Am. J. Hum. Genet. 77:337–345, 2005

Recent Developments in Genomewide Association Scans: A Workshop Summary and Review

Duncan C. Thomas,[1] Robert W. Haile,[1] and David Duggan[2]

[1]Department of Preventive Medicine, University of Southern California, Los Angeles; and [2]Translational Genomics Research Institute (TGen), Phoenix

With the imminent availability of ultra-high-volume genotyping platforms (on the order of 100,000–1,000,000 genotypes per sample) at a manageable cost, there is growing interest in the possibility of conducting genomewide association studies for a variety of diseases but, so far, little consensus on methods to design and analyze them. In April 2005, an international group of >100 investigators convened at the University of Southern California over the course of 2 days to compare notes on planned or ongoing studies and to debate alternative technologies, study designs, and statistical methods. This report summarizes these discussions in the context of the relevant literature. A broad consensus emerged that the time was now ripe for launching such studies, and several common themes were identified—most notably the considerable efficiency gains of multistage sampling design, specifically those made by testing only a portion of the subjects with a high-density genomewide technology, followed by testing additional subjects and/or additional SNPs at regions identified by this initial scan.

Introduction

A traditional means of discovering disease genes begins with family-based linkage scans, looking for regions of the genome that tend to be transmitted through families in a manner that parallels the transmission of the trait, followed by a variety of fine-mapping techniques. This approach has been highly successful for mapping major genes responsible for Mendelian disorders, in part because the breadth of a linkage signal means that a genomewide scan can be accomplished with a few hundred microsatellite or a few thousand SNP markers. However, finer resolution of the putative risk susceptibility loci through linkage analyses will only be feasible with the availability of sufficient recombination events, requiring large pedigrees (Boehnke 1994), and the utility of the linkage approach for identifying multiple low-penetrance variants involved in common diseases has been questioned.

As an alternative, the past decade has seen a rapid escalation in hypothesis-driven candidate gene association studies or fine-mapping studies exploiting linkage disequilibrium (LD), but these have usually been restricted to a few dozen genes. Recent advances in high-volume genotyping technology have now made it possible to consider using empirical LD patterns to search the genome for risk-associated variants. These studies are based on the premise that "unrelated" individuals are more distantly related than subjects from large pedigrees, thus allowing for sufficient recombination events to have taken place (Nordborg and Taveré 2002). Coupled with the efforts by the International Haplotype Mapping ("HapMap") Project (Gibbs et al. 2003) to catalog millions of SNPs and haplotypes across diverse populations and to use these to identify subsets of highly informative "tag" SNPs, genomewide association scans involving hundreds of thousands or more markers on thousands of subjects—first suggested a decade ago by Risch and Merikangas (1996)—are now a real possibility.

Numerous research groups are planning or have underway genomewide searches for a range of disorders and the first reports of such studies (using early versions of high-density SNP chips) are just beginning to appear (Ozaki et al. 2002; Klein et al. 2005). These groups are using a variety of population-based and family-based epidemiological designs or model organisms, but so far there has been little general discussion of how best to design and analyze such studies. In April 2005, an international group of 165 investigators met at the University of Southern California for a 2-day workshop to discuss their efforts and consider various methodological problems. This report provides a brief summary of the major themes addressed at this workshop and a review of the relevant background literature. The reader is also referred to several recent review articles (Hirsch-

Received June 14, 2005; accepted for publication June 20, 2005; electronically published August 1, 2005.

Address for correspondence and reprints: Dr. Duncan C. Thomas, Department of Preventive Medicine, University of Southern California, 1540 Alcazar Street, Los Angeles, CA 90089-9011. E-mail: dthomas@usc.edu

0002-9297/2005/7703-0002$15.00

horn and Daly 2005; Wang et al. 2005; Palmer and Cardon, in press).

The workshop started with descriptions of some ongoing and planned genomewide association studies. Eleven studies were described and the diseases under study included breast, colorectal, and prostate cancer, type I diabetes, age-related macular degeneration (AMD), Parkinson disease, and systemic lupus erethematosis (SLE). A study in northern Finland is collecting data on a wide range of phenotypes (e.g., birth defects, neurological outcomes, mental illness/personality traits, asthma, cardiovascular events, infections, and diabetes), enabling investigators to study a range of phenotypes simultaneously in the same genomewide scan. Most of the studies were not isolated efforts but were integrated into research programs that usually included candidate-gene association studies and sometimes included linkage studies using affected-pair designs. All of the studies either are using or plan to use some version of a multistage design, and none employ pooling of DNA. They differ in the number of stages, ranging from two to four; in criteria for case selection, with some electing to enhance the initial series with "genetically enriched" cases and some not to; and in the nature of the control group, with some using population-based and some using family-based controls. The number of SNPs currently typed at each stage also differed between studies; although all groups expressed interest in eventually using the 500K panel from Affymetrix (and some are using early-access versions now), they wondered about the actual coverage of that panel. The morning session closed with a description of the research opportunities that derive from studying admixed populations, focusing on Hispanic/Latino populations, and with a discussion of the advantages of studying nonhuman model organisms and of how such studies would complement studies on human populations. More-detailed descriptions of each presentation are provided in appendix A (online only).

