Published in final edited form as:
Pharmacogenomics. 2009 February ; 10(2): 243–251. doi:10.2217/14622416.10.2.243.

Methods for analysis in pharmacogenomics: lessons from the Pharmacogenetics Research Network Analysis Group

Balaji S Srinivasan,
Stanford University, CA, USA, PharmGKB, USA

Jinbo Chen,
University of Pennsylvania, USA

Cheng Cheng,
St Jude Children's Research Hospital, TN, USA

David Conti,
University of Southern California, CA, USA

Shiwei Duan,
University of Chicago, IL, USA

Brooke L Fridley,
Mayo Clinic, MN, USA

Xiangjun Gu,
University of Texas, TX, USA

Jonathan L Haines,
Vanderbilt University Medical Center, TN, USA

Eric Jorgenson,
University of California, CA, USA

Aldi Kraja,
Washington University School of Medicine, MO, USA

Jessica Lasky–Su,
Brigham and Women's Hospital, MA, USA

Lang Li,
Indiana University, IN, USA

†Author for correspondence: Vanderbilt University Medical, Center, Nashville, TN, USA, Tel.: +1 615 343 5851; Fax: +1 615 343 8619; ritchie@chgr.mc.vanderbilt.edu.

Financial & competing interests disclosure
BSS was funded by an NSF VIGRE postdoctoral fellowship (NSF grant EMSW21–VIGRE 0502385). BLF was supported in part by National Institutes of Health (NIH) grants R01 GM28157, R01 GM35720, R01 HL71478, R01 NS32352 and U01 GM61388, The Pharmacogenetics Research Network; by a PhRMA Foundation Center of Excellence in Clinical Pharmacology Award; and by American Heart Association (AHA) Grants 56051Z and 0525757Z. EJ was supported by U01 GM061390. DW was supported by Pharmacogenomics and Risk of Cardiovascular Disease (PARC, NIH Grant Number HL69757). JLH and MDR were supported in part by HL65962, the Pharmacogenomics of Arrhythmia Therapy U01 site of the Pharmacogenetics Research Network. ASR was supported in part by NIH grants 5U01GM074492-04 and 5R01HL74735-01. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.
No writing assistance was utilized in the production of this manuscript.

Andrei Rodin,
University of Texas, TX, USA

Dai Wang,
Cedars–Sinai Medical Center, CA, USA

Mike Province, and
Washington University School of Medicine, MO, USA

Marylyn D Ritchie†
Vanderbilt University Medical Center, Nashville, TN, USA, Tel.: +1 615 343 5851; Fax: +1 615 343 8619; ritchie@chgr.mc.vanderbilt.edu

Abstract

Each year, the Pharmacogenetics Research Network (PGRN) holds an analysis workshop for the members of the PGRN to share new methodologies, study design approaches and to discuss real data applications. This event is closed to members of the PGRN, but the methods presented are relevant to others conducting pharmacogenomics research. This special report describes many of the novel approaches discussed at the workshop and provides a resource for investigators in the field performing pharmacogenomics data analysis. While the focus is pharmacogenomics, the methods discussed are far ranging and have relevance to all types of genetic association studies: identifying noncoding variants and tag-SNPs, haplotype analysis, multivariate techniques, quantitative trait analysis, gene–gene and gene–environment interactions, and genome-wide association studies. The goal is to introduce readers to the topics discussed at the workshop and provide a direction for future development of analysis tools and methods for analysis of pharmacogenomic data.

Keywords

gene–environment interactions; gene–gene interactions; genetic determinants; haplotype analysis; pharmacogenomics; QTL analysis; tag SNPs; whole-genome association

Pharmacogenomics is the study of the relationship between individual genetic variation and drug response. One of the major goals of the field is the use of an individual's genomic information in conjunction with other demographic and environmental covariates to personalize a previously generic treatment regimen. Realizing this ambition requires nothing less than the ability to derive a genotype-to-phenotype map for a trait of interest. In the specific case of pharmacogenomics this trait is often a drug dosage, efficacy, toxicity, or a variable indicating response/nonresponse or adverse-event/no adverse-event, and the genotype is frequently a vector of SNP measurements, but progress in the area is intimately tied to progress in the more general search for the genetic determinants of complex traits.

