European Journal of Human Genetics (2008) 16, 554–564

www.nature.com/ejhg

REVIEW

The success of the genome-wide association approach: a brief story of a long struggle

Ku Chee Seng*,[1] and Chia Kee Seng[1]

[1]*Center for Molecular Epidemiology, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore*

The genome-wide association approach has been the most powerful and efficient study design thus far in identifying genetic variants that are associated with complex human diseases. This approach became feasible as the result of several key advancements in genetic knowledge, genotyping technologies, statistical analysis algorithms and the availability of large collections of cases and controls. With all these necessary tools in hand, many genome-wide association studies were recently completed, and many more studies which will explore the genetic basis of various complex diseases and quantitative traits are soon to come. This approach has started to reap the fruits of its labor over the past several months. Publications of genome-wide association studies in several complex diseases such as inflammatory bowel disease, type-2 diabetes, breast cancer and prostate cancer have been abundant in the first half of this year. The aims of this review are firstly, to provide a timely summary for most of the genome-wide association studies that have been published until June/July 2007 and secondly, to evaluate to what extent these results have been validated in subsequent replication studies.
European Journal of Human Genetics (2008) **16,** 554–564; doi:10.1038/ejhg.2008.12; published online 20 February 2008

Keywords: genome-wide association; complex diseases; linkage disequilibrium; haplotype-tagging SNPs; copy number variations; International HapMap Project

Introduction

The genome-wide association (GWA) approach for genetic studies of complex human diseases was first proposed by Risch and Merikangas in 1996,[1] but only became feasible 10 years later. The Human Genome Project (HGP) was launched in 1990 and it took 13 years to finish the sequencing of the human genome. In 2001, both the International Human Genome Sequencing Consortium and Celera Genomics reported draft sequences of the human genome.[2,3] The HGP was deemed complete in April 2003, exactly 50 years after the description of the DNA double helix structure by James Watson and Francis Crick.[4] The HGP revealed that the human genome is composed of ~3 billion base pairs and an estimated 20 000–25 000 protein-coding genes. The completion of HGP, which represents an important milestone in human genomics,[5] was followed by identification and deposition of millions of single nucleotide polymorphisms (SNPs) into public databases by The SNP Consortium (TSC) and International Human Genome Sequencing Consortium.[6] This provided the foundation upon which GWA studies would subsequently build.

*Correspondence: CS Ku or Professor KS Chia, Center for Molecular Epidemiology, Yong Loo Lin School of Medicine, National University of Singapore, 16 Medical Drive, Singapore 117597, Singapore.
Tel: +65 8138 8095, Fax: 6478 9913;
E-mail: cmekcs@nus.edu.sg, cofcks@nus.edu.sg
Received 20 July 2007; revised 19 December 2007; accepted 11 January 2008; published online 20 February 2008

SNPs and copy number variations

SNPs are the most common genetic variations in the human genome; currently >10 million SNPs have been deposited into public databases, most of which are anticipated to have no functional effect. These genetic polymorphisms have proven to be very useful as genetic markers, and can be used to detect the disease variants via

linkage disequilibrium (LD). Owing to the large number of SNPs in the human genome, they provide the highest resolution (in comparison to other genetic markers such as micro-satellites and mini-satellites) and enable researchers to comprehensively interrogate the entire human genome. Nonetheless, SNPs alone can neither explain the total human genetic diversity nor explain the genetic susceptibility to complex diseases and adverse drug reactions. Recently, the discovery of thousands of copy number variations (CNVs), which are ubiquitous in the human genome, has provided another new insight into the complexity of the genetic variations in the human genome. CNVs are expected to play an important role in the genetic basis of complex diseases, and are therefore expected to share the limelight with SNPs in the future GWA studies. CNVs are structural variations or genomic alterations that change the number of copies of DNA involving segments that are larger than 1 kb (including deletions, insertions and duplications). CNVs were first reported ubiquitous in the human genome in 2004.[7,8] The global map of CNVs was finished in 2006 and lead to identification of about 1500 CNV regions.[9]

CNVs and other structural variations, such as inversions, that have thus far been identified have been deposited in the Database of Genomic Variants (http://projects.tcag.ca/variation/). The main objective of this international database is to provide a comprehensive catalog of structural variations in the human genome. The detailed description of CNVs is beyond the scope of this paper; however, a detailed review paper about structural variations is available.[10] CNVs have been reported to affect the gene expression[11] and the importance of CNVs has become recently apparent in susceptibility to complex diseases like autoimmune diseases, autism and bipolar disorder.[12–15]

International HapMap Project and the concept of LD

With the identification of millions of SNPs in the human genome, it remains a daunting task to genotype every single SNP, even with the latest genotyping technologies. To overcome this obstacle, the International HapMap Project was initiated in 2003[16] with the aim of characterizing LD patterns, and identifying haplotype-tagging SNPs in a total of 270 DNA samples that was collected from four major populations of European, African and Asian ancestry. The Phase I and Phase II of the International HapMap Project were completed in 2005 and 2007 respectively.[17,18] The application of the International HapMap Project is evident once we consider tagging SNPs that were identified in this global project were found to be 'transferable' in many populations around the world[19,20] and in isolated populations.[21,22]

At the same time, Perlegen® Sciences genotyped ~1.58 million SNPs on 71 individuals of European, African and Asian ancestry, and reported that these SNPs were able to capture most of the common genetic variations based on LD.[23] The major lesson that geneticists learnt from these two studies is that it is not necessary to genotype every single SNP in the human genome because this would be redundant. SNPs that are close to each other within a genomic region tend to be inherited together more frequent than expected by chance in a block pattern (known as haplotype) due to the presence of LD. Because of this unique relationship among SNPs, genotyping merely a set of informative SNPs to serve as proxy markers (usually called tagging SNPs, with $r^2 > 0.8$) is sufficient to capture most of the genetic information of SNPs, which are not genotyped with only slight loss of statistical power. r^2 is a measurement of 'correlation' or LD between two SNPs whose value ranges from 0 to 1 (r^2 of one indicates complete LD). r^2 depends on both allele frequencies and recombination between the two SNPs. The sample size that is required in a genetic association study is inversely proportional to the r^2 value.[24,25]

Key advancements in genotyping technology and genetic information

With the rapid development of genotyping technologies and decreasing of genotyping costs, currently, genotyping half a million SNPs on thousands of DNA samples is within the capacity of many research institutes. In addition to the fixed content genome-wide genotyping arrays, several custom made genotyping products were also introduced by Illumina® and Affymetrix® to accelerate the fine mapping of the genomic regions identified by GWA studies and linkage analysis.[26–28] The genome-wide genotyping products supplied by Illumina and Affymetrix such as Illumina HumanHap550 and Affymetrix GeneChip 500K offer good coverage of the International HapMap Phase I and Phase II data in both Caucasians and Asians. However, the genomic coverage in Africans was lower due to greater genetic diversity and weaker LD.[29,30]

