthorsdottir, V., Manolescu, A. *et al.* The same sequence variant on 9p21 associates with myocardial infarction, abdominal aortic aneurysm and intracranial aneurysm. *Nat. Genet.* **40,** 217–224 (2008).
43 Pennisi, E. Breakthrough of the year. Human genetic variation. *Science* **318,** 1842–1843 (2007).
44 Parkes, M., Barrett, J. C., Prescott, N. J., Tremelling, M., Anderson, C. A., Fisher, S. A. *et al.* Sequence variants in the autophagy gene IRGM and multiple other replicating loci contribute to Crohn's disease susceptibility. *Nat. Genet.* **39,** 830–832 (2007).
45 Todd, J. A., Walker, N. M., Cooper, J. D., Smyth, D. J., Downes, K., Plagnol, V. *et al.* Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. *Nat. Genet.* **39,** 857–864 (2007).
46 Rioux, J. D., Xavier, R. J., Taylor, K. D., Silverberg, M. S., Goyette, P., Huett, A. *et al.* Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis. *Nat. Genet.* **39,** 596–604 (2007).
47 Easton, D. F., Pooley, K. A., Dunning, A. M., Pharoah, P. D., Thompson, D., Ballinger, D. G. *et al.* Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature* **447,** 1087–1093 (2007).
48 Hunter, D. J., Kraft, P., Jacobs, K. B., Cox, D. G., Yeager, M., Hankinson, S. E. *et al.* A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. *Nat. Genet.* **39,** 870–874 (2007).
49 Sladek, R., Rocheleau, G., Rung, J., Dina, C., Shen, L., Serre, D. *et al.* A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* **445,** 881–885 (2007).
50 Zeggini, E., Weedon, M. N., Lindgren, C. M., Frayling, T. M., Elliott, K. S., Lango, H. *et al.* Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science* **316,** 1336–1341 (2007).
51 Saxena, R., Voight, B. F., Lyssenko, V., Burtt, N. P., de Bakker, P. I., Chen, H. *et al.* Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* **316,** 1331–1336 (2007).
52 Scott, L. J., Mohlke, K. L., Bonnycastle, L. L., Willer, C. J., Li, Y., Duren, W. L. *et al.* A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science* **316,** 1341–1345 (2007).

53 Steinthorsdottir, V., Thorleifsson, G., Reynisdottir, I., Benediktsson, R., Jonsdottir, T., Walters, G. B. *et al.* A variant in CDKAL1 influences insulin response and risk of type 2 diabetes. *Nat. Genet.* **39,** 770–775 (2007).
54 Prokopenko, I., McCarthy, M. I. & Lindgren, C. M. Type 2 diabetes: new genes, new understanding. *Trends Genet.* **24,** 613–621 (2008).
55 Zanke, B. W., Greenwood, C. M., Rangrej, J., Kustra, R., Tenesa, A., Farrington, S. M. *et al.* Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24. *Nat. Genet.* **39,** 989–994 (2007).
56 Tomlinson, I., Webb, E., Carvajal-Carmona, L., Broderick, P., Kemp, Z., Spain, S. *et al.* A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. *Nat. Genet.* **39,** 984–988 (2007).
57 Gudmundsson, J., Sulem, P., Manolescu, A., Amundadottir, L. T., Gudbjartsson, D., Helgason, A. *et al.* Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24. *Nat. Genet.* **39,** 631–637 (2007).
58 Yeager, M., Orr, N., Hayes, R. B., Jacobs, K. B., Kraft, P., Wacholder, S. *et al.* Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nat. Genet.* **39,** 645–649 (2007).
59 Easton, D. F. & Eeles, R. A. Genome-wide association studies in cancer. *Hum. Mol. Genet.* **17,** R109–R115.
60 Thorleifsson, G., Magnusson, K. P., Sulem, P., Walters, G. B., Gudbjartsson, D. F., Stefansson, H. *et al.* Common sequence variants in the LOXL1 gene confer susceptibility to exfoliation glaucoma. *Science* **317,** 1397–1400 (2007).
61 NCI-NHGRI Working Group on Replication in Association StudiesChanock, S. J. Manolio, T. Boehnke, M. Boerwinkle, E., Hunter, D. J. *et al.* Replicating genotype-phenotype associations. *Nature* **447,** 655–660 (2007).