The first afternoon began with a series of presentations about genotyping technologies and bioinformatics support. Appendix A (online only) provides some details about the currently available early-release 100–500K SNP platforms from Illumina, Affymetrix, and other companies, as well as the Genetrix bioinformatics suite.

Epidemiologic Study Design

Although the case-control design has become the workhorse of genetic association studies, there has been considerable discussion about its merits relative to cohort studies, nested case-control or case-cohort designs, and family-based designs (Langholz et al. 1999; Witte et al. 1999; Cardon and Bell 2001; Clayton and McKeigue 2001; Thomas and Witte 2002; Cardon and Palmer 2003). Lyle Palmer noted that one of the advantages of the cohort design is that it allows for many disease endpoints to be considered simultaneously using a common set of controls that he dubbed "universal controls" (Palmer and Cardon, in press).

Sobell et al. (1993) first suggested a two-stage design for association studies, which has recently been extended to genomewide scale by Satagopan and colleagues (2002, 2003, 2004), Lowe et al. (2004), and van den Oord and Sullivan (2003). An initial sample of subjects is tested for a dense set of markers, and then an independent sample is tested only on a subset of the most "significant" markers. They describe methods for optimizing the numbers of subjects and significance levels at each stage to maximize power, subject to a constraint on cost and the overall type I error rate. Jaya Satagopan demonstrated that the two-stage design could yield substantial cost savings over a one-stage design with the same test size and power. She also discussed likelihood inference for a quantitative trait locus (QTL), optimizing by selective sampling of subjects with extreme trait values. A related problem for testing single candidate genes was recently considered by Thomas et al. (2004), in which a relatively small sample was used to select tag SNPs, which were then tested in the full study. Both designs use all the data from both samples in the final analysis. Duncan Thomas and Daniel Stram described some extensions of the Satagopan et al. approach for the design of the CFR and MEC genomewide studies. For these studies, it appears that the optimal design typically entails allocating 80%–90% of the costs to the first stage, with a significance level of ~0.001–0.005 at the first stage. This can be expected to yield ~500–2,500 loci to be tested at the second stage at a significance level of ~0.00001. A sample size of ~2,000 subjects at stage 1 and ~2,000 at stage 2 would be expected to yield 80% power for detecting a locus with allele frequency 10% conferring a relative risk of 1.3, at an overall ("experimentwise") type I error rate of 5% and a total cost of ~$5 million. Extensions required for these calculations involve different costs at the two stages, allowance for the r^2 for prediction of unobserved causal SNPs, and use of additional markers at the second stage (see illustrative calculations in appendix A [online only]).

Other design considerations include whether to use a family-based or population-based design and whether to use stratified sampling to enrich one or more of the stages for genetically predisposed cases. The appropriate choice depends in part on whether one intends a search for common polymorphisms having main effects on disease risk or those having modifying effects on other genes or environmental factors, as well as prior beliefs about the "common disease common variant" (Cargill and Daley 2000; Reich and Lander 2001; Lohmueller et al. 2003) versus "multiple rare variant" (Pritchard

2001; Pritchard and Cox 2002; Wang et al. 2003; Fearnhead et al. 2004) hypotheses. Although restriction of the case series to those with a positive family history can be effective at enriching for genetic susceptibility, it risks introducing cryptic relatedness, since cases may share greater kinship with one another than controls, particularly in small regions or population isolates. Selecting cases on the basis of "severity" could also have the counterproductive effect of enriching environmental as well as genetic factors. Genomewide scans could also be used to identify genes that interact with particular environmental agents or other modifying factors already known to play a major role in the etiology of a particular disease. Stratifying on microsatellite instability (MSI) in colorectal tumors or restricting to MSI-stable cases—or stratifying on family history, age, or any of a number of other factors—might lead to greater etiologic homogeneity and improve power for detecting single-gene effects.

The problem of population stratification has been widely debated (Wacholder et al. 2000; Thomas and Witte 2002; Wacholder et al. 2002; Cardon and Palmer 2003; Freedman et al. 2004). David Clayton noted that this can lead to three distinct problems: confounding; cryptic relatedness, resulting in overdispersion of the test statistic; and selection bias. Unlike some other biases, these problems do not become smaller with increasing sample size—on the contrary, the potential inflation of type I error rates will be much larger in studies of the size needed to demonstrate significance at the genomewide scale. Family-based case-control designs offer protection from population stratification, but at the expense of some loss of power from "overmatching" on genotype. The availability of an enormous number of unlinked markers might provide ample opportunity to control for population stratification by the methods of genomic control (Devlin et al. 2001), structured association (Pritchard et al. 2000), or simple logistic regression (Tang et al. 2004) in studies using unrelated controls. Although population stratification will generally cause overdispersion of test statistics (thereby inflating significance levels overall), the significance of any specific test could be either increased or decreased. Thus, the genomic control method, while yielding a test procedure with the correct type I error rate, may suffer from some power loss. It is not known whether the structured association method, which aims to correct each association by stratifying on individual ancestry estimates, would suffer from power loss to the same extent, but it would require a much more computationally intensive analysis.