With the wealth of genetic information gathered from the HGP and International HapMap Project, the collection of large number of cases and controls, the rapid advancement in genotyping technologies and the advent of powerful analysis algorithms such as PLINK,[31] the GWA approach is rapidly becoming feasible. GWA approach represents the most powerful and efficient study design in genetic dissection of complex diseases in comparison to traditional linkage studies[24,25] and will remain so until we reach the $1000 whole-genome sequencing era.[32]

GWA studies of complex human diseases

The strengths of GWA approach are that it is hypothesis free and that it is able to comprehensively interrogate the

entire human genome. This approach enables investigators to identify novel loci or genes for various diseases and quantitative traits. The achievements of GWA studies have been witnessed in genetic dissection of several complex diseases, namely, age-related macular degeneration (AMD), obesity, inflammatory bowel diseases (IBD), type-2 diabetes (T2D), breast cancer and prostate cancer. Owing to article length constraints, only the GWA studies that were published on these six complex diseases have been selected to be reviewed in this paper. Table 1 summarizes the genes or loci that were identified by GWA studies or consistently replicated for these diseases. In Figure 1, which illustrates the number of GWA publications from 2005 to June/July 2007, we see a sharp increase in the number of publications in 2007 in comparison to the previous 2 years. The aims of this review paper are to provide a timely summary of the GWA studies that were published until June/July 2007 and to evaluate to what extent the results have been replicated and validated. Replication of GWA results is essential to distinguish between 'statistical' artifacts and true associations.[57] This review paper was organized into several sections according to the disease phenotypes and chronologically by the year of publication. The disease sections are followed by a discussion on the determinant factors of a successful GWA study, the future challenges and limitations of GWA approach.

Table 1 The genes or loci that were identified by GWA studies or consistently replicated for complex diseases is reviewed in this paper

Disease	*Gene/locus*	*Reference*
AMD	CFH	33
	HTRA1	34
BMI and obesity	INSIG2	35
	FTO	36
IBD	CARD15/NOD2[a]	37,38
	5q31[a]	37,39
	IL23R[a]	40,37
	ATG16L1[a]	41,37
	10q21[a]	41,37
	5p13.1[a]	37,42
	5q33 (IRGM)	37,43
	3p21 (BSN)	37,43
	10q24.2 (NKX2-3)	37,43
	18p11 (PTPN22)	37,43
T2D	PPARG[b]	44
	KCNJ11[b]	45
	TCF7L2[b]	46
	SLC30A8[b]	47
	LD block contains IDE-KIF11-HHEX[b]	47
	LD block contains EXT2-ALX4	47
	CDKN2A/CDKN2B	48–50
	CDKAL1	48–51
	IGF2BP2	48–50
Breast cancer	FGFR2	52
	TNRC9	52
	MAP3K1	52
	LSP1	52
	FGFR2	53
	2q35	54
	16q12	54
Prostate cancer	8q24	55,56

[a]These genes and loci were successfully replicated by WTCCC GWA study (Reference [37]).
[b]These genes were successfully replicated by GWA studies (Reference: [48–51]).

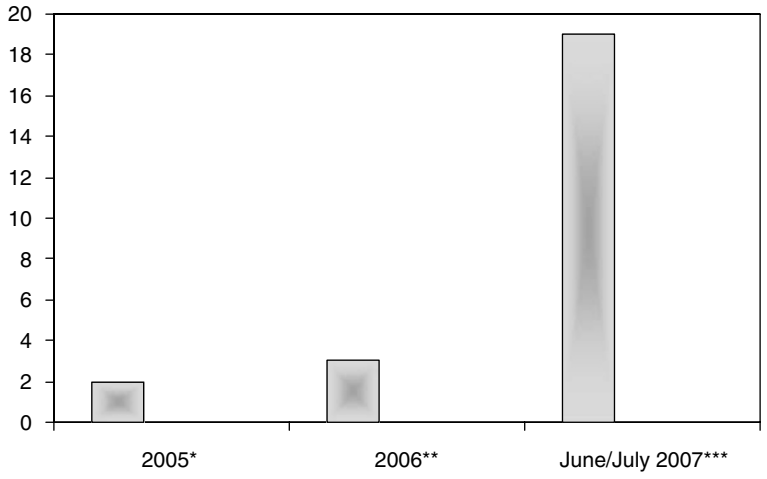
AMD

The success of GWA approach in genetic dissection of complex diseases was apparent in April 2005.[33] The first GWA study which used a commercial genotyping platform to examine the genetic basis of AMD had been published. In the genome-wide scan, Klein *et al*[33] identified a common intronic variant in the complement factor H (CFH) gene that strongly associated with AMD in 96 cases and 50 controls. The *P*-value of this association surpassed the genome-wide significance by Bonferroni correction even with a relatively small sample size. Perhaps both the commonness of the allele and the large genetic effect was reported to have – odds ratio (OR) for homozygous risk allele was 7.4 – contributed to the highly significant finding. In addition, accurate phenotyping may have played a key role in the study, since only AMD patients with the presence of large drusen were recruited, which reduced phenotypic heterogeneity. The genomic region that initially identified was followed by re-sequencing and fine mapping, and finally a nonsynonymous SNP (Y402H) that was strongly associated with AMD was reported. Concurrently, two independent groups[58,59] also reported similar results via fine mapping of the genomic region in 1q31-32 that was identified in previous studies. That three separate studies firmly pinned down the same functional variant speaks of the robustness of the association, which subsequent studies[60] have replicated. This is by far the most robust association that has been derived from a GWA study.

AMD can be classified into two clinical subtypes, dry (non-neovascular) or wet (neovascular). The former subtype, which accounts for ~90% of AMD cases, was associated with the functional variant identified in CFH. However, a novel genetic variant, an SNP (rs11200638) located upstream of the HtrA serine peptidase 1 (HTRA1) putative transcription start site, was also identified for wet subtype of AMD in the study by DeWan *et al*.[34] This association was subsequently replicated in a Caucasian population[61] and in a Japanese population.[62]

Body mass index and obesity

Several GWA studies were published in 2006 after the genetic community witnessed the successful results of



* Ref: 33, 83

** Ref: 34, 35, 40

*** Ref: 36, 37, 41, 42, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 81, 82, 88, 89, 90

This is not a complete file.

Figure 1 The number of GWA publications from 2005 to June/July 2007.

AMD. The first was published in April 2006 to study the genetic basis of obesity. Using the Affymetrix GeneChip 100K to genotype individuals from the Framingham Heart Study cohort, Herbert *et al*[35] identified a novel common genetic variant (rs7566605) near the insulin signaling protein type 2 (INSIG2) gene that was associated with an increased body mass index (BMI) or obesity. They subsequently tested the SNP in five additional cohorts and the association was replicated in all except one. These results provided strong statistical evidence to support the association between INSIG2 gene and obesity; nevertheless, the genetic association failed to be replicated by three independent studies.[63–66] To date, the results from subsequent genetic association studies have been conflicting; the association seems to be reproduced in several but not all the studies.[67] Since then, there has been little success in the identification of genetic determinants of obesity, except for one novel gene that will be discussed in the T2D section of this review.