Pharmacogenomics, similar to other areas of human genetics, has adopted a new strategy for the identification of genetic variation associated with clinical end points: primarily that of genome-wide association studies (GWAS). Recent developments in the large-scale determination of human variation [1] at first promised to make this problem comparatively trivial: simply assay all genomic variants, individually correlate them with the phenotype of interest, and return the loci of maximal effect along with a phenotype prediction function. However, this GWAS approach has proven more difficult than initially envisioned [2]. Perhaps the most unambiguously successful GWAS result to date was the discovery of the T1277C polymorphism in the *CFH* gene in macular degeneration, found simultaneously through GWAS [3] and targeted positional candidate approaches [4,5] due to its atypically large effect size. For example, Haines *et al.* reported that the odds ratio of T1277C homozygotes was 5.57 (95% confidence interval: 2.52–12.27) for carriers of two C alleles with neovascular age-

related macular degeneration (AMD) [4]. A similarly large effect was found for an allele associated with exfoliation glaucoma [6].

In general, though, the successful large-scale GWAS for diseases like coronary heart disease [7–10], breast cancer [11–13], Type II diabetes [10,14–19], and obesity [15,20–22] have discovered SNPs that are reproducibly associated with the trait but have moderate odds ratios in the range of 1.1–2.0. A recent review of 43 such disease associated alleles found that 42 out of 43 had odds ratios below 2, with 35 of these 43 below 1.5 [2]. In general, these small effect sizes mean that variation at these loci is often not diagnostically useful [23] as it accounts for only a small fraction of the variance in outcome. Yet even a small but reproducible effect would be preferable to the outcome of many similar efforts for traits like IQ [24] and Parkinson's disease [25–30], which have been plagued by problems with discoverability and reproducibility despite good study designs with highly heritable phenotypes. GWAS in several pharmaco genetic phenotypes are currently underway, thus a thorough review of the success rates is not available. We speculate that the effect sizes will be comparable in size; however, the success rate may be even lower due to the fact that many pharmaco genetic studies have significantly smaller sample sizes than GWAS in common disease phenotypes.

Several reasons for the inconsistent replication of GWAS are apparent, both experimental and statistical. It is important to note that many of these issues are only apparent in hindsight due to the efforts of the pioneering studies in this area. First, most studies to date have been conducted with either Affymetrix (CA, USA) or Illumina (CA, USA) SNP chips. Because some of the original SNP chips were designed for haplotype mapping rather than direct genomic association [31], the bulk of these SNPs were in nongenic regions of unknown function. As such, these chips privileged exploration over explain-ability. These issues are compounded by the fact that high chip costs limit sample size and the fact that SNPs – being single base alterations – are generally likely to be of small effect, unlike larger DNA lesions like copy-number variations.

From a statistical perspective, this combination of small sample sizes, small effect sizes and 500,000 or more explanatory variables presents significant challenges. Indeed, there is as yet no unified paradigm for the analysis of GWAS data.

One of the most common approaches for the analysis of GWAS case–control data is to use simple statistical tests (e.g., χ^2, Armitage trend, logistic regression) to examine the association between a marker and the case–control status, which essentially tests the differences in marker allele or genotype frequencies between case and control groups. One major criticism of such an initial analytical approach is the large number of expected false-positive results. Using a nominal $p = 0.05$ on the 500,000 SNPs will result in 25,000 false-positive results (even $p = 0.001$ will result in 500 false-positives).

Much is written about the problem of how to correct for the vast number of single locus tests being performed, but consensus has not yet emerged [32,33]. A Bonferroni correction is clearly too conservative for several reasons, including the fact that it assumes the independence of each test even though many of the SNPs are in linkage disequilibrium and thus correlated with each other. Alternative methods, including controlling the false-discovery rate, have been proposed, but none have gained general acceptance and much research is still ongoing [34–37]. As shown by Zaykin *et al.* [38] using multiple-testing correction or false-discovery rate techniques will not affect the overall ranks of test statistics and true associations may not be in the top percentage of test statistics, a phenomenon that has been observed in several recent GWAS [10,16–18]. In general, only the strongest associations can be detected using these traditional approaches with many more genes still to be found [39]. Ultimately, data integration,

replication datasets, or new analytical approaches must be used to filter these results down to a manageable number of the most likely genes.

