American Journal of Medical Genetics Part B (Neuropsychiatric Genetics) 147B:1288–1294 (2008)

Review Article

The Positives, Protocols, and Perils of Genome-Wide Association

Benjamin M. Neale[1,2,3]* and Shaun Purcell[2,4,5]

[1]*Social, Genetic, and Developmental Psychiatry Centre, Institute of Psychiatry, King's College London, De Crespigny Park, London, UK*
[2]*The Broad Institute of Harvard and MIT, Cambridge, Massachusetts*
[3]*Center for Human Genetic Research, Massachusetts General Hospital, Harvard Medical School, Boson, Massachusetts*
[4]*Psychiatric and Neurodevelopmental Genetics Unit, Center for Human Genetic Research, Massachusetts General Hospital, Boston, Massachusetts*
[5]*Department of Psychiatry, Massachusetts General Hospital, Boston, Massachusetts*

Genome-wide association aims to comprehensively survey genetic variation for the purposes of disease and trait mapping. We provide a brief history of the development of genetic technology necessary to realize genome-wide association. From there we identify and review the publicly available resources for conducting such work including the molecular technologies, genomic databases, and analytic tools. Following on from the analytic tools, we highlight common analytic considerations, ranging from study design, quality control, and data cleaning to association analysis and replication. We conclude with a look toward future developments such as the analysis of copy number variation and integration of expression and epigenetic phenomenon into genome-wide association.

KEY WORDS: genome-wide association; linkage disequilibrium; review; SNP chip

Please cite this article as follows: Neale BM, Purcell S. 2008. The Positives, Protocols, and Perils of Genome-Wide Association. Am J Med Genet Part B 147B:1288–1294.

INTRODUCTION

Genome-wide association studies, now applied to a large range of human diseases and traits, are designed to comprehensively survey common genetic variation. The goal is to detect phenotypic associations of modest effect that would have eluded previous linkage and candidate gene approaches. Utilizing new genotyping technologies and genomic resources such as the HapMap [International Hapmap Consortium, 2005], a number of whole genome association studies have identified convincing and replicable disease loci for common diseases [Rioux et al., 2007; Saxena et al., 2007]. The approach looks set to accelerate gene discovery across a range of fields, including neuropsychiatric genetics.

A modern whole genome study typically involves genotyping hundreds of thousands of single nucleotide polymorphisms (SNPs) in thousands of individuals. Although genotyping at this density (on the order of a SNP per 5–10 kb) represents only a small proportion of the total number of known SNPs, it captures the majority of all common genetic variation, as we describe below, due to the extensive correlation between SNPs (linkage disequilibrium, LD). In a sufficiently large sample, this whole genome association study design promises the most extensive look at the genome for uncovering common variation predisposing to disease.

In this article, we briefly describe the history of genome-wide association studies (GWAS, also termed whole genome association studies, WGAS), followed by a review of some currently available resources, including molecular technologies, genomic databases, and analytic tools. We outline some key analytic considerations, such as study design, quality control and data cleaning, analysis and replication. Finally, we look to future developments such as copy number variation (CNV), total coverage and sequencing.

DEVELOPMENTS LEADING TO WHOLE GENOME STUDIES

Large-scale genomic projects paved the way for the shift from candidate gene association to GWAS by cataloguing and understanding genetic variation. Three main projects were critical: the Human Genome Project (HGP), the SNP Consortium and the International HapMap Project (HapMap) [Lander et al., 2001; Sachidanandam et al., 2001; International HapMap Consortium, 2005]. The HGP provides a consensus sequence, which dramatically enhanced the efforts of the SNP Consortium for SNP discovery. With the vast database of identified SNPs, the HapMap project embarked on identifying LD information enabling further development of cost effective genotyping platforms.

The proportion of human variation that needs to be captured for a study to classified as a GWAS is open for debate [Barrett and Cardon, 2006]. For the purposes of this article, a GWAS is required to have genotyped at least 80,000 SNPs or the majority of known non-synonymous variation. The earliest attempts at GWAS were not SNP chip based, but rather high-throughput genotyping of approximately 80,000 gene-centric variants from Yusuke Nakamura's lab [Ohnishi et al., 2001]. This group has published GWAS on myocardial infarction, nephropathy and Crohn's disease [Ozaki et al., 2002; Ohtsubo et al., 2005; Yamazaki et al., 2005]. However, the setup required to execute such a system is extensive and expensive. The subsequent development of comparatively cheap genotyping technologies with little to no overhead required made GWAS readily available, particularly if the investigator is willing to outsource genotyping.

The first major success story of 100K SNP chip GWAS is age-related macular degeneration (AMD), with the identification of variation in the complement factor H gene [Klein et al., 2005].

*Correspondence to: Benjamin M. Neale, Center for Human Genetic Research, Massachusetts General Hospital, Richard B. Simches Research Center, 185 Cambridge Street, Boston, MA 02114. E-mail: b.neale@iop.kcl.ac.uk

Received 17 July 2007; Accepted 4 February 2008

DOI 10.1002/ajmg.b.30747

Published online 23 May 2008 in Wiley InterScience (www.interscience.wiley.com)

Aside from AMD, the use of 100K SNP chips identified variation in NOS1AP (a.k.a. CAPON) influencing QT interval on an electrocardiogram [Arking et al., 2006]. Both of these findings showed significant replication from a number of additional studies, and are almost certainly true associations [Edwards et al., 2005; Hageman et al., 2005; Haines et al., 2005; Arking et al., 2006; Maller et al., 2006; Post et al., 2007]. The rapid success of mapping a significant percentage (~25%) of the risk factors for AMD has not been borne out by other diseases. However, a much smaller fraction of the risk factors for many other diseases have been identified (e.g., types I and II Diabetes, Crohn's disease, prostate and breast cancer).