IBD

The first GWA study on IBD was conducted by the North American IBD Genetics Consortium.[40] IBD is a chronic inflammatory disease of gastro-intestinal tract and it can be divided into two clinical subtypes, namely, Crohn's disease (CD) and ulcerative colitis. The investigators recruited ileal CD cases to minimize the phenotypic heterogeneity. This careful ascertainment of cases is recommended because it will increase the statistical power to detect the disease variants especially if the diverse clinical manifestations are due to genetic heterogeneity. Duerr *et al*[40] identified a novel gene for IBD – a nonsynonymous SNP (Arg381Gln) in interleukin-23 receptor (IL23R) gene, during an interim analysis of their data. This gene encodes a subunit of the receptor for IL23 (pro-inflammatory cytokine) and thus is an interesting and biologically plausible gene for inflammatory diseases. The genetic association was subsequently replicated in a Jewish case–control and in a family-based association study.

The association that they found in the interim analysis was then unequivocally replicated in four independent groups[68–71] and provided compelling evidence to support IL23R as a genuine susceptibility gene for IBD. The results of their complete genome scan were published recently[41] and the investigators were able to uncover several novel loci for IBD, the most notable novel gene was autophagy-related 16-like 1 (ATG16L1). The nonsynonymous SNP (rs2241880) located in the exon 8 of this autophagy gene was previously reported to be associated with CD in a gene-centric GWA study.[72] Likewise, this association was unambiguously replicated by two independent studies from UK.[73,74]

T2D

Although the role of genetic susceptibility in T2D is well established, the results from genetic association studies

have been quite disappointing. Before 2006, there was limited success in genetic studies of T2D. The genes identified, with the exception of PPARG and KCNJ11 gene,[44,45] have been conflicting and inconsistent. In 2006, TCF7L2 gene was first reported to be associated with T2D by deCODE Genetics, who fine mapped a suggestive linkage region identified previously in an Icelandic population.[46] The association has since been consistently replicated in more than 20 studies across different populations with diverse ancestral backgrounds, thereby providing convincing evidence that TCF7L2 was associated with T2D.[75]

The first GWA study on T2D revealed four novel loci, the most notable being a nonsynonymous SNP in solute carrier family 30 member 8 (SLC30A8) gene.[47] It encodes a zinc transporter protein expressed only in β cells, which is also implicated in the final stages of insulin biosynthesis, making this gene a strong biological candidate for T2D. In addition, the investigator also confirmed the previously identified diabetes gene – TCF7L2. However, only two from these four novel loci, were successfully replicated by three diabetes GWA studies that published concurrently in Science (discussed below).

The aim of the GWA study by Frayling *et al*[36] was to identify susceptibility genes for T2D; however, they uncovered a new gene for BMI – fat mass and obesity-associated (FTO) gene. Initially, several SNPs in FTO gene were found strongly associated with T2D. However, the association was abolished after adjustment for BMI in cases and controls. It was therefore concluded that the FTO gene was more likely to be associated with BMI or obesity. To test this hypothesis the investigators attempted to replicate the association in nine independent studies and the results showed that the common variants in FTO gene were reproducibly associated with BMI and the risk of being overweight or obese from childhood to adult. This enormous replication effort has provided strong evidence beyond statistical doubt for the genetic association.

Huge success in genetic dissection of T2D was achieved this year by three GWA studies conducted by the Wellcome Trust Case Control Consortium (WTCCC), Diabetes Genetic Initiative and Finland-United States Investigation of Non-Insulin Dependent Diabetes Mellitus Genetics.[48–50] These diabetes research groups were able to discover three novel loci for T2D. Their success highlights the importance of scientific collaboration and sharing of genome-wide genotyping data among different research groups. For these three GWA studies, the combined sample size exceeded 32 000 samples. This large sample size allowed the investigators to detect variants with modest genetic effects (OR of 1.1–1.2). All three putative candidate genes that were identified are biologically plausible genes for T2D, that is, cyclin-dependent kinase inhibitor 2A/2B (CDKN2A/CDKN2B), CDK5 regulatory subunit associated protein 1-like 1 (CDKAL1) and insulin-like growth factor 2 binding protein 2 (IGF2BP2). In addition to these discoveries, they also successfully replicated the genetic association of several genes known to be associated with diabetes, namely, PPARG, KCNJ11, TCF7L2, SLC30A8 and HHEX. In all, eight loci/genes have been detected and consistently replicated for T2D in Caucasians. Interestingly, all the genetic variants identified have been located in noncoding regions, particularly an SNP (rs10811661) that is located 125 kb away from the nearest annotated genes that is CDKN2A/CDKN2B. In addition to these three studies, an independent GWA study was conducted concurrently in an Icelandic population.[51] The investigators also managed to identify the CDKAL1 gene for T2D. All newly discovered loci by these GWA studies were replicated in a series of studies with large sample size, and are therefore likely to be bona fide loci or genes for T2D. The identification of these novel loci is important to further enhance our understanding on the genetic basis and pathogenesis of T2D.

Breast cancer

The highly penetrant genes – BRCA1 and BRCA2 – only account for $\sim$20% of the total genetic risk of breast cancer. Thus far, the results from the genetic association studies for this cancer have been dissatisfying. However, with the efforts of the Breast Cancer Association Consortium in conducting candidate gene case–control association studies with enhanced statistical power by combining several breast cancer cohorts, two novel genes were identified, namely, CASP8 and TGFB1.[76] Recently, the success of genetic studies of breast cancer has also been seen in three studies[52–54] and their findings are starting to shed some new light on the genetic basis of this cancer. With their three-stage study design and a sample size of more than 50 000 cases and controls, Easton *et al*[52] identified six highly significant SNPs. The most notable genes identified were fibroblast growth factor receptor 2 (FGFR2) and the LD block, which contain TNRC9 gene. FGFR2 encodes a tyrosine kinase receptor, which was overexpressed in breast cancer. Therefore, it is a strong biological candidate gene. This novel gene was simultaneously uncovered by Hunter *et al*[53] who identified four SNPs within intron 2 of FGFR2 that were highly associated with breast cancer. Collectively, these findings indicate a novel susceptibility gene for breast cancer, but further studies are required to fine map and to identify the disease variants.

The third GWA study was done by Stacey *et al*[54] in Icelandic breast cancer cases and controls. Ten SNPs with the most significant *P*-values were tested in additional five cohorts. However, only two SNPs were consistently associated with breast cancer that is one SNP on chromosome 2q35 (rs13387042) and the other one on 16q12 (rs3803662). Interestingly, the 2q35 region contains no

known genes, but the SNP that falls on 16q12 is located near the 5′ region of TNRC9 gene that was simultaneously identified by Easton *et al*[52].