It is this last point – the development of new techniques in genetic epidemiology with specific focus upon pharmacogenomic applications – that is the focus of this report. We discuss the methods and applications presented at a recent meeting of the Pharmacogenomics Research Network (PGRN) Analysis Working Group. The areas covered in this two day workshop can be broken down into four large topics: best practices and software for GWAS data management and analysis, single locus approaches for association, interaction and pathway based approaches for association, and preliminary reports of two recent GWAS of aspirin and statin response.

Genome-wide association studies: data management & analysis

Presentations given by Jonathan Haines (Vanderbilt University, TN, USA) and Marylyn Ritchie (Vanderbilt University, TN, USA), discussed useful heuristics, new software, and analyses of GWAS data. Jonathan Haines discussed methods for quality control in GWAS. He began by noting that as of late 2006, the number of reviews and theoretical papers on GWAS greatly exceeded the number of published, completed GWAS. This observation is not an indictment *per se*; a similar phenomenon occurred with expression microarrays in the late 1990s before the field coalesced. He then discussed numerous approaches to check for genotyping errors, sample mix-ups, and cryptic stratification, along with a program (whole-genome association study pipeline, or WASP) being developed by his laboratory to automate the calculation of these measures as well as generate various diagnostic plots. Table 1 explains many of the important quality control issues to consider.

Marylyn Ritchie discussed the problems associated with GWAS data analysis in the context of her group's new software packages for GWAS analysis. She began by noting that with a GWAS involving 500,000 SNPs and a binary response, a naive calculation of 500,000 χ^2 analyses with a 0.05 type I error rate would result in 25,000 false-positive results. She entertained several possibilities for improving upon this naive approach; focusing first upon her laborotory's implementation of a sequential replication filter (SERF) based approach. The concept here is to directly address the problem of GWA – namely, a failure to replicate – by focusing upon the number of times a functional locus replicates across a simulated study. The idea behind SERF is to determine this replication probability as a function of three parameters (initial group size, p-value threshold, and replication p-value threshold) which are otherwise arbitrarily specified in a stage-wise design. Ritchie then placed SERF in the broader context of SNP filters, which permit selection of SNPs via both within-study statistics (e.g., replication probability or χ^2 association) and prior knowledge (e.g., pathway membership or expression levels). Her group has implemented many such filters in Platform for the Analysis, Translation, and Organization of large scale data (PLATO), a software package for GWAS that is being prepared for release.

Single locus approaches

Three presentations given by Xiangjun Gu (University of Texas, TX, USA), Brooke Fridley (Mayo Clinic, MN, USA), and Eric Jorgenson (UCSF, CA, USA) focused upon methods for analyzing the functional effects of single loci; discussing SNP, haplotype, and even intron variation. Also, Jessica Lasky–Su (Channing Laboratory, MA, USA) presented an approach for screening and replication using the same dataset with an emphasis on single locus statistics. Xiangjun Gu began by discussing results of a simulation study, in which embedding just three causal SNPs in a 115K SNP dataset consisting of 400 individuals raised the number of SNPs with p-values less than 0.05 from approximately 5700 to more than 20,000. The three uncorrelated causal SNPs contribute an average of 10.2, 5.2 and 5.6% total trait variation

separately in 100 replicates. These effects are relatively strong in GWAS with a sample size of thousands of individuals, but they are not all that strong in this simulation study with only 400 individuals because the power when using a Bonferroni approach is 92, 21 and 21%, respectively. Though the specific numbers are highly sensitive to the specification of the genotype-to-phenotype mapping, the general point was that correlations between causal SNPs and other variants can increase the number of false positives in a study. This phenomenon becomes more obvious when causal SNPs have stronger effects and they are correlated with many other SNPs. To deal with this, he proposed a greedy stepwise forward multiple regression model. In each step, the algorithm chooses the SNP that explains the most variance in the trait and discards SNPs which are strongly uncorrelated with the trait (e.g., p-values < 0.05). He then computes the residual variance given this explanatory SNP and repeats the process until no further explanatory SNPs are detected. Results were shown for simulated data with synthetic models, and it will be interesting to see the results of this approach in real datasets.