62 Moffatt, M. F., Kabesch, M., Liang, L., Dixon, A. L., Strachan, D., Heath, S. *et al.* Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. *Nature* **448,** 470–473 (2007).
63 Libioulle, C., Louis, E., Hansoul, S., Sandor, C., Farnir, F., Franchimont, D. *et al.* Novel Crohn disease locus identified by genome-wide association maps to a gene desert on 5p13.1 and modulates expression of PTGER4. *PLoS Genet.* **3,** e58 (2007).
64 Florez, J. C. Clinical review: the genetics of type 2 diabetes: a realistic appraisal in 2008. *J. Clin. Endocrinol. Metab.* **93,** 4633–4642 (2008).
65 de Bakker, P. I., Ferreira, M. A., Jia, X., Neale, B. M., Raychaudhuri, S. & Voight, B. F. Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Hum. Mol. Genet.* **17,** R122–R128 (2008).
66 Ioannidis, J. P., Thomas, G. & Daly, M. J. Validating, augmenting and refining genome-wide association signals. *Nat. Rev. Genet.* **10,** 318–329 (2009).
67 Zeggini, E. & Ioannidis, J. P. Meta-analysis in genome-wide association studies. *Pharmacogenomics* **10,** 191–201 (2009).
68 Zeggini, E., Scott, L. J., Saxena, R., Voight, B. F., Marchini, J. L., Hu, T. *et al.* Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat. Genet.* **40,** 638–645 (2008).
69 Barrett, J. C., Hansoul, S., Nicolae, D. L., Cho, J. H., Duerr, R. H., Rioux, J. D. *et al.* Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat. Genet.* **40,** 955–962 (2008).
70 Cooper, J. D., Smyth, D. J., Smiles, A. M., Plagnol, V., Walker, N. M., Allen, J. E. *et al.* Meta-analysis of genome-wide association study data identifies additional type 1 diabetes risk loci. *Nat. Genet.* **40,** 1399–1401 (2008).
71 Barrett, J. C., Clayton, D. G., Concannon, P., Akolkar, B., Cooper, J. D., Erlich, H. A. *et al.* Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nat. Genet.* **41,** 703–707 (2009).
72 De Jager, P. L., Jia, X., Wang, J., de Bakker, P. I., Ottoboni, L., Aggarwal, N. T. *et al.* Meta-analysis of genome scans and replication identify CD6, IRF8 and TNFRSF1A as new multiple sclerosis susceptibility loci. *Nat. Genet.* **41,** 776–782 (2009).
73 Raychaudhuri, S., Remmers, E. F., Lee, A. T., Hackett, R., Guiducci, C., Burtt, N. P. *et al.* Common variants at CD40 and other loci confer risk of rheumatoid arthritis. *Nat. Genet.* **40,** 1216–1223 (2008).
74 Houlston, R. S., Webb, E., Broderick, P., Pittman, A. M., Di Bernardo, M. C., Lubbe, S. *et al.* Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. *Nat. Genet.* **40,** 1426–1435 (2008).
75 Wang, Y., Broderick, P., Webb, E., Wu, X., Vijayakrishnan, J., Matakidou, A. *et al.* Common 5p15.33 and 6p21.33 variants influence lung cancer risk. *Nat. Genet.* **40,** 1407–1409 (2008).
76 Thomas, G., Jacobs, K. B., Yeager, M., Kraft, P., Wacholder, S., Orr, N. *et al.* Multiple loci identified in a genome-wide association study of prostate cancer. *Nat. Genet.* **40,** 310–315 (2008).
77 Gudmundsson, J., Sulem, P., Rafnar, T., Bergthorsson, J. T., Manolescu, A., Gudbjartsson, D. *et al.* Common sequence variants on 2p15 and Xp11.22 confer susceptibility to prostate cancer. *Nat. Genet.* **40,** 281–283 (2008).
78 Eeles, R. A., Kote-Jarai, Z., Giles, G. G., Olama, A. A., Guy, M., Jugurnauth, S. K. *et al.* Multiple newly identified loci associated with prostate cancer susceptibility. *Nat. Genet.* **40,** 316–321 (2008).
79 Gudmundsson, J., Sulem, P., Gudbjartsson, D. F., Blondal, T., Gylfason, A., Agnarsson, B. A. *et al.* Genome-wide association and replication studies identify four variants associated with prostate cancer susceptibility. *Nat. Genet.* **41,** 1122–1126 (2009).