Since these initial studies, a number of other groups have proceeded with GWAS. Efforts on obesity [Herbert et al., 2006], Parkinson's disease [Maraganore et al., 2005], type 2 diabetes [Saxena et al., 2007; Scott et al., 2007; Sladek et al., 2007; Steinthorsdottir et al., 2007], prostate cancer [Gudmundsson et al., 2007; Yeager et al., 2007], Crohn's disease [Rioux et al., 2007], and breast cancer [Easton et al., 2007] have been published.

Two major initiatives are generating genome-wide association data: the Wellcome Trust Case Control Consortium (WTCCC) and the Genetic Association Information Network (GAIN). The WTCCC is a UK study comprised of 2,000 case sample cohorts for each of the following diseases: tuberculosis, coronary heart disease, type 1 diabetes, type 2 diabetes, rheumatoid arthritis, Crohn's disease, bipolar disorder and hypertension, along with a 3,000 individual shared control sample. The control genotypes are already available at www.wtccc.org.uk and the case genotypes will be made publicly available. Initial results for these scans have recently been published, showing promising results for many, though not all of the disease phenotypes examined [Easton et al., 2007; Frayling et al., 2007; Parkes et al., 2007; Samani et al., 2007; Todd et al., 2007; Wellcome Trust Case Control Consortium, 2007; Zeggini et al., 2007]. In particular, the bipolar scan has shown little in the way of true association, indicating that psychiatric disease may prove more difficult than metabolic disorders. Similarly, a recent genome-wide association scan of bipolar disorder by Sklar and colleagues did not show consistent results with the WTCCC study, indicating that the effect size for risk variation for bipolar is likely to be modest [2008]. GAIN is a United States' National Institutes for Health initiative, generating genotypes on approximately 600K markers for schizophrenia, bipolar disorder, diabetic nephropathy, ADHD, major depression and psoriasis. More information about GAIN can be found at http://www.fnih.org/gain2/home_new.shtml.

RESOURCES

Numerous resources are available to aid whole genome studies, many of which were initially developed for linkage mapping, or have arisen from the HGP and HapMap. Here we present a brief list of some of these resources: further information is available at the websites noted.

SNP Chips

The commercial, technological development of SNP chips has been critical in the development of GWAS. These technologies allow for hundreds of thousands of genotypes per individual to be rapidly and affordably measured. Currently, Affymetrix and Illumina produce genome-wide arrays; Perlegen also provides genotyping, notably for GAIN. Both Affymetrix and Illumina have developed chips to genotype approximately one million SNPs; these products also provide CNV information (see Future Directions Section). More information about these products can be found at www.affymetrix.com and www.illumina.com. The true genomic coverage of these products is considerably greater than merely the number of SNPs because of the LD patterns. Briefly, LD is the non-random assortment of alleles within the population. One consequence of LD is that typing all variation in the genome is unnecessary as SNPs provide information for other loci. Already, the patterns from the HapMap are being used to test SNPs in a multi-marker framework [de Bakker et al., 2005; Pe'er et al., 2006] or to impute unknown SNPs [Marchini et al., 2007]. Generally, Illumina coverage tends to be slightly deeper because of the utilization of HapMap LD information.

Other studies have employed DNA pooling methodologies to reduce costs, estimating allele frequencies in cases and controls rather than individual genotyping. Examples of this approach have been published for nicotine dependence [Bierut et al., 2007; Uhl et al., 2007], bipolar disorder [Baum et al., 2008], osteoarthritis [Abel et al., 2006], supranuclear palsy [Melquist et al., 2007], and lung cancer [Spinola et al., 2007]. Other studies have focused only on non-synonymous variation at a genome-wide level: for example, Crohn's disease [Hampe et al., 2007], type 1 diabetes [Smyth et al., 2006], and Alzheimer's [Grupe et al., 2007].

Online Resources

A number of internet resources provide information for accessing and understanding the results of GWAS. The National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/) hosts a number of relevant resources, such as dbGAP, which hosts GWAS genotypes and results including GAIN. Such online databases, particularly when linked to existing resources such as PubMed (a searchable index of publications), GenBank (a genetic sequence database) and Entrez (a search engine for nucleotide, protein, structure, taxonomy, genome, expression, and chemical databases) will provide a powerful means to store, share, and mine GWAS data. The University of California at Santa Cruz hosts a genome browser at http://genome.ucsc.edu/cgi-bin/hggateway tailored more for comparative genomics. HapMap also provides a genome browser, annotated with LD information, useful for the identification of tagging SNPs and correlation across associated regions (www.hapmap.org).

A number of shared controls sets are available on the internet. The WTCCC provides their control dataset (pending an application process). All of the GAIN controls will be made available (but with use in publication delayed until after a nine month proprietary period as well as an application process). The Coriell Institute (https://queue.coriell.org/q) provides a case/control study of amyotrophic lateral sclerosis [Schymick et al., 2007].