Prostate cancer

Just as with other cancers; identifying the genetic variants with modest effect for prostate cancer has been proven difficult. However, some success was achieved recently in genetic studies of prostate cancer. Chromosome 8q24 region was first identified via the genome-wide linkage analysis in Icelandic families with prostate cancer. Allele-8 of the microsatellite DG8S737 was consistently associated with the disease in a series of studies.[77] One year later, the investigators reported the second genetic variant in the 8q24 region for prostate cancer through GWA study.[55] Yeager *et al*[56] also applied the same approach in genetic exploration of prostate cancer. Haiman *et al*[78] followed up on their initial admixture results by extensively fine mapping the region. The results from these three independent studies suggest that 8q24 is implicated in prostate cancer and that this genomic region, may be harboring susceptibility genes for prostate cancer. Further studies are needed to discern the disease variants and genes within this region.

WTCCC

WTCCC is the largest ever GWA study to explore the genetics of seven common diseases.[37] In total, 17 000 individuals that is 2000 cases for each of the seven common diseases and 3000 shared controls were genotyped by Affymetrix GeneChip 500K. This is a huge success because many novel loci were uncovered for these common diseases and many of them have been successfully replicated in other independent sample sets, namely, T2D (as discussed above), CD,[43] and type-1 diabetes.[79] As for CD, the second autophagy gene – immunity-related GTPase family, M – was revealed in the study by Parkes *et al*.[43] In addition to the discovery of many novel loci, almost all of the genes that identified by previous studies for these common diseases were successfully replicated by WTCCC, for example, HLA-DRB1, INS, CTLA4, PTPN22, IFIH1 and IL2RA[80] genes were replicated for type-1 diabetes. As a whole, these results suggest that there may be two important pathways for the pathogenesis of CD that is IL23 and autophagy pathways.

From the list of novel loci or genes revealed by WTCCC, one of the most intriguing results, perhaps, is the association of coronary artery disease with the region on chromosome 9p21 as this region contains the coding sequences for CDKN2A/CDKN2B, which are genes associated with T2D as well. At the same time, the same region was found to be associated with coronary heart disease and myocardial infarction by other two independent studies respectively.[81,82] These results might help to explain why some individuals are more susceptible to these two closely related common diseases.

Parkinson disease and other neurological diseases

The papers that were discussed above are examples of successful GWA studies, which yielded spectacular results (except the association of INSIG2 gene with obesity). On the other hand, there have been examples of GWA studies of complex diseases where validation of initial results has not been achieved yet. For instance, a two-stage GWA study by Maraganore *et al*[83] identified 13 SNPs strongly associated with Parkinson disease. Nevertheless, lack of replication was observed subsequently[84–86] despite a large sample size of $\sim$10 000 from a large scale international study.[87]

Hitherto three genome-wide scans using either Illumina HumanHap300 or HumanHap550 genotyping platform were completed for these complex diseases, namely, Parkinson disease, ischemic stroke and amyotrophic lateral sclerosis[88–90] and the data was released into public domain. None of the SNPs achieved or surpassed genome-wide significance in these three studies. There are several possible explanations: (1) the Bonferroni correction is too stringent, which may have overcorrected the significance threshold, (2) there is a lack of statistical power to detect common variants with modest effect in these genome-wide scans because of relatively small sample sizes and finally, (3) perhaps there is no disease variant with large genetic effect (like the case of CFH gene for AMD) for these neurological diseases. Nevertheless, the availability of these resources will allow researchers to access the genome-wide scan data and will thus accelerate the pace of discovery to identify novel genetic variants for these neurological diseases.

Determinant factors for a successful GWA study

From the experience of recent GWA studies, we learn that one of the major determinants of the success seems to be the requirement of a large sample size to provide adequate statistical power to detect genetic variants with modest effect that is OR$<$1.5. The statistical power of genetic association study is basically a function of sample size, magnitude of genetic effect, and allele frequency. As the latter two factors are unknown until the genetic variants are uncovered, sample size is the major controllable factor in the determination of statistical power. In addition, power also depends on the tag SNPs selected that is the genome coverage.[91] Both these factors are modifiable, thus, increasing both the sample size and the genome coverage will increase the statistical power of the study.

The large sample size can be more easily achieved through consortia or collaboration as demonstrated in the breast cancer (BCAC) and T2D studies respectively, and the Genetic Association Information Network is a good example of these efforts. Genetic Association Information Network is a public–private partnership that was established to interrogate the genetic basis of common diseases through a series of collaborative GWA studies.[92]

Almost all the SNPs that appeared highly significant in WTCCC GWA study either had strong prior evidence of association with the diseases studied or were successfully confirmed in the subsequent replication studies. A large sample size with an enriched statistical power is both crucial and essential in GWA study to ensure that genuine associations rank at the top of the SNPs according to the *P*-values.[93] Researchers should be cautious when applying stringent significance thresholds (like Bonferroni corrections) or when weighing the SNPs for replication in a two-stage study design to control the false-positive associations, as they run the risk of overcorrecting and subsequently having type II error (because many true positive associations with modest *P*-values would have been excluded in the genome-wide scan). Stringent significance threshold is only reasonable if the sample size in the genome-wide scan is large and the statistical power is adequate enough to allow most of the true signals to rank at the top of the list.

All reviewed studies used commercially available genotyping platforms from Illumina, Affymetrix or Perlegen. Good genome coverage is of utmost importance in GWA studies because the underlying principle of this approach is based on LD to detect the disease variants. In those regions with a scarce number of SNPs or which are poorly covered by markers, genuine disease variants might be missed since these disease variants were not in strong LD ($r^2 > 0.8$) with any of the SNPs that were genotyped on the array. Furthermore, extensive replication is essential to declare a bona fide association. Reproducing strong associations that were found in the GWA does not seem to be a major issue in the recent GWA studies discussed above (with few exceptions), as all were successfully validated in a series of replication studies and provided strong evidence to support the genetic association.

Proper study design also plays a key role in determining the success of GWA studies. Researchers need to pay more attention on the methodology and analysis issues such as applying stringent control of the quality of genotyping data to minimize the genotyping errors which could both produce spurious associations and mask the true associations. It is also recommended to use other genotyping platforms such as Sequenom iPlex to validate the SNPs from genome-wide scan since this genotyping technology uses a totally different principle in allelic discrimination or genotyping, that is, MALDI-TOF MS (matrix assisted laser desorption/ionization, time-of-flight mass spectrometry)[94] *versus* hybridization and fluorescent intensity measurement methods employed in both Illumina and Affymetrix platforms. The accuracy of classification of disease phenotype is equally important for a successful GWA study; the importance of this criterion was demonstrated in the AMD and Crohn's disease studies by Klein *et al* and Duerr *et al*[33,40] respectively.