Brooke Fridley presented a study in which a vector of repeated patient measurements was regressed upon haplotype variation in a candidate gene. Specifically, her group measured blood levels of epinephrine and norepinephrine at eight time points in 75 patients before, during and after a workout. In addition, these patients were genotyped at 12 SNPs spanning a particular locus with four common haplotypes. She specified a repeated measures haplotype model in which haplotype k had an effect upon the metabolite level of patient i at time j, as well as two more traditional models in which means and slopes of the metabolite time series were regressed upon haplotype variation. No highly significant results were found, but the general concept of compiling a rich vector of patient measurements is certainly advisable.

Eric Jorgenson's presentation dealt with the hunt for pharmacogenomically-relevant variation in introns within membrane-transporter genes. Past work on exonic variation had shown that nonsynonymous sites had lower variation and that variants with decreased function had lower allelic frequencies. His group's work extended this analysis to consider intronic variation within 50 bp of the intron–exon boundary; such sequence is known to encompass functionally relevant positions (e.g., splice sites) and is thus a natural candidate for in-depth analysis. Firstly, he used a Hidden Markov Model-based approach to define splice sites and branch points within the intronic sequence, and showed via a receiver operating characteristic (ROC) plot that predicted branch points matched prior knowledge. Then he calculated population genetic statistics for each position and noted that these varied between splice sites and branch points across two different datasets. Due to Encyclopedia of DNA Elements (ENCODE) [40] and related efforts, this kind of analysis is just taking off and the analysis of functional variation in intronic regions promises to be a very exciting area in GWAS for years to come.

Jessica Lasky–Su proposed a strategy developed for case–control studies that implements both screening and testing of SNP-trait associations using the same dataset. The screening step is constructed so that it is statistically independent of the association tests that are computed in the testing step. Therefore, the most promising SNPs identified by the screening step can be tested for association in the testing step without the need to adjust the significance level for the analysis conducted in the screening step. In simulation studies for 100K SNP scans, they observed significant differences in power between the proposed testing strategy and the standard Bonferroni correction. The practical relevance of the approach was illustrated by applications to a GWAS (100K), in which SNPs reaching genome-wide significance were identified that would not have been detected by standard adjustments for multiple testing. This methodology will be interesting to prospectively validate by conducting a GWAS with positive controls, to determine whether it is in fact possible to augment power by separating the screening and association steps.

Gene–gene & gene–environment interaction approaches

Four presentations given by Jinbo Chen (University of Pennsylvania, PA, USA), Aldi Kraja (Washington University, MO, USA), Shiwei Duan (University of Chicago, IL, USA), and Lang Li (Indiana University, IN, USA) discussed methods for multivariate analyses of gene–gene and gene–environment interactions in the context of candidate gene studies. Jinbo Chen presented a new class of semiparametric regression models for exploring gene–gene and gene–environment effects [41]. These partially linear tree-based regression models aim to combine the best aspects of linear models for dealing with additive main effects and tree-based models for investigating higher order gene–gene interactions. Chen applied the partially linear tree-based regression model to assess the association between biliary stone risk and 53 SNPs in the inflammation pathway in a population-based case–control study. The analysis yielded an interesting parsimonious summary of the joint effect of all SNPs. The method may be useful for candidate gene studies with many subjects and a limited number of explanatory variables.

Aldi Kraja introduced a new application of index selection, an established multivariate statistical technique developed in plant and animal breeding research designed to find phenotypes of sets of genes that operate together. He applied it to the problem of discovering the genetic basis of cytotoxicity response in cancer therapy and compared the results to those produced by traditional methods. In a preliminary analysis, he found 68 significant SNPs in genic regions and noted that the overall correlation between the index selection and the observed viability of Centre d' Etude du Polymorphisme Humain (CEPH) cell lines was 0.814. A follow-up analysis is planned to quantify the predicted individual response to chemotherapy.

Shiwei Duan integrated genotype, gene expression and daunorubicin sensitivity data on 176 HapMap cell lines to identify genetic variants that contributed to chemotherapeutic agent-induced cytotoxicity. Approximately, 200 total SNPs were found to be associated with daunorubicin-induced cytotoxicity in the HapMap populations, with about 30 of these SNPs identifed as expression qualitative trait locus (eQTLs). Moreover, a large proportion (~35.7–53.3%) of the mRNA level of the transcripts regulated by the eQTLs were significantly associated with daunorubicin-induced cytotoxicity (uncorrected $p < 0.05$). These results are important and interesting as they show strong concordance between three different assay types. Moreover, they demonstrate that bringing more data of different kinds to bear is likely to yield higher dividends than tweaking the formulas for correlation.