Collaboration and Consortia

Essential for the success of GWAS is increasing sample size to detect variants of small effect. The WTCCC is an excellent example of collaboration with this aim. Benefits of such collaboration include the pooling of case samples from across the UK as well as drawing from the experience of analysts, genetics, and clinicians on major collections [Wellcome Trust Case Control Consortium, 2007]. Additionally, ongoing collaboration between the major Type II Diabetes projects (DGI, FUSION, and Novartis samples) [Saxena et al., 2007] has dramatically improved the detection of risk alleles. Similarly, the GAIN initiative is also comprised of collaborative efforts such as the International Multi-centre ADHD Genetics (IMAGE) project [Brookes et al., 2006; Kuntsi et al., 2006], the Major Depression Disorder project, and the Bipolar and Schizophrenia work.

Software

One difficult and important aspect of GWAS is managing the data. Given a sample size of five thousand individuals, with a million SNPs, such datasets contain five billion genotype datapoints. This presents a computational as well as a statistical burden of multiple testing. Adding multiple phenotypes, covariates and modifiers to the basic analysis adds further burden.

One approach is to use general statistical packages such as R, Stata, and SAS, which offer extensive statistical tests and models, but more limited genetic analyses (e.g., support for family-based studies, or haplotype analysis, for example). An R package, snpMatrix, is available to handle GWAS data and perform basic tests [Clayton and Leung, 2007]. A number of computational tools have been developed specifically for large-scale or whole-genome association studies: PLINK (www.pngu.mgh.harvard.edu/~purcell/plink) [Purcell et al., 2007]; PBAT (http://www.biostat.harvard.edu/~clange/default.htm) [Lange et al., 2004; Van Steen and Lange, 2005]; snptest (http://www.stats.ox.ac.uk/~marchini/software/gwas/snptest.html) [Marchini et al., 2007; Wellcome Trust Case Control Consortium, 2007]; and EIGENSTRAT/EIGENSOFT (http://genepath.med.harvard.edu/~reich/eigenstrat.htm) [Patterson et al., 2006; Price et al., 2006]. Haploview 4.0 (http://www.broad.mit.edu/mpg/haploview/) [Barrett et al., 2005] has been extended to provide a browser for GWAS results integrated with PLINK; it also will download the HapMap data to generate LD and tagging information for a specific region of the genome.

ANALYTIC CONSIDERATIONS

Study Design

Many aspects of WGAS study design are similar to candidate gene association analysis. Both case–control and family-based association study designs can be employed. Thus far, most WGAS are case–control, primarily because of the increased power per genotype compared to family-based designs [McGinnis et al., 2002]. Good experimental procedure such as randomization of case and controls across plates are important to protect against bias. Matching of controls to the cases, with a particular focus on ancestry is recommended. The magnitude of WGAS datasets brings some other study design issues into play, however. One is the utility of multi-stage designs, which have been suggested as an approach to control costs [Van den Oord, 1999; Skol et al., 2006], although the relative costs of different genotyping platforms are constantly changing. Because these datasets are expensive to collect and a fixed marker set is employed (for a given genotyping platform), the idea of using shared control datasets is both desirable and feasible. This factor also brings some difficult challenges however: the ability to ensure consistency across different samples, genotype calling algorithms and/or laboratory procedures; the trade-off in terms of power and false positives between adding increasingly less well-matched controls; the interpretation of replication if two studies use different case samples but the same control sample [Hamer and Sirota, 2000].

Quality Control

Ensuring the quality of the genotype data from GWAS is essential for drawing accurate conclusions from association analysis. Considering a dataset of a million SNPs, if only 0.5% of the SNPs are systematically biased assays, this still corresponds to 5,000 biased tests, potentially yielding an unacceptable false positive rate. To control for such pitfalls, data quality thresholds are applied. In general, the key motivation behind quality control is that as the prior probability of a SNP showing true significance is low, discarding SNPs for reasons such as missingness, minor allele frequency, mendelization errors, and Hardy–Weinberg disequilibrium, is unlikely to remove true associations. Many of the cleaning quality metrics described below are consistent with previous WGAS [Saxena et al., 2007; Wellcome Trust Case Control Consortium, 2007] and review of good experimental procedure for such studies [Manolio et al., 2007]. A tension between genotype information and controlling for bias still exists, but with procedures such as imputation, such concerns are assuaged [Marchini et al., 2007].

A good indicator of genotype probe performance for SNP chips is the call rate across the sample. We recommend examining the distribution of missingness across the sample to identify problematic SNPs. In addition to a global missing threshold, comparing missingness between cases and controls, via a Chi square, is suggested. Similar considerations for the level of genotyping of each individual are also recommended, as low genotyping rate is a marker for poor DNA quality. As an example of problem of missingness, the second highest SNP from the AMD GWAS, rs10272438, was a false positive due to missingness [Klein et al., 2005]. Approximately 15% of the genotypes failed which when genotyped using another technology showed no association. In fact, differential missing rates between cases and controls can induce false positive association [Clayton et al., 2005].

Another key measure of the quality of the genotypes is reproducibility, as assessed through intentional sample duplication. For example, HapMap samples can be used to generate quality control metrics based on sample concordance with the existing genotyping. Additionally, HapMap individuals are uniquely identifiable, and so can act as positive controls for potential laboratory mishandling (e.g., plate orientation). If a family-based design is adopted, then Mendelian checks also provide a first pass at sample integrity. As some random errors are generally expected, the thresholds for Mendelian inconsistencies and sample duplication mismatch tend to be less conservative, such that the probability of observing the number of errors is unlikely to be due to chance.