It is important to address the issue of population stratification, even when the study was conducted in a relatively homogenous population, or if the cases and controls were well matched and recruited from the same geographical location, because these techniques cannot totally eliminate the effect of population stratification. The effect of this confounding factor will be intensified in the GWA studies where tens of thousands of samples are needed.[95] Freeware such as EIGENSTRAT[96] (available online) should be applied in GWA studies to identify outliers with different ancestry backgrounds and to exclude them from further analysis. This issue has been receiving attention from researchers in their GWA studies as discussed above to exclude entirely the possibility that the positive associations observed are attributed to population stratification.

Challenges waiting ahead

Once the researchers establish the association beyond statistical doubt, three additional challenges are still waiting ahead. First, although many novel loci were identified for complex diseases, the task of identifying the actual functional disease variants remains ahead. It is often difficult to discern the disease variants, especially when the surrounding markers are in perfect or nearly perfect LD ($r^2 > 0.9$), because they will give almost similar strength of association. Therefore, re-sequencing is usually needed after identifying the genomic region that potentially harbors the disease variants. Re-sequencing strategies will enable investigators to uncover novel and uncommon variants. Often, functional studies are also required, but these studies are only feasible for those genes or regions, which are well-characterized. In most cases, GWA approach is unlikely to directly reveal the functional variants for the disease. This is demonstrated by the AMD study where DeWan and coworkers first identified an intragenic SNP in their genome-wide scan before they unraveled the functional variant that affects the transcription of HTRA1 gene.[34]

Second, it remains difficult to establish the functional role of the disease variants, for example, how the disease variants affect the structure and function of the genes (and its end product – proteins), and also transcription regulation. This is especially challenging for SNPs located in genes, which are not well-characterized or unknown functions, noncoding regions and gene deserts since our knowledge about the functional elements in the human

genome is still very limited. For instance, a strong association for a cluster of SNPs on chromosome 5p13.1 was consistently found for CD.[37,42] Interestingly this region was located within a 1.2 Mb gene desert and the nearest annotated gene is prostaglandin E receptor 4 (PTGER4). So, one wonders how these 'long-distance' variants affect the function of the disease genes? Perhaps we can get some answers from the pilot phase of the ENCODE Project.[97] The ENCODE Projects found that regulatory regions or elements of a gene can be located far from it and yet still be able to affect expression and function the gene. This project was initiated after the completion of HGP with the aim to identify and characterize all the functional elements within the entire human genome.[98] Although the pilot phase of the ENCODE Project was finished; there is still 99% of human genome that needs to be explored. Investigating the functional roles of those SNPs located within noncoding regions and gene deserts pose tremendous challenges while promising the possibility of great rewards, the identification of novel functional elements previously uncharacterized. Lastly, it is not a trivial task to elucidate the molecular pathway based on the results derived from GWA studies of the disease, especially for genes or proteins with unknown function.

Limitations of GWA studies

Nowadays, GWA approach is the 'best' medicine in dissecting the genetic basis of complex diseases, but it is not the 'panacea'. There are several limitations and problems with this study design. In GWA approach, several hundred thousands of SNP markers throughout the entire genome are analyzed at once, creating a multiple-hypothesis problem which can lead to substantial type I error. To minimize the false-positive results, statistical adjustment like Bonferroni correction is applied and a very stringent P-value is needed, usually at the significance level of $10-7$. This translates into the requirement of a large sample size of tens of thousands of samples for both genome-wide scan and replication studies. This requirement is hard and is not likely to be attained in a single study; hence, collaborations and the establishment of consortia are of utmost importance.

GWA approach is based on the principle of LD; therefore, the genetic markers that identified are unlikely to be the disease variants. Extensive re-sequencing and fine mapping are required to discern the disease variants; it is a great challenge in fine mapping when the SNPs within the genomic region are in strong LD. This is a double-edged sword, although strong LD helps to reduce the number of markers to genotype in the genome-wide scan, it also limits the ability to 'resolve' the association and creates difficulty to identify the 'culprits'. Biological studies are therefore required to determine the functional roles of the disease variants.

The GWA study design is hypothesis-generating rather than hypothesis-testing; therefore, replication is paramount in GWA studies to confirm the results, and replication has been widely accepted as the gold standard to discern genuine genetic associations. Lack of replication or conflicting result is still a problem in some GWA studies of diseases such as Parkinson's disease[83] and INSIG2 gene for obesity.[35] These and other problems in GWA studies have been well addressed by Shriner *et al*[99] and Williams *et al*[100] in their Letters to Science.

Conclusions

The genetic spectrum of complex human diseases has yet to be elucidated, but the recent achievements in genetic studies of various complex diseases have provided some new insights. Most of the genetic variants that have been consistently identified for the diseases studied are common (minor allele frequency $>5\%$) and confer only modest genetic effect (OR<1.5). Does it mean that the genetic spectrum of complex diseases comprises only of common variants with modest effect? The answer is probably no; rare variants, such as IL23R,[40] are likely to contribute to disease risk as well as influence the quantitative trait for example, high-density lipoprotein cholesterol level.[101,102] The relative proportion of common variants *versus* rare variants in the total genetic contribution to both complex diseases and quantitative traits is still largely unknown.

Less than expected success in identifying rare genetic variants might be due to the fact that current genotyping platforms have poor coverage for rare variants. The SNP markers included in both Illumina and Affymetrix genotyping arrays are biased toward common alleles. As a result, these markers are in weak LD (r^2) with rare variants, mainly because of the discrepancy between their frequencies. To get a high r^2 value, the frequencies of the two SNPs must be comparable in addition to no recombination that occurred between them, so the proxy marker could predict the nongenotyped SNP via LD. Since statistical power drops drastically for those rare SNPs, a larger sample size than the figure reported in the recent GWA studies might be needed to detect rare variants.

According to the common disease common variant hypothesis,[24,25] common disease such as T2D is likely due to common genetic variants with modest effect. Since the genetic variants are common, they are likely to be shared across different populations with diverse ancestry backgrounds. Most of the GWA and replication studies were conducted in Caucasian populations; less replication effort has been devoted in other populations like Asians and Africans. So it would be interesting to determine and investigate how many loci or genes identified by these

GWA studies are also associated with the disease phenotypes in other populations. Well-designed replication studies are crucial to either validate or refute the initial positive association. The guidelines to conduct replication studies were suggested by NCI-NHGRI Working Group on Replication in Association Studies.[57]

The most successful studied diseases thus far are T2D and CD; about 10 loci or genes have been consistently identified in each of the diseases from the findings of GWA studies. Does this signal the end of genetic association studies of these diseases? This is probably only the beginning; it has been predicted that there are still plenty of genetic variants or genes underlying the genome for the researchers to uncover.