Population Selection

White populations have hitherto been the primary focus of most association studies, and one of the populations being intensively studied in the HapMap and other genomic variation projects. An open question remains about the "transferability" to other populations of a panel of tag SNPs that has been optimized for whites (Carlson et al. 2003; Mueller et al. 2005). Duncan Thomas described preliminary simulation studies suggesting that a testing procedure that combines a test of overall race-adjusted association and a test of race heterogeneity (each tested at significance level $\alpha/2$) could yield higher power than either test alone (at level α), even under the hypothesis that the relative risk for an unobserved causal variant was the same across populations. The true relative risk for a causal variant might also be expected to vary across populations because of interactions with other genes or environmental factors with differing prevalence.

Itsik Pe'er discussed the utility of isolated populations for association studies because of their reduced genetic diversity, longer LD, and extreme phenotype frequencies for particular conditions. As an example, he described efforts to construct a haplotype map for the Kosrae population of Micronesia, an isolate for ~2,000 years with European admixture beginning in the 19th century (Wijsman et al. 2003). These data are based on typing 100,000 SNPs with the 100K GeneChip in 30 parent-child trios. Although the general patterns of allele frequencies and decay of LD were similar to the non-African HapMap data, there was a striking excess of single-copy alleles, with 14 individuals carrying 80% of these rare alleles and a tendency for them to cluster along the genome, suggesting the effects of recent European admixture. This implies a modest improvement in power for single-marker associations and less diversity in long-range haplotypes.

Marker Selection

Eric Jorgenson and John Witte elaborated on the relative merits of "map-based" (i.e., uniformly spaced and or tagSNPs) versus function or "gene-based" (i.e., occurring only in coding, splice site, regulatory regions, and or highly conserved intronic regions) approaches to whole genome association studies (Collins et al. 1997; Tabor et al. 2002; Botstein and Risch 2003; Carlson et al. 2004; Neale and Sham 2004; Palmer and Cardon, in press). They described sample size and cost calculations, concluding that the gene-based approach would be considerably less expensive, because of the reduction in the number of SNPs that must be genotyped and the resulting smaller sample sizes required, but would undoubtedly miss some potentially relevant regions (e.g., enhancers). A map-based approach would also be likely to miss effects of some variation in genes, unless the panel included adequate density of markers within genes (see appendix A [online only] for discussion).

Statistical Analyses

Two main approaches have been advocated for testing gene associations: a "direct" method, based on a simple χ^2 test for association, and an "indirect" method, based on associations with haplotypes inferred from unphased multilocus genotypes (Schaid et al. 2002; Zaykin et al. 2002; Stram et al. 2003), the haplotypes being assumed to carry information about possibly unobserved causal variants in the region. In a genomewide context, either approach involves testing an enormous number of hypotheses simultaneously, thereby raising the problem of multiple comparisons. Bonferroni correction is one commonly used approach to address this problem, requiring an extremely small P value (say, $0.05/500{,}000 = 10^{-7}$) to claim genomewide significance for any particular SNP—or an even smaller P value if multiple subgroups, additional markers, or multiple methods of analysis (e.g., SNPs and haplotypes) are considered. Others have suggested a Bayesian approach, such as the False Positive Report Probability (Wacholder et al. 2004), requiring explicit consideration of the prior probability for each hypothesis under consideration. Under the assumption that there could be many true positive associations, however, the Bonferroni correction is too conservative, and a variety of methods based on the False Discovery Rate have been advocated (Benjamini and Hochberg 1995; Efron and Tibshirani 2002; Sabatti et al. 2003; Storey and Tibshirani 2003).

Two papers have recently proposed analytic approaches for genomewide association studies that go well beyond simple exhaustive testing of all SNPs separately. Lin et al. (2004) proposed exhaustive testing of haplotype associations over all possible windows of segments, using a computationally efficient permutation procedure to assess the significance of these correlated tests. Marchini et al. (2005) proposed exhaustive testing of all possible pairwise gene-gene interactions. Nelson Freimer introduced the idea of haplotype sharing among case-case pairs as an alternative to case-control association; Duncan Thomas provided a formal test of haplotype sharing and showed how this test could be decomposed into principal components representing case-control associations with clusters of similar haplotypes, in the spirit of Tzeng et al. (2003), Nyholt (2004), and Lin and Altman (2004).

It is generally agreed that no amount of association testing in epidemiological studies alone can distinguish between the true-positive and false-positive signals obtained in a multistage genomewide scan (Page et al. 2003). Approaches that could be taken at this stage might include comparative genomics (Sidow 2002; Bejerano et al. 2004), linkage analysis of expression data (Morley et al. 2004), or computational approaches to predicting function (Ng and Henikoff 2003; Taylor and Greene 2003; Livingston et al. 2004; Xi et al. 2004; Zhu et al. 2004), before launching into the labor-intensive and time-consuming process of developing functional tests. Eleazar Eskin illustrated this by incorporating predictions of variation function, using as an example the Chromogranin A gene (*CHGA*) involved in hypertension. By use of the HAP phasing algorithm (Hinds et al. 2005), six common haplotypes were identified, one of which appeared to be strongly associated with the trait. A combination of comparative genomic analysis and known binding-site analysis identified a specific variant that could be responsible, G462A, shown by in vitro assay to alter reporter expression.