Potential batch effects are also important to examine. Often times, all samples are not done with the same product at the same time suggesting the possibility of batch effects. Considering the availability of shared control sets, such phenomena are commonplace for WGAS. Other lines of enquiry for batch effects include: different DNA sources (e.g., blood vs. buccal vs. saliva), different extraction techniques, different centers contributing DNA, different technical procedures, or plate effects. A look at the data in chronological genotyping order may also yield insight into potential sources of error, as stock changes in the lab may prove important.

Minor allele frequency (MAF) thresholds are also recommended as many studies do not have the power to detect significant association for very rare variation. Of course, the MAF threshold is dependent on the sample size, but a decent rule of thumb is observing at least 20–30 copies of the minor allele in the total sample. Current genotyping calling algorithms rely on clustering points on an intensity scale, and so rare genotypes are also more prone to error (e.g., it is difficult to define a cluster with only one observation). Comparing observed genotype frequencies in controls against the HapMap allele frequency can also provide evidence for bias.

Testing for deviation from Hardy–Weinberg equilibrium (HWE) may provide further information about the validity of the genotypes from a SNP. However, such endeavors are confounded by both population stratification and true association signal. Therefore, markers passing all criteria except for HWE ought to be considered carefully rather than discarded out of hand. Another approach is to define a more stringent

threshold, such as 0.000001 for deviation from HWE. HWE tests can be calculated on only the controls or in the entire sample. The justification for considering only the controls for HWE is that positive association may confound the HWE test [Sham, 1997].

Beyond these initial cleaning techniques, further checks for family structure are suggested in the case of family-based data. Non-paternity is a potential problem for trio and sibship designs, which can be easily detected by looking at identity-by-state (IBS). For case–control designs, the same IBS information can be used to determine identity-by-descent (IBD) information across the sample (see Purcell et al. [2007] for more details on the relationship between IBS and IBD at a population level). Examining both IBS and IBD information can identify sample mix-up (via different IBD patterns), cryptic relatedness (high IBD sharing), and sample contamination (excess heterozygosity and IBD).

For case–control and population-based quantitative analysis, population stratification is a key potential confounding factor. With whole-genome association data, however, the ability to identify population structure is dramatically improved. PLINK includes routines to cluster individuals based on IBS sharing for population classification. Aside from assigning individuals to clusters, a correction to the inflation of the association statistic can be applied by principal components analysis [Price et al., 2006].

For a given associated SNP, it is worthwhile to see whether nearby SNPs or haplotypes that are correlated with the variant also show association with disease; if the associated SNP is rare or has a high missing rate, confirming that the association is also seen with haplotypes formed by common, high genotyping SNPs is, whenever possible, desirable. A SNP that shows a strong association but for which all the correlated, neighboring variants are not associated, is more likely to represent an artifact.

As a final check, the distribution of association test statistics is a useful indicator for sources of bias. Gross enrichment of the distribution of the association evidence is a hallmark sign of bias. Furthermore, extremely significant *P*-values, such at 10^{-60} are more likely than not due to batch effects, non-random missingness or data-handling errors. For further information about data cleaning considerations, we recommend a recent feature in Nature from NCI-NHGRI [Chanock et al., 2007] and the WTCCC manuscript [Wellcome Trust Case Control Consortium, 2007].

ANALYSIS

Three most common analytic techniques for case–control analysis are the χ^2 test of allele counts, trend tests (where a multiplicative model is assumed for the regression based on genotype category, coded as 0, 1, and 2), and a 2 degree of freedom genotypic model (where one genotype category is assumed as baseline and the effects of the other two categories are modeled). For family-based analysis, the TDT [Spielman et al., 1993] for trios and the sib-TDT [Spielman and Ewens, 1998] (using siblings discordant for disease) are obvious choices. Quantitative methods include regression models for population-data, following the similar parameterization as the case–control, while quantitative approaches have been developed for families [Rabinowitz, 1997; Fulker et al., 1999; Lange et al., 2004].

As well as testing directly genotyped SNPs, consideration of haplotype structure enables one to test ungenotyped variation. One approach would be to specify haplotypes based on sliding windows of SNPs, or on haplotype blocks based on the LD structure of the observed data. An alternate approach is to use information from the HapMap to specify more precise haplotype tests specifically for the HapMap SNPs that were not directly genotyped in the study. For example, for a fixed genotyping platform, Pe'er et al. [2006] compiled lists of single SNPs and two and three SNP haplotypes that are in strong LD with ungenotyped HapMap SNPs.

Beyond these initial tests, a number of other techniques are frequently employed. Based on the tagging information from HapMap, tests of two and three marker haplotypes which are proxies for known variants can be conducted. Furthermore, different imputation methods are being developed to generate genotypes at untyped loci jointly with information from a reference panel based on LD patterns in the HapMap and further untyped variation based on ancestral recombination graphs [Marchini et al., 2007]. The benefit of imputation is still to be fully evaluated: with increasing chip densities, the majority of common variation may well be directly captured. Perhaps one particularly useful application of imputation will be to reconcile results and merge data for WGAS studies that have used different genotyping platforms. Finally, multi-marker tests that consider whole pathways and genes simultaneously, instead of single variants, are another area of promise.