80 Gudmundsson, J., Sulem, P., Steinthorsdottir, V., Bergthorsson, J. T., Thorleifsson, G., Manolescu, A. *et al.* Two variants on chromosome 17 confer prostate cancer risk, and the one in TCF2 protects against type 2 diabetes. *Nat. Genet.* **39,** 977–983 (2007).
81 Shete, S., Hosking, F. J., Robertson, L. B., Dobbins, S. E., Sanson, M., Malmer, B. *et al.* Genome-wide association study identifies five susceptibility loci for glioma. *Nat. Genet.* **41,** 899–904 (2009).
82 Bishop, D. T., Demenais, F., Iles, M. M., Harland, M., Taylor, J. C., Corda, E. *et al.* Genome-wide association study identifies three loci associated with melanoma risk. *Nat. Genet.* **41,** 920–925 (2009).
83 Stacey, S. N., Sulem, P., Masson, G., Gudjonsson, S. A., Thorleifsson, G., Jakobsdottir, M. *et al.* New common variants affecting susceptibility to basal cell carcinoma. *Nat. Genet.* **41,** 909–914 (2009).
84 Eeles, R. A., Kote-Jarai, Z., Al Olama, A. A., Giles, G. G., Guy, M., Severi, G. *et al.* Identification of seven new prostate cancer susceptibility loci through a genome-wide association study. *Nat. Genet.* **41,** 1116–1121 (2009).
85 Tomlinson, I. P., Webb, E., Carvajal-Carmona, L., Broderick, P., Howarth, K., Pittman, A. M. *et al.* A genome-wide association study identifies colorectal cancer susceptibility loci on chromosomes 10p14 and 8q23.3. *Nat. Genet.* **40,** 623–630 (2008).
86 Tenesa, A., Farrington, S. M., Prendergast, J. G., Porteous, M. E., Walker, M., Haq, N. *et al.* 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 (2008).
87 Broderick, P., Carvajal-Carmona, L., Pittman, A. M., Webb, E., Howarth, K., Rowan, A. *et al.* A genome-wide association study shows that common alleles of SMAD7 influence colorectal cancer risk. *Nat. Genet.* **39,** 1315–1317 (2007).
88 Hung, R. J., McKay, J. D., Gaborieau, V., Boffetta, P., Hashibe, M., Zaridze, D. *et al.* A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. *Nature* **452,** 633–637 (2008).
89 Amos, C. I., Wu, X., Broderick, P., Gorlov, I. P., Gu, J., Eisen, T. *et al.* Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. *Nat. Genet.* **40,** 616–622 (2008).
90 Thorgeirsson, T. E., Geller, F., Sulem, P., Rafnar, T., Wiste, A., Magnusson, K. P. *et al.* A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. *Nature* **452,** 638–642 (2008).
91 McKay, J. D., Hung, R. J., Gaborieau, V., Boffetta, P., Chabrier, A., Byrnes, G. *et al.* Lung cancer susceptibility locus at 5p15.33. *Nat. Genet.* **40,** 1404–1406 (2008).
92 Rafnar, T., Sulem, P., Stacey, S. N., Geller, F., Gudmundsson, J., Sigurdsson, A. *et al.* Sequence variants at the TERT-CLPTM1 L locus associate with many cancer types. *Nat. Genet.* **41,** 221–227 (2009).
93 Chanock, S. J. & Hunter, D. J. Genomics: when the smoke clears. *Nature* **452,** 537–538 (2008).
94 Thomas, G., Jacobs, K. B., Kraft, P., Yeager, M., Wacholder, S., Cox, D. G. *et al.* A multistage genome-wide association study in breast cancer identifies two new risk alleles at 1p11.2 and 14q24.1 (RAD51L1). *Nat. Genet.* **41,** 579–584 (2009).
95 Zheng, W., Long, J., Gao, Y. T., Li, C., Zheng, Y., Xiang, Y. B. *et al.* Genome-wide association study identifies a new breast cancer susceptibility locus at 6q25.1. *Nat. Genet.* **41,** 324–328 (2009).