Lang Li proposed a mixture model approach that concurrently detects main and interaction effects of genetic variables through a likelihood ratio test, and performs phenotype cluster analysis based on genetic variable combinations. Its performance was demonstrated with four examples: *ESR2* effects on hot flashes in a tamoxifen trial; *ABCB1/ABCG2* interaction effects on patient survival in a docetaxel trial; *CYP2D6* polymophism effects on tamoxifen metabolite; and *CYP2B6* polymorphism effects on protein expression. Their method was promising in that it can perform genotype clustering and hypothesis testing simultaneously when investigating genotype/phenotype associations in pharmacogenetic studies. Importantly, though model based, they show that the approach is robust with respect to distribution misspecifcation [42].

Function & pathway based approaches

As noted above, multiple testing problems and small sample sizes are already major issues in GWAS when considering univariate associations between variants and traits. These problems are vastly exacerbated in multivariate analysis. Consider for example a case–control study with 500 subjects of each class and 500,000 SNPs. Naive consideration of all pairs of SNPs would result in 125 billion hypotheses to test. Moreover, the sparsity of data would increase, as the contingency tables would move from six cells (two trait values × three genotypes) with on

average 1000/6 counts per cell to a far sparser table with 18 cells (two traits × three genotypes × three genotypes) and 1000/18 counts per cell. Therefore, it is clear that this kind of exhaustive approach will not work, and that some kind of prior knowledge must be brought to bear. In particular, it is now increasingly apparent that genetic variation must be situated in a pathway context [43,44] to be properly understood. Two talks at the workshop were particularly focused in this regard, by Andrei Rodin (University of Texas, TX, USA) and David Conti (USC, CA, USA). In addition, Cheng Cheng (St. Jude Children's Research Hospital, TN, USA) also reviewed the literature on combining SNP and microarray data. Andrei Rodin presented work on reverse engineering pathways from pharmacogenomic association datasets. He first applied a Bayesian belief network (BN) approach to predict plasma lipid levels from *APOE* variation in three populations of 702, 854 and 286 patients, respectively, where variation was assessed in approximately 20 *APOE* SNPs in each population. He continued by applying the BN technique to a genome-wide association dataset containing 104K genome-spanning SNPs from the Genetic Epidemiology of Responses to Antihypertensives (GERA) study of blood pressure response to a thiazide diuretic. The presentation concluded with the discussion of general and technical aspects of his groups' BN software implementation. These included:

- Alternative discretization methods
- Hybrid probability models, incorporating both discrete and continuous variables
- Increasing scalability via pairwise SNP pretesting
- Balancing overfitting and underfitting
- Incorporating prior (expert) knowledge

The latter is especially important, as it allows one to incorporate already known pathway information into the Pharmacogenomic Evaluation of Antihypertensive Responses (PEAR) and GERA analyses, thereby controlling for overfitting.

David Conti gave a two-part talk on computational identification of tag SNPs and subsequent application to the analysis of genetic variation upon nicotine addiction. In the first part of his talk, he discussed a new program (SNAGGER) for computationally efficient selection of tag-SNPs. SNAGGER is advantageous in that it requires only pairwise linkage disequilibrium measurements rather than full haplotype inference and allows the user to incorporate SNPs with *a priori* importance. In the second half of his presentation, he discussed the analysis of an association study of nicotine addiction in which prior information was used to determine which genetic predictors would be retained, via Bayes model averaging using stochastic variable selection [45,46]. His approach was similar to that used in functional genomics [47, 48] for encoding information on gene function (e.g., from gene ontology [GO] or phenotypic quality ontology [PATO]) in such a way as to limit the number of terms used in each model fit.