All of the above methods fall broadly under traditional association analyses and are targeted at the common diseases/common variant hypothesis (CDCV), that variation predisposing to disease within the population will be common within the population and of modest effect. In contrast, the multiple rare variant (MRV) hypothesis states that variation predisposing to disease is rare and of small to modest effect (with the extreme example being that every case for a given disease has a set of private mutations). In all likelihood, both the CDCV and the MRV are likely to be true for the etiology of common disease within the population. How WGAS studies of common SNPs will fare when the MRV is true for a substantial proportion of the genetic variation for a particular disease is unclear. New methods and models are being developed that might partially address this problem. For example, comparing LD information between cases and controls may shed insight on rare variation [Zaykin et al., 2006]. Alternatively, using WGAS data, one might look for regions of increased ancestral sharing between cases, as individuals sharing the same rare variant are also likely to share an extended, surrounding region [Purcell et al., 2007]. Homozygosity and admixture analyses are additional lines of enquiry for the mapping of risk-conferring variation [Lander and Green, 1987; Reich and Patterson, 2005]. Ultimately, sequence data will likely become routinely available, to complement common polymorphism data and drive the investigation of rare variation.

Multiple Testing

The number of association tests for WGAS is staggering. Standard approaches for multiple testing including Bonferroni and False Discovery Rate (FDR) can be used to control the error rate of the study. For family-based association, one potential analytic possibility is to condition on the between family information to select SNPs for the within family test, to reduce the necessary multiple testing burden [Lange et al., 2004]. By selecting, the necessary number of SNPs for genome-wide significance under Bonferroni is reduced to the number of SNPs analyzed in the within test. However, as the between and within information are independent, it may be more efficient to combine the evidence for association [Skol et al., 2006]. Risch and Merikangas [1996] proposed a threshold of 10^{-6} based on the number of known SNPs at the time, though a more realistic threshold is perhaps on the order of 10^{-7} or even 10^{-8} assuming approximately a million testable variants using the Šidák Correction [Sidák, 1967]. Permutation analysis is also an avenue for generating an appropriate experiment-wide *P*-values, but such efforts may not appropriately control for all SNPs potentially tested.

Replication and follow-up studies are essential for determining whether identified variants are true or false positives (although it is worth remembering that if hundreds or even thousands of SNPs are followed up, then a predictable proportion will replicate purely by chance also). With replication and follow-up come the difficulties of meta-analysis. Ideally, data sharing is encouraged to provide maximal information about the association evidence. If this is not possible, then combining evidence based on the direction and magnitude of the effect is encouraged. As a last resort, Fisher's combination of *P*-values can be utilized.

FUTURE DIRECTIONS

Recently, a coalition of clinicians, geneticists, and analysts have formed the Psychiatric GWAS Consortium (PGC) [The Psychiatric GWAS Consortium, submitted], which aims to encourage data-sharing and perform a comprehensive meta-analysis of genome-wide association studies of psychiatric disease. The current focus is on ADHD, autism, bipolar, major depression, and schizophrenia, looking both within and across disorders. In total, there will be in excess of 25 billion genotypes for meta-analysis, representing the largest genetic study in psychiatry ever conducted.

Genetics as a field continues to develop technologies for studying the human genome at finer and finer scales. The most recent SNP Chip technologies provide some insight into CNVs. CNVs are loosely defined as approximately 1 kb or longer regions of the genome which show variation in the number of copies as compared to a given reference sequence [Feuk et al., 2006]. Already a handful of studies have been published on the effects of CNVs on gene expression and phenotypes [McCarroll et al., 2006; Sebat et al., 2007; Stranger et al., 2007; Wong et al., 2007].

Complete coverage of the genome with respect to LD and eventually full sequence information will be available for analysis. Such extensive information will require even more careful data management. However, many of the existing tools for analysis can be applied to such data. Sequencing enables examination of rarer variation as a potential cause of disease. The analysis of such variation will likely require the development of new statistical models. In addition to identifying the genetic code, expression and epigenetic information will also reduce in cost. For an excellent review of global gene expression see Rockman and Kruglyak [2006], encompassing the genetics of global gene expression thus far, features of regulatory sequence variation, and genomic effects such as *cis*-acting, *trans*-acting, *cis*-regulatory, and protein-coding on gene expression. Epigenetics examines DNA structure (e.g., histone placement) and methylation patterns; for a review see van Vliet et al. [2007].

CONCLUSIONS

WGAS promise the most extensive look at the genome for uncovering variation predisposing to disease. Technology will continue to develop yielding a wealth of data for identifying the etiology of disease. While WGAS will not identify all of the genetic factors, new biochemical pathways will be identified for investigation. For many diseases, which are known to be strongly heritable, finding even one or two true disease genes could potentially transform the research in that disease area, even if the majority of genetic determinants elude detection in that particular study. Given the difficulty of mapping genetic variation for neuropsychiatric disease, even greater care is necessary for successful association mapping.

REFERENCES

Abel K, Reneland R, Kammerer S, Mah S, Hoyal C, Cantor CR, Nelson MR, Braun A. 2006. Genome-wide SNP association: Identification of susceptibility alleles for osteoarthritis. Autoimmun Rev 5(4):258–263.

Arking DE, Pfeufer A, Post W, Kao WH, Newton-Cheh C, Ikeda M, West K, Kashuk C, Akyol M, Perz S, et al. 2006. A common genetic variant in the NOS1 regulator NOS1AP modulates cardiac repolarization. Nat Genet 38(6):644–651.

Barrett JC, Cardon LR. 2006. Evaluating coverage of genome-wide association studies. Nat Genet 38(6):659–662.

Barrett JC, Fry B, Maller J, Daly MJ. 2005. Haploview: Analysis and visualization of LD and haplotype maps. Bioinformatics 21(2):263–265.