Chiara Sabatti discussed the interpretation of stretches of homozygosity, using data on Costa Rican case-parent trios with Bipolar-1 disorder, typed at ~3,000 SNPs on chromosome 22. She described a hidden Markov model approach to estimation of the inbreeding coefficient from genomic data (Leutenegger et al. 2003), which showed that all but three of parents had estimated inbreeding coefficients of zero. Another possible explanation for long homozygous stretches is large-scale copy number variation (Iafrate et al. 2004; Sebat et al. 2004). She discussed the applicability of methods used to deal with genomic losses in cancer (Newton et al. 1998) to other types of phenotypes, but she concluded that, without a good model for instability, such techniques were more useful for evaluating the likelihood of seeing particular stretches of large-scale copy number variation than for detecting their existence.

Power

David Clayton showed the sample-size requirements for a single-stage association study using both the direct and indirect approaches (see appendix A [online only]). At a marker density on the order of 1 every 6 kb (500,000 markers), we expect that most associations would be detected indirectly by LD rather than with causal variants directly. Under the assumption that an average of 8 tag SNPs would yield an r^2 of 0.8 and with the use of an 8-df test, the sample sizes required for such indirect associations would be slightly less than double those that would be needed for a direct association.

Paul de Bakker summarized a comparison of various SNP selection and analysis methods, using simulations based on the HapMap-ENCODE regions, representative 500-kb regions of the genome with complete ascertainment of common variation in 270 individuals (ENCODE Project Consortium 2004). By nominating all common SNPs as a causal allele, one by one, they generated simulated case-control data sets, from which they computed the power to detect an association under different tagging and testing scenarios. They found that choosing tag SNPs from a 5-kb panel (such as the Phase

I HapMap) gave surprisingly good power for common (>5% frequency) causal alleles, and that specified haplotype-based tests further improved genotyping efficiency—a 33% reduction in the number of tag SNPs required, with no loss of power. Additional sliding windows of haplotypes did not help for common causal alleles, once the increase in the number of tests was allowed for, but there was some improvement in power when the causal allele was rare (minor allele frequency [MAF] < 5%).

Daniel Stram described similar simulations, focusing on the power of some very simple analyses of whole genome scans using tag SNPs. Power is determined by the noncentrality parameter, a function of the causal allele frequency, its true relative risk, and the r^2 for the prediction of the unobserved causal variant by nearby SNP(s). For relatively small regions, a Bonferroni-adjusted single-SNP analysis is generally more powerful than a multivariate test of association, but, on a genomewide scale, the effective number of "independent" tests is a function of the extent of LD. By determining the block structure, choosing tag SNPs within blocks, conducting multivariate tests within each block, and applying a Bonferroni correction for the number of blocks instead, he showed that this method yielded better power than simply using all SNPs with Bonferroni correction for the number of SNPs.

Genotyping Errors

David Clayton noted that genotyping errors are generally assumed to be nondifferential (i.e., not related to phenotype), leading to some loss of power and bias in relative risk estimates towards the null, but no increase in type I errors (except in case-parent trio designs). However, he pointed out that it may be difficult to ensure that all aspects of DNA processing and analysis are the same for cases and controls, particularly if the ascertainment of these two groups is not concurrent. To test this assumption, he showed data from a study of nonsynonymous SNPs (nsSNPs) and type I diabetes, which revealed that some of the overdispersion of association tests that were not obviously true positives could be explained by questionable allele calls or by those not replicated on another platform, as well as by regional stratification and substructure (see appendix A [online only] for additional details of this analysis). Particularly disturbing were shifts between cases and controls in the point clouds corresponding to each genotype, presumably due to differences in DNA processing. Standard laboratory practice of using blinded samples to determine the parameters for allele calling could thus lead to differential misclassification (with consequent inflation of type I error rates and relative risk estimates biased away from the null). Instead, it would appear that, to minimize such misclassification, it would be necessary to calibrate the software separately for each group.

Derek Gordon showed the effects of nondifferential genotyping error on both family-based and population-based tests of association. For case-parent trio data, even nondifferential errors can inflate type I error rates (Mitchell et al. 2003). To overcome this problem, Gordon et al. (2001) introduced the likelihood-based TDTae ("adjusted for errors"). Similar issues have been addressed for case-control designs using unrelated individuals (Rice and Holmans 2003). Phenotyping errors would be expected to have similar effects. Recent work on the use of double sampling, combining fallible methods on a large sample with a "gold standard" method on a subset, appears to improve power for tests of association (Gordon et al. 2004). Methods that formally incorporate information on accuracy of genotype or haplotype calls into the statistical test of association (Hao and Wang 2004) have some potential for extension to the whole-genome scale.