96 Ong, K. K., Elks, C. E., Li, S., Zhao, J. H., Luan, J., Andersen, L. B. *et al.* Genetic variation in LIN28B is associated with the timing of puberty. *Nat. Genet.* **41,** 729–733 (2009).
97 Sulem, P., Gudbjartsson, D. F., Rafnar, T., Holm, H., Olafsdottir, E. J., Olafsdottir, G. H. *et al.* Genome-wide association study identifies sequence variants on 6q21 associated with age at menarche. *Nat. Genet.* **41,** 734–738 (2009).
98 Perry, J. R., Stolk, L., Franceschini, N., Lunetta, K. L., Zhai, G., McArdle, P. F. *et al.* Meta-analysis of genome-wide association data identifies two loci influencing age at menarche. *Nat. Genet.* **41,** 648–650 (2009).
99 Stolk, L., Zhai, G., van Meurs, J. B., Verbiest, M. M., Visser, J. A., Estrada, K. *et al.* Loci at chromosomes 13, 19 and 20 influence age at natural menopause. *Nat. Genet.* **41,** 645–647 (2009).
100 He, C., Kraft, P., Chen, C., Buring, J. E., Paré, G., Hankinson, S. E. *et al.* Genome-wide association studies identify loci associated with age at menarche and age at natural menopause. *Nat. Genet.* **41,** 724–728 (2009).
101 Rapley, E. A., Turnbull, C., Al Olama, A. A., Dermitzakis, E. T., Linger, R., Huddart, R. A. *et al.* A genome-wide association study of testicular germ cell tumor. *Nat. Genet.* **41,** 807–810 (2009).
102 Kanetsky, P. A., Mitra, N., Vardhanabhuti, S., Li, M., Vaughn, D. J., Letrero, R. *et al.* Common variation in KITLG and at 5q31.3 predisposes to testicular germ cell cancer. *Nat. Genet.* **41,** 811–815 (2009).
103 Oksenberg, J. R., Baranzini, S. E., Sawcer, S. & Hauser, S. L. The genetics of multiple sclerosis: SNPs to pathways to pathogenesis. *Nat. Rev. Genet.* **9,** 516–526 (2008).
104 Plenge, R. M. Recent progress in rheumatoid arthritis genetics: one step towards improved patient care. *Curr. Opin. Rheumatol.* **21,** 262–271 (2009).
105 Graham, R. R., Hom, G., Ortmann, W. & Behrens, T. W. Review of recent genome-wide association scans in lupus. *J. Intern. Med.* **265,** 680–688 (2009).
106 Budarf, M. L., Labbé, C., David, G. & Rioux, J. D. GWA studies: rewriting the story of IBD. *Trends Genet.* **25,** 137–146 (2009).
107 Heap, G. A. & van Heel, D. A. The genetics of chronic inflammatory diseases. *Hum. Mol. Genet.* **18,** R101–106 (2009).
108 Lettre, G. & Rioux, J. D. Autoimmune diseases: insights from genome-wide association studies. *Hum. Mol. Genet.* **17,** R116–121 (2008).
109 Asano, K., Matsushita, T., Umeno, J., Hosono, N., Takahashi, A., Kawaguchi, T. *et al.* A genome-wide association study identifies three new susceptibility loci for ulcerative colitis in the Japanese population. *Nat. Genet.* **41,** 1325–1329 (2009).
110 Franke, A., Balschun, T., Karlsen, T. H., Sventoraityte, J., Nikolaus, S., Mayr, G. *et al.* Sequence variants in IL10, ARPC2 and multiple other loci contribute to ulcerative colitis susceptibility. *Nat. Genet.* **40,** 1319–1323 (2008).
111 Silverberg, M. S., Cho, J. H., Rioux, J. D., McGovern, D. P., Wu, J., Annese, V. *et al.* Ulcerative colitis-risk loci on chromosomes 1p36 and 12q15 found by genome-wide association study. *Nat. Genet.* **41,** 216–220 (2009).

112 UK IBD Genetics ConsortiumBarrett, J. C. Lee, J. C. Lees, C. W., Prescott, N. J., Anderson, C. A. *et al.* Genome-wide association study of ulcerative colitis identifies three new susceptibility loci, including the HNF4A region. *Nat. Genet.* **41,** 1330–1334 (2009).