Baum AE, Akula N, Cabanero M, Cardona I, Corona W, Klemens B, Schulze TG, Cichon S, Rietschel M, Nothen MM, et al. 2008. A genome-wide association study implicates diacylglycerol kinase eta (DGKH) and several other genes in the etiology of bipolar disorder. Mol Psychiatry 13(2):197–207.

Bierut LJ, Madden PA, Breslau N, Johnson EO, Hatsukami D, Pomerleau OF, Swan GE, Rutter J, Bertelsen S, Fox L, et al. 2007. Novel genes identified in a high-density genome wide association study for nicotine dependence. Hum Mol Genet 16(1):24–35.

Brookes K, Xu X, Chen W, Zhou K, Neale B, Lowe N, Anney R, Franke B, Gill M, Ebstein R, et al. 2006. The analysis of 51 genes in DSM-IV combined type attention deficit hyperactivity disorder: Association signals in DRD4, DAT1 and 16 other genes. Mol Psychiatry 11(10):934–953.

Chanock SJ, Manolio T, Boehnke M, Boerwinkle E, Hunter DJ, Thomas G, Hirschhorn JN, Abecasis G, Altshuler D, Bailey-Wilson JE, et al. 2007. Replicating genotype-phenotype associations. Nature 447(7145):655–660.

Clayton D, Leung HT. 2007. An R package for analysis of whole-genome association studies. Hum Hered 64(1):45–51.

Clayton DG, Walker NM, Smyth DJ, Pask R, Cooper JD, Maier LM, Smink LJ, Lam AC, Ovington NR, Stevens HE, et al. 2005. Population structure, differential bias and genomic control in a large-scale, case-control association study. Nat Genet 37(11):1243–1246.

de Bakker PI, Yelensky R, Pe'er I, Gabriel SB, Daly MJ, Altshuler D. 2005. Efficiency and power in genetic association studies. Nat Genet 37(11): 1217–1223.

Easton DF, Pooley KA, Dunning AM, Pharoah PD, Thompson D, Ballinger DG, Struewing JP, Morrison J, Field H, Luben R, et al. 2007. Genome-wide association study identifies novel breast cancer susceptibility loci. Nature 447(7148):1087–1093.

Edwards AO, Ritter R III, Abel KJ, Manning A, Panhuysen C, Farrer LA. 2005. Complement factor H polymorphism and age-related macular degeneration. Science 308(5720):421–424.

Feuk L, Marshall CR, Wintle RF, Scherer SW. 2006. Structural variants: Changing the landscape of chromosomes and design of disease studies. Hum Mol Genet 15(Spec no. 1):R57–R66.

Frayling TM, Timpson NJ, Weedon MN, Zeggini E, Freathy RM, Lindgren CM, Perry JR, Elliott KS, Lango H, Rayner NW, et al. 2007. A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. Science 316(5826):889–894.

Fulker DW, Cherny SS, Sham PC, Hewitt JK. 1999. Combined linkage and association sib-pair analysis for quantitative traits. Am J Hum Genet 64:259–267.

Grupe A, Abraham R, Li Y, Rowland C, Hollingworth P, Morgan A, Jehu L, Segurado R, Stone D, Schadt E, et al. 2007. Evidence for novel susceptibility genes for late-onset Alzheimer's disease from a genome-wide association study of putative functional variants. Hum Mol Genet 16(8):865–873.

Gudmundsson J, Sulem P, Manolescu A, Amundadottir LT, Gudbjartsson D, Helgason A, Rafnar T, Bergthorsson JT, Agnarsson BA, Baker A, et al. 2007. Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24. Nat Genet 39(5):631–637.

Hageman GS, Anderson DH, Johnson LV, Hancox LS, Taiber AJ, Hardisty LI, Hageman JL, Stockman HA, Borchardt JD, Gehrs KM, et al. 2005. A common haplotype in the complement regulatory gene factor H (HF1/CFH) predisposes individuals to age-related macular degeneration. Proc Natl Acad Sci USA 102(20):7227–7232.

Haines JL, Hauser MA, Schmidt S, Scott WK, Olson LM, Gallins P, Spencer KL, Kwan SY, Noureddine M, Gilbert JR, et al. 2005. Complement factor H variant increases the risk of age-related macular degeneration. Science 308(5720):419–421.

Hamer D, Sirota L. 2000. Beware the chopsticks gene. Mol Psychiatry 5(1):11–13.

Hampe J, Franke A, Rosenstiel P, Till A, Teuber M, Huse K, Albrecht M, Mayr G, De La Vega FM, Briggs J, et al. 2007. A genome-wide association scan of nonsynonymous SNPs identifies a susceptibility variant for Crohn disease in ATG16 L1. Nat Genet 39(2):207–211.

Herbert A, Gerry NP, McQueen MB, Heid IM, Pfeufer A, Illig T, Wichmann HE, Meitinger T, Hunter D, Hu FB, et al. 2006. A common genetic variant is associated with adult and childhood obesity. Science 312 (5771):279–283.

International Hapmap Consortium. 2005. A haplotype map of the human genome. Nature 437:1299–1320.

Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, Henning AK, SanGiovanni JP, Mane SM, Mayne ST, et al. 2005. Complement factor H polymorphism in age-related macular degeneration. Science 308(5720): 385–389.

Kuntsi J, Neale BM, Chen W, Faraone SV, Asherson P. 2006. The IMAGE project: Methodological issues for the molecular genetic analysis of ADHD. Behav Brain Funct 2:27.

Lander ES, Green P. 1987. Construction of multilocus genetic linkage maps in humans. Proc Natl Acad Sci USA 84(8):2363–2367.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al. 2001. Initial sequencing and analysis of the human genome. Nature 409(6822):860–921.