DNA pooling has been suggested as an efficient procedure for screening many samples for differences in allele frequencies at many loci (Bansal et al. 2002; Sham et al. 2002). Various authors have discussed experimental design for such studies (Barratt et al. 2002; Pfeiffer et al. 2002; Sham et al. 2002), including the number of pools and pool sizes needed for accurate allele frequency determination. The general sense of the participants was that this approach is not sufficiently reliable for use on a genomewide scale at this time, despite its obvious cost appeal.

Conclusions

All genomewide association studies of human populations that have been described above are using or plan to use a multistage design, and none are proposing to use DNA pooling. Studies differ in the number of stages and in the nature of cases and controls selected for each stage. Some of the studies in the United Kingdom and the CFRs employ a strategy that aims to enhance genetically caused cases, while others (e.g., the AMD study) choose not to employ such a strategy. There was general agreement that it is probably helpful to decrease heterogeneity in the case series, either by exclusion of selected subgroups (e.g., exclusion of colorectal cancer cases on the basis of MSI-H tumors or evidence of a germline mutation in a mismatch repair gene) or by a stratified selection that would ensure sufficient sample size in the major strata of interest. In the study of cancer, there is growing recognition of the value of using molecular markers derived from the tumor to define sources of heterogeneity. Similar markers are under development with other diseases. Thus, even for diseases like cancer that are traditionally analyzed as simple dichotomous

phenotypes, there are often several dimensions on which to characterize cases; for other diseases, like diabetes, the number of variables needed to fully characterize the phenotype can be very large. A genomewide scan for genomic determinants of gene expression levels, for example, would entail potentially many billion comparisons; several such scans are currently underway.

Advantages and disadvantages of alternative control series, usually discussed in the context of candidate gene studies (e.g., trade-offs between power and control of population stratification), are also relevant to genomewide association studies. Some studies—for example, the CFRs—are in a position to use both types of controls and currently plan to use unrelated controls in stage 1 to enhance power and family-based controls in stage 2. If the main objective of a second stage is to replicate the findings of a first stage, then one needs to be mindful about introducing possible sources of heterogeneity (e.g., by using different types of cases or controls) between stages, which will complicate interpretations of results.

Most studies are not or do not plan to incorporate information on environmental exposures in the early stages of the genomewide studies. This concerned some investigators, since genomewide scans could miss important genetic causes where the effect of the gene is only detectable when information on the relevant environmental exposure(s) is incorporated into the analyses, particularly since common genetic variants for common diseases may plausibly interact with environmental exposures.

Several very general issues were raised in the concluding discussion. Alice Whittemore began by asking (1) Is the technology driving the science? (2) Can we afford the technology? (3) When is an association scan unwarranted? and (4) When it is warranted, how can the epidemiology and biology of the disease drive our choice of design? An example where such an approach might not be warranted is Hodgkin disease, in which the risk to DZ twins is very low and the risk to MZ twins nearly 100% (Mack et al. 1995), suggesting multiple rare variants (Risch 1990), a scenario not amenable to association mapping. Robert Haile asked whether the time was ripe in terms of the technology development, the need for coordination by the many investigators who are likely to be proposing such studies in the near future for a range of diseases and even for different species, and how best to deal with the problems of etiologic heterogeneity and complexity. Nevertheless, there seemed to be a broad consensus that the time was indeed ripe for launching the first generation of genomewide association studies, but that each would require careful justification and coordination among groups studying similar conditions to ensure optimal allocation of the limited resources available for such expensive undertakings.

Acknowledgments

This workshop was supported by the University of Southern California (USC) Center of Excellence in Genomic Sciences (grant 1P50 HG002790), the Southern California Environmental Health Sciences Center (grant 5P30 ES07048), and the USC Keck School of Medicine. Invited speakers included Habib Ahsan (Columbia University), Fernando Arena (National Cancer Institute, National Institutes of Health), Paul de Bakker (Massachusetts General Hospital), Timothy Bishop (Leeds University), Jonathan Buckley (University of Southern California), Graham Casey (Cleveland Clinic Foundation), David Clayton (Cambridge Institute of Medical Research), Mariza de Andrade (Mayo Clinic), David Duggan (TGen), Eleazar Eskin (University of California at San Diego), Nelson Freimer (University of California at Los Angeles), Ellen Goode (Mayo Clinic), Derek Gordon (Rockefeller University), Robert Haile (University of Southern California), Brian Henderson (University of Southern California), John Hopper (University of Melbourne), Eric Jorgenson (University of California at San Francisco), Magnus Nordborg (University of Southern California), Lyle Palmer (University of Western Australia), Itsik Pe'er (Broad Institute), Chiara Sabatti (University of California at Los Angeles), Jaya Satagopan (Memorial Sloan Kettering Cancer Center), Nik Schork (University of California at San Diego), Daniela Seminara (National Cancer Institute, National Institutes of Health), Susan Service (University of California at Los Angeles), Dan Stram (University of Southern California), Simon Tavaré (University of Southern California), Nicole Tedeschi (University of Southern California), David Van Den Berg (University of Southern California), Alice Whittemore (Stanford University), and John Witte (University of California at San Francisco).