Note

The pace of development in GWA studies is at an unprecedented speed, such that during the submission and revision of this review paper, a substantial number of GWA studies were also published. However, it is certainly beyond the scope and the length in this review paper. However, we think that it is important to briefly highlight the GWA studies published during July–December 2007 (although the list is incomplete); the complex diseases that were interrogated by these GWA studies include the restless legs syndrome (periodic limb movements),[103,104] coronary artery disease,[105] multiple sclerosis,[106] gallstone disorder,[107] exfoliation glaucoma,[108] colorectal cancer,[109,110] HIV,[111] type 1 diabetes,[112] childhood asthma,[113] atrial fibrillation,[114] sporadic amyotrophic lateral sclerosis[115,116] and rheumatoid arthritis.[117,118]

Acknowledgements

The authors declare no conflicts of interest. We are grateful to Dr Sonia Davila (Genome Institute of Singapore) for her valuable comments and kind assistance in revising this paper. We are thankful to Kaavya Narasimhalu (Center for Molecular Epidemiology, National University of Singapore) for critical proofreading of this review paper.

References

1 Risch N, Merikangas K: The future of genetic studies of complex human diseases. *Science* 1996; **273**: 1516–1517.
2 International Human Genome Sequencing Consortium: Initial sequencing and analysis of the human genome. *Nature* 2001; **409**: 860–921.
3 Venter JC, Adams MD, Myers EW *et al*: The sequence of the human genome. *Science* 2001; **291**: 1304–1351.
4 Watson JD, Crick FH: Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* 1953; **171**: 737–738.
5 International Human Genome Sequencing Consortium: Finishing the euchromatic sequence of the human genome. *Nature* 2004; **431**: 931–945.
6 International SNP Map Working Group: A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* 2001; **409**: 928–933.
7 Sebat J, Lakshmi B, Troge J *et al*: Large-scale copy number polymorphism in the human genome. *Science* 2004; **305**: 525–528.
8 Iafrate AJ, Feuk L, Rivera MN *et al*: Detection of large-scale variation in the human genome. *Nat Genet* 2004; **36**: 949–951.
9 Redon R, Ishikawa S, Fitch KR *et al*: Global variation in copy number in the human genome. *Nature* 2006; **444**: 444–454.
10 Feuk L, Carson AR, Scherer SW: Structural variation in the human genome. *Nat Rev Genet* 2006; **7**: 85–97.
11 Stranger BE, Forrest MS, Dunning M *et al*: Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* 2007; **315**: 848–853.
12 Fanciulli M, Norsworthy PJ, Petretto E *et al*: FCGR3B copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity. *Nat Genet* 2007; **39**: 721–723.
13 Yang Y, Chung EK, Wu YL *et al*: Gene copy-number variation and associated polymorphisms of complement component C4 in human systemic lupus erythematosus (SLE). *Am J Hum Genet* 2007; **80**: 1037–1054.
14 Sebat J, Lakshmi B, Malhotra D *et al*: Strong association of *de novo* copy number mutations with autism. *Science* 2007; **316**: 445–449.
15 Lachman HM, Pedrosa E, Petruolo OA *et al*: Increase in GSK3beta gene copy number variation in bipolar disorder. *Am J Med Genet B Neuropsychiatr Genet* 2007; **144**: 259–265.
16 The International HapMap Consortium: The international HapMap project. *Nature* 2003; **426**: 789–796.
17 The International HapMap Consortium: A haplotype map of the human genome. *Nature* 2005; **437**: 1299–1320.
18 The International HapMap Consortium: A second generation human haplotype map of over 3.1 million SNPs. *Nature* 2007; **449**: 851–861.
19 de Bakker PI, Burtt NP, Graham RR *et al*: Transferability of tag SNPs in genetic association studies in multiple populations. *Nat Genet* 2006; **38**: 1298–1303.
20 Conrad DF, Jakobsson M, Coop G *et al*: A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. *Nat Genet* 2006; **38**: 1251–1260.
21 Bonnen PE, Pe'er I, Plenge RM *et al*: Evaluating potential for whole genome studies in Kosrae, an isolated population in Micronesia. *Nat Genet* 2006; **38**: 214–217.
22 Service S, International Collaborative Group on Isolated Populations, Sabatti C, Freimer N: Tag SNPs chosen from HapMap perform well in several population isolates. *Genet Epidemiol* 2007; **31**: 189–194.
23 Hinds DA, Stuve LL, Nilsen GB *et al*: Whole-genome patterns of common DNA variation in three human populations. *Science* 2005; **307**: 1072–1079.
24 Wang WY, Barratt BJ, Clayton DG, Todd JA: Genome-wide association studies: theoretical and practical concerns. *Nat Rev Genet* 2005; **6**: 109–118.
25 Hirschhorn JN, Daly MJ: Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet* 2005; **6**: 95–108.
26 Gunderson KL, Steemers FJ, Lee G, Mendoza LG, Chee MS: A genome-wide scalable SNP genotyping assay using microarray technology. *Nat Genet* 2005; **37**: 549–554.
27 Steemers FJ, Gunderson KL: Whole genome genotyping technologies on the BeadArray platform. *Biotechnol J* 2007; **2**: 41–49.
28 Kennedy GC, Matsuzaki H, Dong S *et al*: Large-scale genotyping of complex DNA. *Nat Biotechnol* 2003; **21**: 1233–1237.
29 Barrett JC, Cardon LR: Evaluating coverage of genome-wide association studies. *Nat Genet* 2006; **38**: 659–662.
30 Eberle MA, Ng PC, Kuhn K *et al*: Power to detect risk alleles using genome-wide tag SNP panels. *PLoS Genet* 2007; **3**: 1827–1837.
31 Purcell S, Neale B, Todd-Brown K *et al*: PLINK: a toolset for whole genome association and population based linkage analyses. *Am J Hum Genet* 2007; **81**: 559–575.
32 Bentley DR: Whole-genome re-sequencing. *Curr Opin Genet Dev* 2006; **16**: 545–552.