GWAS design

In addition to the PGRN reports on methodology, two talks were given with preliminary results of recent GWAS, by Haiqing Shen (University of Maryland, MD, USA) and Dai Wang (Cedars–Sinai Medical Center, CA, USA). First, Haiqing Shen presented preliminary results from a GWAS of aspirin response in the Amish Heredity and Phenotype Intervention (HAPI) Heart Study. It has been known for sometime that aspirin's anti-aggregatory effect on platelet function may benefit patients with cardiovascular disease by inoculating against thromboembolic event, but wide individual variations in the response to aspirin treatment have complicated the therapeutic use of this drug. In the HAPI study, 886 Amish subjects were characterized with respect to cardio vascular disease and atherosclerosis and subject to four different short-term interventions targeted at different cardiovascular outcomes. One of these

interventions was aspirin therapy with monitoring of subsequent change in platelet function. Shen *et al.* found aspirin response to be significantly heritable, reporting a heritability of 25%. Preliminary results of a GWAS indicated a few putatively associated SNPs, though final confirmation must wait until the completion of the study.

Dai Wang reported preliminary results of the first stage of the Pharmacogenomics and Risk of Cardiovascular Disease (PARC) multistage study of the pharmacogenomics of statins. In the first stage of this design, 317K SNPs were assayed in 305 Caucasians treated with simvastatin and 675 Caucasians treated with pravastatin, with the goal of selecting 12,000 SNPs to do follow-up genotyping in 350 simvastatin and 650 pravastatin patients. They considered both the approach of using all subjects simultaneously to compute p-values as well as the possibility of independently computing p-values for each SNP in the simvastatin and pravastatin patients and then using Fisher's method to combine the p-values. As with the Shen presentation, the results were preliminary as genotyping had not yet been completed.

Expert commentary

Quality control procedures are essential prior to a thorough data analysis.

The future of GWAS in general and particularly in pharmacogenomics must focus more closely upon the systematic use of prior biological knowledge to boost power and bound possible epistatic effects. Otherwise, the combination of small effect sizes and massively multiple testing results in an intractable statistical problem.

One of the ways to incorporate prior knowledge is to use information from a pathway database in a Bayesian framework. This is preferable to the current practice in which pathways are primarily used in an a posteriori fashion to rationalize the top-ranked SNPs or genes as biologically significant.

It is likely that some combination of main effects, gene–gene and gene–environment interactions will be important for complex phenotypes, including drug response outcomes, and when robust methods are employed for multivariate or interaction analysis, interesting models can be identified.

It is important to gather as much data as possible before beginning the statistical analysis. This increment in data collection should not simply be limited to sample size or SNP count. Comprehensive measurement of as many relevant patient variables as possible is critical.

Moreover, if possible, positive controls should be included by measuring ubiquitous traits of known or partially known genetic etiology such as lactose tolerance or eye color. In general, standard practice should be to include as many such positive controls as possible, many of which can be cheaply assayed.

Future perspective

Genome-wide association where only SNPs are considered is just the beginning of a comprehensive analysis in pharmacogenomics. From a statistical perspective, we need to include many more biological covariates in our prediction functions. Specifically, we need to design studies that simultaneously measure many different kinds of biological data [48], which change at different time scales. SNP and copy-number variation chip measurements are currently popular because genome content is mostly constant across lifespan (modulo transpositions, insertions and deletions). However, as costs continue to plummet, it will become economically feasible to include multiple measurements of expression levels, metabolomic profiles, and possibly other variables like methylation states. In particular, the manifest

relevance of correlating metabolomic variation in particular with pharmacogenomic response variables should be apparent. Rather than methods that focus upon analytical manipulations within a given study, a statistically rigorous means of combining rich predictor data with known prior information is likely to be the key to deriving a robust genotype-to-phenotype mapping function for an arbitrary trait – and thus the key to pharmacogenomics.

Executive summary

- Pharmacogenomics as a field is increasingly dominated by genome-wide association studies, but technical challenges abound.
- Data management in genome-wide association itself is a nontrivial burden, and it is useful for practitioners to avail themselves of packages for quality control.
- Even given intact data, naive approaches to analyzing large scale genome-wide association data on a SNP-by-SNP basis encounter problems with multiple testing and high false-positive rates.
- Of these approaches, it is likely that a Bayesian approach that incorporates prior knowledge on genetic association is likely to be the most successful.

Acknowledgements

The authors thank the Pharmacogenetics Research Network (PGRN) Publications Committee for their thoughtful review of the manuscript.