Web Resources

The URLs for data presented herein are as follows:

Affymetrix, http://www.affymetrix.com/index.affx
Epicenter Software, http://www.epicentersoftware.com/genetrix.php (for Genetrix)
Illumina, http://www.illumina.com
ParAllele BioScience, http://www.parallelebio.com (for 100K panel)
Sequenom, http://www.sequenom.com (for 75K and 100K panels)

References

Bansal A, van den Boom D, Kammerer S, Honisch C, Adam G, Cantor CR, Kleyn P, Braun A (2002) Association testing by DNA pooling: an effective initial screen. Proc Natl Acad Sci USA 99:16871–16874

Barratt BJ, Payne F, Rance HE, Nutland S, Todd JA, Clayton DG (2002) Identification of the sources of error in allele

frequency estimations from pooled DNA indicates an optimal experimental design. Ann Hum Genet 66:393–405
Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, Haussler D (2004) Ultraconserved elements in the human genome. Science 304:1321–1325
Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc B 57:289–300
Boehnke M (1994) Limits of resolution of genetic linkage studies: implications for the positional cloning of human disease genes. Am J Hum Genet 55:379–390
Botstein D, Risch N (2003) Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. Nat Genet 33:228–237
Cardon LR, Bell JI (2001) Association study designs for complex diseases. Nat Rev Genet 2:91–99
Cardon LR, Palmer LJ (2003) Population stratification and spurious allelic association. Lancet 361:598–604
Cargill M, Daley GQ (2000) Mining for SNPs: putting the common variants—common disease hypothesis to the test. Pharmacogenomics 1:27–37
Carlson CS, Eberle MA, Kruglyak L, Nickerson DA (2004) Mapping complex disease loci in whole-genome association studies. Nature 429:446–452
Carlson CS, Eberle MA, Rieder MJ, Smith JD, Kruglyak L, Nickerson DA (2003) Additional SNPs and linkage-disequilibrium analyses are necessary for whole-genome association studies in humans. Nat Genet 33:518–521
Clayton DG, McKeigue PM (2001) Epidemiological methods for studying genes and environmental factors in complex diseases. Lancet 358:1357–1360
Collins FS, Guyer MS, Charkravarti A (1997) Variations on a theme: cataloging human DNA sequence variation. Science 278:1580–1581
Devlin B, Roeder K, Wasserman L (2001) Genomic control, a new approach to genetic-based association studies. Theor Pop Biol 60:155–160
Efron B, Tibshirani R (2002) Empirical Bayes methods and false discovery rates for microarrays. Genet Epidemiol 23:70–86
ENCODE Project Consortium (2004) The ENCODE (ENCyclopedia Of DNA Elements) Project. Science 306:636–640
Fearnhead NS, Wilding JL, Winney B, Tonks S, Bartlett S, Bicknell DC, Tomlinson IP, Mortensen NJ, Bodmer WF (2004) Multiple rare variants in different genes account for multifactorial inherited susceptibility to colorectal adenomas. Proc Natl Acad Sci USA 101:15992–15997
Freedman ML, Reich D, Penney KL, McDonald GJ, Mignault AA, Patterson N, Gabriel SB, Topol EJ, Smoller JW, Pato CN, Pato MT, Petryshen TL, Kolonel LN, Lander ES, Sklar P, Henderson B, Hirschhorn JN, Altshuler D (2004) Assessing the impact of population stratification on genetic association studies. Nat Genet 36:388–393
Gibbs RA, Belmont JW, Hardenbol P, Willis TD, Yu F, Yang H, Ch'ang LY, et al (2003) The International HapMap Project. Nature 426:789–796
Gordon D, Heath SC, Liu X, Ott J (2001) A transmission/disequilibrium test that allows for genotyping errors in the analysis of single-nucleotide polymorphism data. Am J Hum Genet 69:371–380
Gordon D, Yang Y, Haynes C, Finch SJ, Mendell NR, Brown AM, Haroutunian V (2004) Increasing power for tests of genetic association in the presence of phenotype and/or genotype error by use of double-sampling. Stat Appl Genet Mol Biol 3:1–35
Hao K, Wang X (2004) Incorporating individual error rate into association test of unmatched case-control design. Hum Hered 58:154–163
Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E, Ballinger DG, Frazer KA, Cox DR (2005) Whole-genome patterns of common DNA variation in three human populations. Science 307:1072–1079
Hirschhorn JN, Daly MJ (2005) Genome-wide association studies for common disease and complex traits. Nat Rev Genet 6:95–108
Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW, Lee C (2004) Detection of large-scale variation in the human genome. Nat Genet 36:949–951
Klein RJ, Zeiss C, Chew EY, Tsai J-Y, Sackler RS, Haynes C, Henning AK, SanGiovanni JP, Mane SM, Mayne ST, Bracken MB, Ferris FL, Ott J, Barnstable C, Hoh J (2005) Complement factor H polymorphism in age-related macular degeneration. Science 308:385–389
Langholz B, Rothman N, Wacholder S, Thomas D (1999) Cohort studies for characterizing measured genes. Monogr Natl Cancer Inst 26:39–42
Leutenegger AL, Prum B, Genin E, Verny C, Lemainque A, Clerget-Darpoux F, Thompson EA (2003) Estimation of the inbreeding coefficient through use of genomic data. Am J Hum Genet 73:516–523
Lin Z, Altman RB (2004) Finding haplotype tagging SNPs by use of principal components analysis. Am J Hum Genet 75:850–861
Lin S, Chakravarti A, Cutler DJ (2004) Exhaustive allelic transmission disequilibrium tests as a new approach to genome-wide association studies. Nat Genet 36:1181–1188
Livingston RJ, von Niederhausern A, Jegga AG, Crawford DC, Carlson CS, Rieder MJ, Gowrisankar S, Aronow BJ, Weiss RB, Nickerson DA (2004) Pattern of sequence variation across 213 environmental response genes. Genome Res 14:1821–1831
Lohmueller KE, Pearce CL, Pike MC, Lander ES, Hirschhorn JN (2003) Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. Nat Genet 33:177–182
Lowe CE, Cooper JD, Chapman JM, Barratt BJ, Twells RC, Green EA, Savage DA, Guja C, Ionescu-Tirgoviste C, Tuomilehto-Wolf E, Tuomilehto J, Todd JA, Clayton DG (2004) Cost-effective analysis of candidate genes using htSNPs: a staged approach. Genes Immun 5:301–305
Mack TM, Cozen W, Shibata DK, Weiss LM, Nathwani BN, Hernandez AM, Taylor CR, Hamilton AS, Deapen DM, Rappaport EB (1995) Concordance for Hodgkin's disease in identical twins suggesting genetic susceptibility to the young-adult form of the disease. N Engl J Med 332:413–418
Marchini J, Donnelly P, Cardon LR (2005) Genome-wide strat-