113 Kugathasan, S., Baldassano, R. N., Bradfield, J. P., Sleiman, P. M., Imielinski, M., Guthery, S. L. *et al.* Loci on 20q13 and 21q22 are associated with pediatric-onset inflammatory bowel disease. *Nat. Genet.* **40,** 1211–1215 (2008).
114 Imielinski, M., Baldassano, R. N., Griffiths, A., Russell, R. K., Annese, V., Dubinsky, M. *et al.* Common variants at five new loci associated with early-onset inflammatory bowel disease. *Nat. Genet.* **41,** 1335–1340 (2009).
115 Satake, W., Nakabayashi, Y., Mizuta, I., Hirota, Y., Ito, C., Kubo, M. *et al.* Genome-wide association study identifies common variants at four loci as genetic risk factors for Parkinson's disease. *Nat. Genet.* **41,** 1303–1307 (2009).
116 Simon-Sanchez, J., Schulte, C., Bras, J. M., Sharma, M., Gibbs, J. R., Berg, D. *et al.* Genome-wide association study reveals genetic risk underlying Parkinson's disease. *Nat. Genet.* **41,** 1308–1312 (2009).
117 Winckler, W., Weedon, M. N., Graham, R. R., McCarroll, S. A., Purcell, S., Almgren, P. *et al.* Evaluation of common variants in the six known maturity-onset diabetes of the young (MODY) genes for association with type 2 diabetes. *Diabetes* **56,** 685–693 (2007).
118 Coon, K. D., Myers, A. J., Craig, D. W., Webster, J. A., Pearson, J. V., Lince, D. H. *et al.* A high-density whole-genome association study reveals that APOE is the major susceptibility gene for sporadic late-onset Alzheimer's disease. *J. Clin. Psychiatry* **68,** 613–618 (2007).
119 Webster, J. A., Myers, A. J., Pearson, J. V., Craig, D. W., Hu-Lince, D., Coon, K. D. *et al.* Sorl1 as an Alzheimer's disease predisposition gene? *Neurodegener. Dis.* **5,** 60–64 (2007).
120 Li, H., Wetten, S., Li, L., St Jean, P. L., Upmanyu, R., Surh, L. *et al.* Candidate single-nucleotide polymorphisms from a genomewide association study of Alzheimer disease. *Arch. Neurol.* **65,** 45–53 (2007).
121 Harold, D., Abraham, R., Hollingworth, P., Sims, R., Gerrish, A., Hamshere, M. L. *et al.* Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease. *Nat. Genet.* **41,** 1088–1093 (2009).
122 Lambert, J. C., Heath, S., Even, G., Campion, D., Sleegers, K., Hiltunen, M. *et al.* Genome-wide association study identifies variants at CLU and CR1 associated with Alzheimer's disease. *Nat. Genet.* **41,** 1094–1099 (2009).
123 Shi, J., Levinson, D. F., Duan, J., Sanders, A. R., Zheng, Y., Pe'er, I. *et al.* Common variants on chromosome 6p22.1 are associated with schizophrenia. *Nature* **460,** 753–757 (2009).
124 Stefansson, H., Ophoff, R. A., Steinberg, S., Andreassen, O. A., Cichon, S., Rujescu, D. *et al.* Common variants conferring risk of schizophrenia. *Nature* **460,** 744–747 (2009).
125 International Schizophrenia Consortium. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460,** 748–752 (2009).
126 Owen, M. J., Williams, H. J. & O'Donovan, M. C. Schizophrenia genetics: advancing on two fronts. *Curr. Opin. Genet. Dev.* **19,** 266–270 (2009).
127 Cho, J. H. The genetics and immunopathogenesis of inflammatory bowel disease. *Nat. Rev. Immunol.* **8,** 458–466 (2008).
128 Pomerantz, M. M., Ahmadiyeh, N., Jia, L., Herman, P., Verzi, M. P., Doddapaneni, H. *et al.* The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. *Nat. Genet.* **41,** 882–884 (2009).
129 Tuupanen, S., Turunen, M., Lehtonen, R., Hallikas, O., Vanharanta, S., Kivioja, T. *et al.* The common colorectal cancer predisposition SNP rs6983267 at chromosome 8q24 confers potential to enhanced Wnt signaling. *Nat. Genet.* **41,** 885–890 (2009).