Lange C, DeMeo D, Silverman EK, Weiss ST, Laird NM. 2004. PBAT: Tools for family-based association studies. Am J Hum Genet 74(2):367–369.

Maller J, George S, Purcell S, Fagerness J, Altshuler D, Daly MJ, Seddon JM. 2006. Common variation in three genes, including a noncoding variant in CFH, strongly influences risk of age-related macular degeneration. Nat Genet 38(9):1055–1059.

Manolio TA, Rodriguez LL, Brooks L, Abecasis G, Ballinger D, Daly M, Donnelly P, Faraone SV, Frazer K, Gabriel S, et al. 2007. New models of collaboration in genome-wide association studies: The Genetic Association Information Network. Nat Genet 39(9):1045–1051.

Maraganore DM, de Andrade M, Lesnick TG, Strain KJ, Farrer MJ, Rocca WA, Pant PV, Frazer KA, Cox DR, Ballinger DG. 2005. High-resolution whole-genome association study of Parkinson disease. Am J Hum Genet 77(5):685–693.

Marchini J, Howie B, Myers S, McVean G, Donnelly P. 2007. A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet 39(7):906–913.

McCarroll SA, Hadnott TN, Perry GH, Sabeti PC, Zody MC, Barrett JC, Dallaire S, Gabriel SB, Lee C, Daly MJ, et al. 2006. Common deletion polymorphisms in the human genome. Nat Genet 38(1):86–92.

McGinnis R, Shifman S, Darvasi A. 2002. Power and efficiency of the TDT and case-control design for association scans. Behav Genet 32(2):135–144.

Melquist S, Craig DW, Huentelman MJ, Crook R, Pearson JV, Baker M, Zismann VL, Gass J, Adamson J, Szelinger S, et al. 2007. Identification of a novel risk locus for progressive supranuclear palsy by a pooled genomewide scan of 500,288 single-nucleotide polymorphisms. Am J Hum Genet 80(4):769–778.

Ohnishi Y, Tanaka T, Ozaki K, Yamada R, Suzuki H, Nakamura Y. 2001. A high-throughput SNP typing system for genome-wide association studies. J Hum Genet 46(8):471–477.

Ohtsubo S, Iida A, Nitta K, Tanaka T, Yamada R, Ohnishi Y, Maeda S, Tsunoda T, Takei T, Obara W, et al. 2005. Association of a single-nucleotide polymorphism in the immunoglobulin mu-binding protein 2 gene with immunoglobulin A nephropathy. J Hum Genet 50(1):30–35.

Ozaki K, Ohnishi Y, Iida A, Sekine A, Yamada R, Tsunoda T, Sato H, Sato H, Hori M, Nakamura Y, et al. 2002. Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction. Nat Genet 32(4):650–654.

Parkes M, Barrett JC, Prescott NJ, Tremelling M, Anderson CA, Fisher SA, Roberts RG, Nimmo ER, Cummings FR, Soars D, et al. 2007. Sequence variants in the autophagy gene IRGM and multiple other replicating loci contribute to Crohn's disease susceptibility. Nat Genet 39(7):830–832.

Patterson N, Price AL, Reich D. 2006. Population structure and eigenanalysis. PLoS Genet 2(12):e190.

Pe'er I, de Bakker PI, Maller J, Yelensky R, Altshuler D, Daly MJ. 2006. Evaluating and improving power in whole-genome association studies using fixed marker sets. Nat Genet 38(6):663–667.

Post W, Shen H, Damcott C, Arking DE, Kao WH, Sack PA, Ryan KA, Chakravarti A, Mitchell BD, Shuldiner AR. 2007. Associations between genetic variants in the NOS1AP (CAPON) gene and cardiac repolarization in the old order Amish. Hum Hered 64(4):214–219.

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38(8):904–909.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, et al. 2007. PLINK: A tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81(3):559–575.

Rabinowitz D. 1997. A transmission disequilibrium test for quantitative trait loci. Hum Hered 47(6):342–350.

Reich D, Patterson N. 2005. Will admixture mapping work to find disease genes? Philos Trans R Soc Lond B Biol Sci 360(1460):1605–1607.

Rioux JD, Xavier RJ, Taylor KD, Silverberg MS, Goyette P, Huett A, Green T, Kuballa P, Barmada MM, Datta LW, et al. 2007. Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis. Nat Genet 39(5):596–604.

Risch N, Merikangas K. 1996. The future of genetic studies of complex human diseases. Science 273(5281):1516–1517.

Rockman MV, Kruglyak L. 2006. Genetics of global gene expression. Nat Rev Genet 7(11):862–872.

Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, Sherry S, Mullikin JC, Mortimore BJ, Willey DL, et al. 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 409(6822):928–933.

Samani NJ, Erdmann J, Hall AS, Hengstenberg C, Mangino M, Mayer B, Dixon RJ, Meitinger T, Braund P, Wichmann HE, et al. 2007. Genome-wide association analysis of coronary artery disease. N Engl J Med 357(5):443–453.

Saxena R, Voight BF, Lyssenko V, Burtt NP, de Bakker PI, Chen H, Roix JJ, Kathiresan S, Hirschhorn JN, Daly MJ, et al. 2007. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. Science 316(5829):1331–1336.