egies for detecting multiple loci that influence complex diseases. Nat Genet 37:413–417
Mitchell AA, Cutler DJ, Chakravarti A (2003) Undetected genotyping errors cause apparent overtransmission of common alleles in the transmission/disequilibrium test. Am J Hum Genet 72:598–610
Morley M, Molony CM, Weber TM, Devlin JL, Ewens KG, Spielman RS, Cheung VG (2004) Genetic analysis of genome-wide variation in human gene expression. Nature 430: 743–747
Mueller JC, Lohmussaar E, Magi R, Remm M, Bettecken T, Lichtner P, Biskup S, Illig T, Pfeufer A, Luedemann J, Schreiber S, Pramstaller P, Pichler I, Romeo G, Gaddi A, Testa A, Wichmann HE, Metspalu A, Meitinger T (2005) Linkage disequilibrium patterns and tagSNP transferability among European populations. Am J Hum Genet 76:387–398
Neale BM, Sham PC (2004) The future of association studies: gene-based analysis and replication. Am J Hum Genet 75: 353–362
Newton MA, Gould MN, Reznikoff CA, Haag JD (1998) On the statistical analysis of allelic-loss data. Stat Med 17:1425–1445
Ng PC, Henikoff S (2003) SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res 31:3812–3814
Nordborg M, Taveré S (2002) Linkage disequilibrium: what history has to tell us. Trends Genet 18:83–90
Nyholt DR (2004) A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. Am J Hum Genet 74:765–769
Ozaki K, Ohnishi Y, Iida A, Sekine A, Yamada R, Tsunoda T, Sato H, Hori M, Nakamura Y, Tanaka T (2002) Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction. Nat Genet 32:650–654
Page GP, George V, Go RC, Page PZ, Allison DB (2003) "Are we there yet?": deciding when one has demonstrated specific genetic causation in complex diseases and quantitative traits. Am J Hum Genet 73:711–719
Palmer LJ, Cardon LR. Shaking the tree: mapping complex disease genes using linkage disequilibrium. Lancet (in press)
Pfeiffer RM, Rutter JL, Gail MH, Struewing J, Gastwirth JL (2002) Efficiency of DNA pooling to estimate joint allele frequencies and measure linkage disequilibrium. Genet Epidemiol 22:94–102
Pritchard JK (2001) Are rare variants responsible for susceptibility to complex diseases? Am J Hum Genet 69:124–137
Pritchard JK, Cox NJ (2002) The allelic architecture of human disease genes: common disease-common variant…or not? Hum Mol Genet 11:2417–2423
Pritchard JK, Stephens M, Rosenberg NA, Donnelly P (2000) Association mapping in structured populations. Am J Hum Genet 67:170–181
Reich DE, Lander ES (2001) On the allelic spectrum of human disease. Trends Genet 17:502–510
Rice KM, Holmans P (2003) Allowing for genotyping error in analysis of unmatched case-control studies. Ann Hum Genet 67:165–174
Risch N (1990) Linkage strategies for genetically complex traits. I. Multilocus models. Am J Hum Genet 46:222–228
Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. Science 273:1616–1617
Sabatti C, Service S, Freimer N (2003) False discovery rate in linkage and association genome screens for complex disorders. Genetics 164:829–833
Satagopan JM, Elston RC (2003) Optimal two-stage genotyping in population-based association studies. Genet Epidemiol 25:149–157
Satagopan JM, Venkatraman ES, Begg CB (2004) Two-stage designs for gene-disease association studies with sample size constraints. Biometrics 60:589–597
Satagopan JM, Verbel DA, Venkatraman ES, Offit KE, Begg CB (2002) Two-stage designs for gene-disease association studies. Biometrics 58:163–170
Schaid DJ, Rowland CM, Tines DE, Jacobson RM, Poland GA (2002) Score tests for association between traits and haplotypes when linkage phase is ambiguous. Am J Hum Genet 70:425–434
Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, Maner S, Massa H, Walker M, Chi M, Navin N, Lucito R, Healy J, Hicks J, Ye K, Reiner A, Gilliam TC, Trask B, Patterson N, Zetterberg A, Wigler M (2004) Large-scale copy number polymorphism in the human genome. Science 305:525–528
Sham P, Bader JS, Craig I, O'Donovan M, Owen M (2002) DNA pooling: a tool for large-scale association studies. Nat Rev Genet 3:862–871
Sidow A (2002) Sequence first. Ask questions later. Cell 111: 13–16
Sobell JL, Heston LL, Sommer SS (1993) Novel association approach for determining the genetic predisposition to schizophrenia: case-control resource and testing of a candidate gene. Am J Med Genet 48:28–35
Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. Proc Natl Acad Sci USA 100:9440–9445
Stram DO, Pearce CL, Bretsky P, Freedman M, Hirschhorn JN, Altshuler D, Kolonel LN, Henderson BE, Thomas DC (2003) Modeling and E-M estimation of haplotype-specific relative risks from genotype data for a case-control study of unrelated individuals. Hum Hered 55:179–190
Tabor HK, Risch NJ, Myers RM (2002) Opinion: candidate-gene approaches for studying complex genetic traits: practical considerations. Nat Rev Genet 3:391–397
Tang H, Quertermous T, Rodriguez B, Kardia SL, Zhu X, Brown A, Pankow JS, Province MA, Hunt SC, Boerwinkle E, Schork NJ, Risch NJ (2004) Genetic structure, self-identified race/ethnicity, and confounding in case-control association studies. Am J Hum Genet 76:268–275
Taylor NE, Greene EA (2003) PARSESNP: A tool for the analysis of nucleotide polymorphisms. Nucleic Acids Res 31: 3808–11
Thomas DC, Witte JS (2002) Point: Population stratification: A problem for case-control studies of candidate gene associations? Cancer Epidemiol Biomarkers Prev 11:505–512
Thomas DC, Xie R, Gebregziabher M (2004) Two-stage sampling designs for gene association studies. Genet Epidemiol 27:401–414
Tzeng JY, Devlin B, Wasserman L, Roeder K (2003) On the identification of disease mutations by the analysis of hap-