130 Cookson, W., Liang, L., Abecasis, G., Moffatt, M. & Lathrop, M. Mapping complex disease traits with global gene expression. *Nat. Rev. Genet.* **10,** 184–194 (2009).
131 Maher, B. The case of the missing heritability. *Nature* **456,** 18–21 (2008).
132 Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J. *et al.* Finding the missing heritability of complex diseases. *Nature* **461,** 747–753 (2009).
133 Pharoah, P. D., Antoniou, A. C., Easton, D. F. & Ponder, B. A. Polygenes, risk prediction, and targeted prevention of breast cancer. *N. Engl. J. Med.* **358,** 2796–2803 (2008).
134 Zheng, S. L., Sun, J., Wiklund, F., Smith, S., Stattin, P., Li, G. *et al.* Cumulative association of five genetic variants with prostate cancer. *N. Engl. J. Med.* **358,** 910–919 (2008).
135 Goldstein, D. B. Common genetic variation and human traits. *N. Engl. J. Med.* **360,** 1696–1698 (2009).
136 Hirschhorn, J. N. Genomewide association studies—illuminating biologic pathways. *N. Engl. J. Med.* **360,** 1699–1701 (2009).
137 Kraft, P. & Hunter, D. J. Genetic risk prediction—are we there yet? *N. Engl. J. Med.* **360,** 1701–1703 (2009).
138 Zhernakova, A., van Diemen, C. C. & Wijmenga, C. Detecting shared pathogenesis from the shared genetics of immune-related diseases. *Nat. Rev. Genet.* **10,** 43–55 (2009).
139 Study Group of Millennium Genome Project for CancerSakamoto, H. Yoshimura, K. Saeki, N. Katai, H. Shimoda, T., Matsuno, Y. *et al.* Genetic variation in PSCA is associated with susceptibility to diffuse-type gastric cancer. *Nat. Genet.* **40,** 730–740 (2008).
140 Wu, X., Ye, Y., Kiemeney, L. A., Sulem, P., Rafnar, T., Matullo, G. *et al.* Genetic variation in the prostate stem cell antigen gene PSCA confers susceptibility to urinary bladder cancer. *Nat. Genet.* **41,** 991–995 (2009).
141 Adeyemo, A. & Rotimi, C. Genetic variants associated with complex human diseases show wide variation across multiple populations. *Public Health Genomics* **13,** 72–79 (2010).
142 Han, J. W., Zheng, H. F., Cui, Y., Sun, L. D., Ye, D. Q., Hu, Z. *et al.* Genome-wide association study in a Chinese Han population identifies nine new susceptibility loci for systemic lupus erythematosus. *Nat. Genet.* **41,** 1234–1237 (2009).
143 Zhang, X. J., Huang, W., Yang, S., Sun, L. D., Zhang, F. Y., Zhu, Q. X. *et al.* Psoriasis genome-wide association study identifies susceptibility variants within LCE gene cluster at 1q21. *Nat. Genet.* **41,** 205–210 (2009).
144 Nair, R. P., Duffin, K. C., Helms, C., Ding, J., Stuart, P. E., Goldgar, D. *et al.* Genome-wide scan reveals association of psoriasis with IL-23 and NF-kappaB pathways. *Nat. Genet.* **41,** 199–204 (2009).
145 Levy, S., Sutton, G., Ng, P. C., Feuk, L., Halpern, A. L., Walenz, B. P. *et al.* The diploid genome sequence of an individual human. *PLoS Biol.* **5,** e254 (2007).
146 Wheeler, D. A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A. *et al.* The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452,** 872–876 (2008).
147 Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456,** 53–59 (2008).
148 Wang, J., Wang, W., Li, R., Li, Y., Tian, G., Goodman, L. *et al.* The diploid genome sequence of an Asian individual. *Nature* **456,** 60–65 (2008).
149 Kim, J. I., Ju, Y. S., Park, H., Kim, S., Lee, S., Yi, J. H. *et al.* A highly annotated whole-genome sequence of a Korean individual. *Nature* **460,** 1011–1015 (2009).
150 Ahn, S. M., Kim, T. H., Lee, S., Kim, D., Ghang, H., Kim, D. S. *et al.* The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. *Genome Res.* **19,** 1622–1629 (2009).
151 Pushkarev, D., Neff, N. F. & Quake, S. R. Single-molecule sequencing of an individual human genome. *Nat. Biotechnol.* **27,** 847–852 (2009).