33 Klein RJ, Zeiss C, Chew EY *et al*: Complement factor H polymorphism in age-related macular degeneration. *Science* 2005; **308**: 385–389.
34 DeWan A, Liu M, Hartman S *et al*: HTRA1 promoter polymorphism in wet age-related macular degeneration. *Science* 2006; **314**: 989–992.
35 Herbert A, Gerry NP, McQueen MB *et al*: A common genetic variant is associated with adult and childhood obesity. *Science* 2006; **312**: 279–283.
36 Frayling TM, Timpson NJ, Weedon MN *et al*: A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science* 2007; **316**: 889–894.
37 The Wellcome Trust Case Control Consortium: Genome-wide association study of 14 000 cases of seven common diseases and 3000 shared controls. *Nature* 2007; **447**: 661–678.
38 Hugot JP, Chamaillard M, Zouali H *et al*: Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature* 2001; **411**: 599–603.
39 Rioux JD, Daly MJ, Silverberg MS *et al*: Genetic variation in the 5q31 cytokine gene cluster confers susceptibility to Crohn disease. *Nat Genet* 2001; **29**: 223–228.
40 Duerr RH, Taylor KD, Brant SR *et al*: A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. *Science* 2006; **314**: 1461–1463.
41 Rioux JD, Xavier RJ, Taylor KD *et al*: Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis. *Nat Genet* 2007; **39**: 596–604.
42 Libioulle C, Louis E, Hansoul S *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* 2007; **3**: e58.
43 Parkes M, Barrett JC, Prescott NJ *et al*: Sequence variants in the autophagy gene IRGM and multiple other replicating loci contribute to Crohn's disease susceptibility. *Nat Genet* 2007; **39**: 830–832.
44 Altshuler D, Hirschhorn JN, Klannemark M *et al*: The common PPARγ Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. *Nat Genet* 2000; **26**: 76–80.
45 Gloyn AL, Weedon MN, Owen KR *et al*: Large-scale association studies of variants in genes encoding the pancreatic beta-cell KATP channel subunits Kir6.2 (KCNJ11) and ABCC8 confirm that the KCNJ11 E23K variant is associated with type-2 diabetes. *Diabetes* 2006; **52**: 568–572.
46 Grant SF, Thorleifsson G, Reynisdottir I *et al*: Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes. *Nat Genet* 2006; **38**: 320–323.
47 Sladek R, Rocheleau G, Rung J *et al*: A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* 2007; **445**: 881–885.
48 Zeggini E, Weedon MN, Lindgren CM *et al*: Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science* 2007; **316**: 1336–1341.
49 Diabetes Genetic Initiative (DGI) of Broad Institute of Harvard and MIT, Lund University and Novartis Institute for Biomedical Research: Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* 2007; **316**: 1331–1336.
50 Scott LJ, Mohlke KL, Bonnycastle LL *et al*: A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science* 2007; **316**: 1341–1345.
51 Steinthorsdottir V, Thorleifsson G, Reynisdottir I *et al*: A variant in CDKAL1 influences insulin response and risk of type 2 diabetes. *Nat Genet* 2007; **39**: 770–775.
52 Easton DF, Pooley KA, Dunning AM *et al*: Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature* 2007; **447**: 1087–1093.
53 Hunter DJ, Kraft P, Jacobs KB *et al*: A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. *Nat Genet* 2007; **39**: 870–874.
54 Stacey SN, Manolescu A, Sulem P *et al*: Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer. *Nat Genet* 2007; **39**: 865–869.
55 Gudmundsson J, Sulem P, Manolescu A *et al*: Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24. *Nat Genet* 2007; **39**: 631–637.
56 Yeager M, Orr N, Hayes RB *et al*: Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nat Genet* 2007; **39**: 645–649.
57 NCI-NHGRI Working Group on Replication in Association Studies: Replicating genotype-phenotype associations. *Nature* 2007; **447**: 655–660.
58 Haines JL, Hauser MA, Schmidt S *et al*: Complement factor H variant increases the risk of age-related macular degeneration. *Science* 2005; **308**: 419–421.
59 Edwards AO, Ritter R, Abel KJ *et al*: Complement factor H polymorphism and age-related macular degeneration. *Science* 2005; **308**: 421–424.
60 Thakkinstian A, Han P, McEvoy M *et al*: Systematic review and meta-analysis of the association between complement factor H Y402H polymorphisms and age-related macular degeneration. *Hum Mol Genet* 2006; **15**: 2784–2790.
61 Yang Z, Camp NJ, Sun H *et al*: A variant of the HTRA1 gene increases susceptibility to age-related macular degeneration. *Science* 2006; **314**: 992–993.
62 Yoshida T, Dewan A, Zhang H *et al*: HTRA1 promoter polymorphism predisposes Japanese to age-related macular degeneration. *Mol Vis* 2007; **4**: 545–548.
63 Dina C, Meyre D, Samson C *et al*: Comment on 'A common genetic variant is associated with adult and childhood obesity'. *Science* 2007; **315**: 187b.
64 Loos RJF, Barroso I, Rahilly SO, Wareham NJ: Comment on 'A common genetic variant is associated with adult and childhood obesity'. *Science* 2007; **315**: 187c.
65 Rosskopf D, Bornhorst A, Rimmbach C *et al*: Comment on 'A common genetic variant is associated with adult and childhood obesity'. *Science* 2007; **315**: 187d.
66 Herbert A, Gerry NP, McQueen MB *et al*: Response to Comments on 'A common genetic variant is associated with adult and childhood obesity'. *Science* 2007; **315**: 187e.
67 Lyon HN, Emilsson V, Hinney A *et al*: The association of a SNP upstream of INSIG2 with body mass index is reproduced in several but not all cohorts. *PLoS Genet* 2007; **3**: e61.
68 Cummings JR, Ahmad T, Geremia A *et al*: Contribution of the novel inflammatory bowel disease gene IL23R to disease susceptibility and phenotype. *Inflamm Bowel Dis* 2007; **13**: 1063–1068.
69 Van Limbergen JE, Russell RK, Nimmo ER *et al*: IL23R Arg381Gln is associated with childhood onset inflammatory bowel disease in Scotland. *Gut* 2007; **56**: 1173–1174.
70 Tremelling M, Cummings F, Fisher SA *et al*: IL23R variation determines susceptibility but not disease phenotype in inflammatory bowel disease. *Gastroenterology* 2007; **132**: 1657–1664.
71 Dubinsky MC, Wang D, Picornell Y *et al*: IL-23 receptor (IL-23R) gene protects against pediatric Crohn's disease. *Inflamm Bowel Dis* 2007; **13**: 511–515.
72 Hampe J, Franke A, Rosentiel P *et al*: A genome-wide association scan of nonsynonymous SNPs identifies a susceptibility variant for Crohn disease in ATG16L1. *Nat Genet* 2007; **39**: 207–211.
73 Cummings JR, Cooney R, Pathan S *et al*: Confirmation of the role of ATG16l1 as a Crohn's disease susceptibility gene. *Inflamm Bowel Dis* 2007; **13**: 941–946.
74 Prescott NJ, Fisher SA, Franke A *et al*: A nonsynonymous SNP in ATG16L1 predisposes to ileal Crohn's disease and is independent of CARD15 and IBD5. *Gastroenterology* 2007; **132**: 1665–1671.
75 Cauchi S, El Achhab Y, Choquet H *et al*: TCF7L2 is reproducibly associated with type 2 diabetes in various ethnic groups: a global meta-analysis. *J Mol Med* 2007; **85**: 777–782.