Bibliography

Papers of special note have been highlighted as:

• of interest

•• of considerable interest

1. International HapMap Consortium: A haplotype map of the human genome. Nature 2005;437:1299–1320.1320 [PubMed: 16255080]• Excellent resource to learn more about the HapMap. The HapMap is an incredible resource for the genetics and pharmacogenetics community.
2. Iles MM, Fisher E. What can genome-wide association studies tell us about the genetics of common disease. PLoS Genet 2008;4:E33. [PubMed: 18454206]
3. Klein RJ, Zeiss C, Chew EY, et al. Complement factor H polymorphism in age-related macular degeneration. Science 2005;308:385–389. [PubMed: 15761122]
4. Haines JL, Hauser MA, Schmidt S, et al. Complement factor H variant increases the risk of age-related macular degeneration. Science 2005;308:419–421. [PubMed: 15761120]
5. Edwards AO, Ritter R, Abel KJ, Manning A, Panhuysen C, Farrer LA. Complement factor H polymorphism and age-related macular degeneration. Science 2005;308:421–424. [PubMed: 15761121]
6. Thorleifsson G, Magnusson K, Sulem P, et al. Common sequence variants in the *LOXL1* gene confer susceptibility to exfoliation glaucoma. Science 2007;317:1397–1400. [PubMed: 17690259]
7. Samani NJ, Erdmann J, Hall AS, et al. Genomewide association analysis of coronary artery disease. N Engl. J. Med 2007;357:443–453. [PubMed: 17634449]
8. McPherson R, Pertsemlidis A, Kavaslar N, et al. A common allele on chromosome 9 associated with coronary heart disease. Science 2007;316:1488–1491. [PubMed: 17478681]
9. Helgadottir A, Thorleifsson G, Manolescu A, et al. A common variant on chromosome 9p21 affects the risk of myocardial infarction. Science 2007;316:1491–1493. [PubMed: 17478679]

10. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 2007;447:661–678.678 [PubMed: 17554300]•• Shows one approach for genome-wide association study (GWAS) design and analysis that was successful for several common, complex disease phenotypes.
11. Easton DF, Pooley KA, Dunning AM, et al. Genome-wide association study identifies novel breast cancer susceptibility loci. Nature 2007;447:1087–1093. [PubMed: 17529967]
12. Stacey SN, Manolescu A, Sulem P, et al. Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer. Nat. Genet 2007;39:865–869. [PubMed: 17529974]
13. Hunter DJ, Kraft P, Jacobs KB, et al. A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. Nat. Genet 2007;39:870–874. [PubMed: 17529973]
14. Sladek R, Rocheleau G, Rung J, et al. A genome-wide association study identifies novel risk loci for Type 2 diabetes. Nature 2007;445:881–885. [PubMed: 17293876]
15. Frayling TM, Timpson NJ, Weedon MN, et al. A common variant in the *FTO* gene is associated with body mass index and predisposes to childhood and adult obesity. Science 2007;316:889–894. [PubMed: 17434869]
16. Zeggini E, Weedon MN, Lindgren CM, et al. Replication of genome-wide association signals in UK samples reveals risk loci for Type 2 diabetes. Science 2007;316:1336–1341. [PubMed: 17463249]
17. Scott LJ, Mohlke KL, Bonnycastle LL, et al. A genome-wide association study of Type 2 diabetes in Finns detects multiple susceptibility variants. Science 2007;316:1341–1345. [PubMed: 17463248]
18. Diabetes Genetics Initiative of Broad Institute of Harvard and MIT, Lund University, and Novartis Institutes of BioMedical Research. Genome-wide association analysis identifies loci for Type 2 diabetes and triglyceride levels. Science 2007;316:1331–1336. [PubMed: 17463246]
19. Zeggini E, Scott LJ, Saxena R, et al. Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for Type 2 diabetes. Nat. Genet 2008;40(5):638–645.645 [PubMed: 18372903]• GWAS can find even small to modest effects and this paper demonstrates how effects can be identified and/or missed depending on the study design.
20. Loos RJ, Bouchard C. *FTO*: the first gene contributing to common forms of human obesity. Obes. Rev 2008;9:246–250. [PubMed: 18373508]
21. Dina C, Meyre D, Gallina S, et al. Variation in *FTO* contributes to childhood obesity and severe adult obesity. Nat. Genet 2007;39:724–726. [PubMed: 17496892]
22. Freathy RM, Timpson NJ, Lawlor DA, et al. Common variation in the *FTO* gene alters diabetes-related metabolic traits to the extent expected, given its effect on BMI. Diabetes 2008;57(5):1419–1426. [PubMed: 18346983]
23. Pepe MS, Janes H, Longton G, Leisenring W, Newcomb P. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. Am. J. Epidemiol 2004;159(9):882–890. [PubMed: 15105181]
24. Butcher LM, Davis OS, Craig IW, Plomin R. Genome-wide quantitative trait locus association scan of general cognitive ability using pooled DNA and 500K single nucleotide polymorphism microarrays. Genes Brain Behav 2007;7(4):435–446. [PubMed: 18067574]
25. Clarimon J, Scholz S, Fung HC, et al. Conflicting results regarding the semaphorin gene (*SEMA5A*) and the risk for Parkinson disease. Am. J. Hum. Genet 2006;78:1082–1084. [PubMed: 16685660]
26. Elbaz A, Nelson LM, Payami H, et al. Lack of replication of thirteen single-nucleotide polymorphisms implicated in Parkinson's disease: a large-scale international study. Lancet Neurol 2006;5:917–923. [PubMed: 17052658]
27. Fung HC, Scholz S, Matarin M, et al. Genome-wide genotyping in Parkinson's disease and neurologically normal controls: first stage analysis and public release of data. Lancet Neurol 2006;5:911–916. [PubMed: 17052657]
28. Goris A, Williams-Gray CH, Foltynie T, Compston DA, Barker RA, Sawcer SJ. No evidence for association with Parkinson disease for 13 single-nucleotide polymorphisms identified by whole-genome association screening. Am. J. Hum. Genet 2006;78:1088–1084. [PubMed: 16685662]