152 Kuehn, B. M. 1000 Genomes Project promises closer look at variation in human genome. *JAMA* **300,** 2715 (2008).
153 Frazer, K. A., Murray, S. S., Schork, N. J. & Topol, E. J. Human genetic variation and its contribution to complex traits. *Nat. Rev. Genet.* **10,** 241–251 (2009).
154 Stranger, B. E., Forrest, M. S., Dunning, M., Ingle, C. E., Beazley, C., Thorne, N. *et al.* Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* **315,** 848–853 (2007).
155 Henrichsen, C. N., Vinckenbosch, N., Zöllner, S., Chaignat, E., Pradervand, S., Schütz, F. *et al.* Segmental copy number variation shapes tissue transcriptomes. *Nat. Genet.* **41,** 424–429 (2009).
156 Cahan, P., Li, Y., Izumi, M. & Graubert, T. A. The impact of copy number variation on local gene expression in mouse hematopoietic stem and progenitor cells. *Nat. Genet.* **41,** 430–437 (2009).
157 Hinds, D. A., Kloek, A. P., Jen, M., Chen, X. & Frazer, K. A. Common deletions and SNPs are in linkage disequilibrium in the human genome. *Nat. Genet.* **38,** 82–85 (2006).
158 McCarroll, S. A., Hadnott, T. N., Perry, G. H., Sabeti, P. C., Zody, M. C., Barrett, J. C. *et al.* Common deletion polymorphisms in the human genome. *Nat. Genet.* **38,** 86–92 (2006).
159 McCarroll, S. A., Kuruvilla, F. G., Korn, J. M., Cawley, S., Nemesh, J., Wysoker, A. *et al.* Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat. Genet.* **40,** 1166–1174 (2008).
160 McCarroll, S. A., Huett, A., Kuballa, P., Chilewski, S. D., Landry, A., Goyette, P. *et al.* Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease. *Nat. Genet.* **40,** 1107–1112 (2008).
161 Conrad, D. F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y. *et al.* Origins and functional impact of copy number variation in the human genome. *Nature* (2009) (e-pub ahead of print).
162 Hollox, E. J., Huffmeier, U., Zeeuwen, P. L., Palla, R., Lascorz, J., Rodijk-Olthuis, D. *et al.* Psoriasis is associated with increased beta-defensin genomic copy number. *Nat. Genet.* **40,** 23–25 (2008).
163 Gonzalez, E., Kulkarni, H., Bolivar, H., Mangano, A., Sanchez, R., Catano, G. *et al.* The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. *Science* **307,** 1434–1440 (2005).
164 Walsh, T., McClellan, J. M., McCarthy, S. E., Addington, A. M., Pierce, S. B., Cooper, G. M. *et al.* Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science* **320,** 539–543 (2008).
165 Sebat, J., Lakshmi, B., Malhotra, D., Troge, J., Lese-Martin, C., Walsh, T. *et al.* Strong association of *de novo* copy number mutations with autism. *Science* **316,** 445–449 (2007).
166 Wain, L. V., Armour, J. A. & Tobin, M. D. Genomic copy number variation, human health, and disease. *Lancet* **374,** 340–350 (2009).
167 Hannan, A. J. Tandem repeat polymorphisms: modulators of disease susceptibility and candidates for 'missing heritability'. *Trends Genet.* **26,** 59–65 (2009).
168 International Schizophrenia Consortium. Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature* **455,** 237–241 (2008).
169 Cohen, J. C., Kiss, R. S., Pertsemlidis, A., Marcel, Y. L., McPherson, R. & Hobbs, H. H. Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science* **305,** 869–872 (2004).
170 Romeo, S., Pennacchio, L. A., Fu, Y., Boerwinkle, E., Tybjaerg-Hansen, A., Hobbs, H. H. *et al.* Population-based resequencing of ANGPTL4 uncovers variations that reduce triglycerides and increase HDL. *Nat. Genet.* **39,** 513–516 (2007).
171 Nejentsev, S., Walker, N., Riches, D., Egholm, M. & Todd, J. A. Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. *Science* **324,** 387–389 (2009).
172 Ng, S. B., Turner, E. H., Robertson, P. D., Flygare, S. D., Bigham, A. W., Lee, C. *et al.* Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461,** 272–276 (2009).