76 Cox A, Dunning AM, Garcia-Closas M *et al*: A common coding variant in CASP8 is associated with breast cancer risk. *Nat Genet* 2007; **39**: 352–358.
77 Amundadottir LT, Sulem P, Gudmundsson J *et al*: A common variant associated with prostate cancer in European and African populations. *Nat Genet* 2006; **38**: 652–658.
78 Haiman CA, Patterson N, Freedman ML *et al*: Multiple regions within 8q24 independently affect risk for prostate cancer. *Nat Genet* 2007; **39**: 638–644.
79 Todd JA, Walker NM, Cooper JD *et al*: Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. *Nat Genet* 2007; **39**: 857–864.
80 Lowe CE, Cooper JD, Brusko T *et al*: Large-scale genetic fine mapping and genotype-phenotype associations implicate polymorphism in the IL2RA region in type 1 diabetes. *Nat Genet* 2007; **39**: 1074–1082.
81 McPherson R, Pertsemlidis A, Kavaslar N *et al*: A common allele on chromosome 9 associated with coronary heart disease. *Science* 2007; **316**: 1488–1491.
82 Helgadottir A, Thorleifsson G, Manolescu A *et al*: A common variant on chromosome 9p21 affects the risk of myocardial infarction. *Science* 2007; **316**: 1491–1493.
83 Maraganore DM, de Andrade M, Lesnick TG *et al*: High-resolution whole-genome association study of Parkinson disease. *Am J Hum Genet* 2005; **77**: 685–693.
84 Clarimon J, Scholz S, Fung HC *et al*: Conflicting results regarding the semaphorin gene (SEMA5A) and the risk for Parkinson disease. *Am J Hum Genet* 2006; **78**: 1082–1084.
85 Goris A, Williams-Gray CH, Foltynie T *et al*: No evidence for association with Parkinson disease for 13 single-nucleotide polymorphisms identified by whole-genome association screening. *Am J Hum Genet* 2006; **78**: 1088–1090.
86 Li Y, Rowland C, Schrodi S *et al*: A case-control association study of the 12 single-nucleotide polymorphisms implicated in Parkinson disease by a recent genome scan. *Am J Hum Genet* 2006; **78**: 1090–1092.
87 Elbaz A, Nelson LM, Payami H *et al*: Lack of replication of thirteen single-nucleotide polymorphisms implicated in Parkinson's disease: a large-scale international study. *Lancet Neurol* 2006; **5**: 917–923.
88 Fung HC, Scholz S, Matarin M *et al*: Genome-wide genotyping in Parkinson's disease and neurologically normal controls: first stage analysis and public release of data. *Lancet Neurol* 2006; **5**: 911–916.
89 Matarin M, Brown WM, Scholz S *et al*: A genome-wide genotyping study in patients with ischaemic stroke: initial analysis and data release. *Lancet Neurol* 2007; **6**: 414–420.
90 Schymick JC, Scholz SW, Fung HC *et al*: Genome-wide genotyping in amyotrophic lateral sclerosis and neurologically normal controls: first stage analysis and public release of data. *Lancet Neurol* 2007; **6**: 322–328.
91 Klein RJ: Power analysis for genome-wide association studies. *BMC Genet* 2007; **8**: 58.
92 The GAIN Collaborative Research Group: New models of collaboration in genome-wide association studies: the Genetic Association Information Network. *Nat Genet* 2007; **39**: 1045–1051.
93 Zaykin DV, Zhivotovsky LA: Ranks of genuine associations in whole-genome scans. *Genetics* 2005; **171**: 813–823.
94 Ragoussis J, Elvidge GP, Kaur K *et al*: Matrix-assisted laser desorption/ionisation, time-of-flight mass spectrometry in genomics research. *PLoS Genet* 2006; **2**: e100.
95 Marchini J, Cardon LR, Phillips MS *et al*: The effects of human population structure on large genetic association studies. *Nat Genet* 2004; **36**: 512–517.
96 Price AL, Patterson NJ, Plenge RM *et al*: Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 2006; **38**: 904–909.
97 The ENCODE Project Consortium: Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 2007; **447**: 799–816.
98 The ENCODE Project Consortium: The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* 2004; **306**: 636–640.
99 Shriner D, Vaughan LK, Padilla MA, Tiwari HK: Problems with genome-wide association studies. *Science* 2007; **316**: 1840–1842.
100 Williams SM, Canter JA, Crawford DC *et al*: Problems with genome-wide association studies. *Science* 2007; **316**: 1840–1842.
101 Cohen JC, Kiss RS, Pertsemlidis A *et al*: Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science* 2004; **305**: 869–872.
102 Romeo S, Pennacchio LA, Fu Y *et al*: Population-based resequencing of ANGPTL4 uncovers variations that reduce triglycerides and increase HDL. *Nat Genet* 2007; **39**: 513–516.
103 Stefansson H, Rye DB, Hicks A *et al*: A genetic risk factor for periodic limb movements in sleep. *N Engl J Med* 2007; **357**: 639–647.
104 Winkelmann J, Schormair B, Lichtner P *et al*: Genome-wide association study of restless legs syndrome identifies common variants in three genomic regions. *Nat Genet* 2007; **39**: 1000–1006.
105 Samani NJ, Erdmann J, Hall AS *et al*: Genome wide association analysis of coronary artery disease. *N Engl J Med* 2007; **357**: 443–453.
106 The International Multiple Sclerosis Genetics Consortium: Risk alleles for multiple sclerosis identified by a genome wide study. *N Engl J Med* 2007; **357**: 851–862.
107 Buch S, Schafmayer C, Volzke H *et al*: A genome-wide association scan identifies the hepatic cholesterol transporter ABCG8 as a susceptibility factor for human gallstone disease. *Nat Genet* 2007; **39**: 995–999.
108 Thorleifsson G, Magnusson KP, Sulem P *et al*: Common sequence variants in the LOXL1 gene confer susceptibility to exfoliation glaucoma. *Science* 2007; **317**: 1397–1400.
109 Tomlinson I, Webb E, Carvajal-Carmona L *et al*: A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. *Nat Genet* 2007; **39**: 984–988.
110 Zanke BW, Greenwood CM, Rangrej J *et al*: Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24. *Nat Genet* 2007; **39**: 989–994.
111 Fellay J, Shianna KV, Ge D *et al*: A whole-genome association study of major determinants for host control of HIV-1. *Science* 2007; **317**: 944–947.
112 Hakonarson H, Grant SF, Bradfield JP *et al*: A genome-wide association study identifies KIAA0350 as a type 1 diabetes gene. *Nature* 2007; **448**: 591–594.
113 Moffatt MF, Kabesch M, Liang L *et al*: Genetic variants regulating ORMDL3 expression contributes to the risk of childhood asthma. *Nature* 2007; **448**: 470–473.
114 Gudbjartsson DF, Arnar DO, Helgadottir A *et al*: Variants conferring risk of atrial fibrillation on chromosome 4q25. *Nature* 2007; **448**: 353–357.
115 van Es MA, Van Vught PW, Blauw HM *et al*: ITPR2 as a susceptibility gene in sporadic amyotrophic lateral sclerosis: a genome-wide association study. *Lancet Neurol* 2007; **6**: 869–877.
116 Dunckley T, Huentelman MJ, Craig DW *et al*: Whole-genome analysis of sporadic amyotrophic lateral sclerosis. *N Engl J Med* 2007; **357**: 775–788.
117 Plenge RM, Seielstad M, Padyukov L *et al*: TRAF1-C5 as a risk locus for rheumatoid arthritis – a genome wide study. *N Engl J Med* 2007; **357**: 1199–1209.
118 Plenge RM, Cotsapas C, Davies L *et al*: Two independent alleles at 6q23 associated with risk of rheumatoid arthritis. *Nat Genet* 2007; **39**: 1477–1482.