lotype similarity and goodness of fit. Am J Hum Genet 72: 891–902

van den Oord EJ, Sullivan PF (2003) A framework for controlling false discovery rates and minimizing the amount of genotyping in the search for disease mutations. Hum Hered 56:188–199

Wacholder S, Chanock S, Garcia-Closas M, El Ghormli L, Rothman N (2004) Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. J Natl Cancer Inst 96:434–442

Wacholder S, Rothman N, Caporaso N (2000) Population stratification in epidemiologic studies of common genetic variants and cancer: quantification of bias. J Natl Cancer Inst 92:1151–8

Wacholder S, Rothman N, Caporaso N (2002) Counterpoint: Bias from population stratification is not a major threat to the validity of conclusions from epidemiologic studies of common polymorphisms and cancer. Cancer Epidemiol Biomarkers Prev 11:513–520

Wang WY, Cordell HJ, Todd JA (2003) Association mapping of complex diseases in linked regions: estimation of genetic effects and feasibility of testing rare variants. Genet Epidemiol 24:36–43

Wang WYS, Barratt BJ, Clayton DG, Todd JA (2005) Genome-wide association studies: theoretical and practical concerns. Nat Rev Genet 6:109–118

Wijsman EM, Rosenthal EA, Hall D, Blundell ML, Sobin C, Heath SC, Williams R, Brownstein MJ, Gogos JA, Karayiorgou M (2003) Genome-wide scan in a large complex pedigree with predominantly male schizophrenics from the island of Kosrae: evidence for linkage to chromosome 2q. Mol Psychiatry 8:695–705, 643

Witte JS, Gauderman WJ, Thomas DC (1999) Asymptotic bias and efficiency in case-control studies of candidate genes and gene-environment interactions: Basic family designs. Am J Epidemiol 148:693–705

Xi T, Jones IM, Mohrenweiser HW (2004) Many amino acid substitution variants identified in DNA repair genes during human population screenings are predicted to impact protein function. Genomics 83:970–979

Zaykin DV, Westfall PH, Young SS, Karnoub MA, Wagner MJ, Ehm MG (2002) Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. Hum Hered 53:79–91

Zhu Y, Spitz MR, Amos CI, Lin J, Schabath MB, Wu X (2004) An evolutionary perspective on single-nucleotide polymorphism screening in molecular cancer epidemiology. Cancer Res 64:2251–2257