Schymick JC, Scholz SW, Fung HC, Britton A, Arepalli S, Gibbs JR, Lombardo F, Matarin M, Kasperaviciute D, Hernandez DG, et al. 2007. Genome-wide genotyping in amyotrophic lateral sclerosis and neurologically normal controls: First stage analysis and public release of data. Lancet Neurol 6(4):322–328.

Scott LJ, Mohlke KL, Bonnycastle LL, Willer CJ, Li Y, Duren WL, Erdos MR, Stringham HM, Chines PS, Jackson AU, et al. 2007. A genome-wide association study of type 2 diabetes in finns detects multiple susceptibility variants. Science 316(5829):1341–1345.

Sebat J, Lakshmi B, Malhotra D, Troge J, Lese-Martin C, Walsh T, Yamrom B, Yoon S, Krasnitz A, Kendall J, et al. 2007. Strong association of de novo copy number mutations with autism. Science 316(5823):445–449.

Sham P. 1997. Statistics in human genetics. London New York: Arnold; Wiley. Vol. viii, 290 p.

Šidák Z. 1967. Rectangular confidence regions for the means of multivariate distributions. J Am Stat Assoc 62:626–633.

Sklar P, Smoller JW, Fan J, Ferreira MA, Perlis RH, Chambert K, Nimgaonkar VL, McQueen MB, Faraone SV, Kirby A, et al. 2008. Whole-genome association study of bipolar disorder. Mol Psychiatry DOI: 10.1038/sj.mp.4002151.

Skol AD, Scott LJ, Abecasis GR, Boehnke M. 2006. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. Nat Genet 38:209–213.

Sladek R, Rocheleau G, Rung J, Dina C, Shen L, Serre D, Boutin P, Vincent D, Belisle A, Hadjadj S, et al. 2007. A genome-wide association study identifies novel risk loci for type 2 diabetes. Nature 445(7130):881–885.

Smyth DJ, Cooper JD, Bailey R, Field S, Burren O, Smink LJ, Guja C, Ionescu-Tirgoviste C, Widmer B, Dunger DB, et al. 2006. A genome-wide association study of nonsynonymous SNPs identifies a type 1 diabetes locus in the interferon-induced helicase (IFIH1) region. Nat Genet 38(6): 617–619.

Spielman RS, Ewens WJ. 1998. A sibship test for linkage in the presence of association: The Sib Transmission/Disequilibrium Test. Am J Hum Genet 62:450–458.

Spielman RS, McGinnis RE, Ewens WJ. 1993. Transmission test for linkage disequilibrium: The insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet 52(3):506–516.

Spinola M, Leoni VP, Galvan A, Korsching E, Conti B, Pastorino U, Ravagnani F, Columbano A, Skaug V, Haugen A, et al. 2007. Genome-wide single nucleotide polymorphism analysis of lung cancer risk detects the KLF6 gene. Cancer Lett 251(2):311–316.

Steinthorsdottir V, Thorleifsson G, Reynisdottir I, Benediktsson R, Jonsdottir T, Walters GB, Styrkarsdottir U, Gretarsdottir S, Emilsson V, Ghosh S, et al. 2007. A variant in CDKAL1 influences insulin response and risk of type 2 diabetes. Nat Genet 39(6):770–775.

Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N, Redon R, Bird CP, de Grassi A, Lee C, et al. 2007. Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science 315(5813):848–853.

Todd JA, Walker NM, Cooper JD, Smyth DJ, Downes K, Plagnol V, Bailey R, Nejentsev S, Field SF, Payne F, et al. 2007. Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. Nat Genet 39(7):857–864.

Uhl GR, Liu QR, Drgon T, Johnson C, Walther D, Rose JE. 2007. Molecular genetics of nicotine dependence and abstinence: Whole genome association using 520,000 SNPs. BMC Genet 8:10.

Van den Oord EJCG. 1999. A comparison between different designs and tests to detects QTLs in association studies. Behav Genet 29(4):245–256.

Van Steen K, Lange C. 2005. PBAT: A comprehensive software package for genome-wide association analysis of complex family-based studies. Hum Genomics 2(1):67–69.

van Vliet J, Oates NA, Whitelaw E. 2007. Epigenetic mechanisms in the context of complex diseases. Cell Mol Life Sci 64(12):1531–1538.

Wong KK, deLeeuw RJ, Dosanjh NS, Kimm LR, Cheng Z, Horsman DE, MacAulay C, Ng RT, Brown CJ, Eichler EE, et al. 2007. A comprehensive analysis of common copy-number variations in the human genome. Am J Hum Genet 80(1):91–104.

Wellcome Trust Case Control Consortium. 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447(7145):661–678.

Yamazaki K, McGovern D, Ragoussis J, Paolucci M, Butler H, Jewell D, Cardon L, Takazoe M, Tanaka T, Ichimori T, et al. 2005. Single nucleotide polymorphisms in TNFSF15 confer susceptibility to Crohn's disease. Hum Mol Genet 14(22):3499–3506.

Yeager M, Orr N, Hayes RB, Jacobs KB, Kraft P, Wacholder S, Minichiello MJ, Fearnhead P, Yu K, Chatterjee N, et al. 2007. Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. Nat Genet 39(5):645–649.

Zaykin DV, Meng Z, Ehm MG. 2006. Contrasting linkage-disequilibrium patterns between cases and controls as a novel association-mapping method. Am J Hum Genet 78(5):737–746.

Zeggini E, Weedon MN, Lindgren CM, Frayling TM, Elliott KS, Lango H, Timpson NJ, Perry JR, Rayner NW, Freathy RM, et al. 2007. Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. Science 316(5829):1336–1341.