29. Li Y, Rowland C, Schrodi S, et al. A case–control association study of the 12 single-nucleotide polymorphisms implicated in Parkinson disease by a recent genome scan. Am. J. Hum. Genet 2006;78:1090–1094. [PubMed: 16685663]
30. Myers RH. Considerations for genomewide association studies in Parkinson disease. Am. J. Hum. Genet 2006;78:1081–1082. [PubMed: 16685659]
31. Jorgenson E, Witte JS. A gene-centric approach to genome-wide association studies. Nat. Rev. Genet 2006;7:885–891.891 [PubMed: 17047687]•• Discusses how prior knowledge can be used for GWAS.
32. Hirschhorn JN, Daly MJ. Genome-wide association studies for common diseases and complex traits. Nat. Rev. Genet 2005;6:95–108.108 [PubMed: 15716906]• Good review for GWAS for researchers in genetics and pharmacogenetics.
33. Wang WY, Barratt BJ, Clayton DG, Todd JA. Genome-wide association studies: theoretical and practical concerns. Nat. Rev Genet 2005;6:109–118. [PubMed: 15716907]
34. Bahlo M, Stankovich J, Speed TP, Rubio JP, Burfoot RK, Foote SJ. Detecting genome wide haplotype sharing using SNP or microsatellite haplotype data. Hum. Genet 2006;119:38–50. [PubMed: 16362347]
35. Dalmasso C, Broët P, Moreau T. A simple procedure for estimating the false discovery rate. Bioinformatics 2005;21:660–668. [PubMed: 15479710]
36. Sun L, Craiu RV, Paterson AD, Bull SB. Stratified false discovery control for large-scale hypothesis testing with application to genome-wide association studies. Genet. Epidemiol 2006;30:519–530. [PubMed: 16800000]
37. Yang Q, Cui J, Chazaro I, Cupples LA, Demissie S. Power and type I error rate of false discovery rate approaches in genome-wide association studies. BMC Genet 2005;6:S134. [PubMed: 16451593]
38. Zaykin DV, Zhivotovsky LA. Ranks of genuine associations in whole-genome scans. Genetics 2005;171:813–823. [PubMed: 16020784]
39. Couzin J, Kaiser J. Genome-wide association. Closing the net on common disease genes. Science 2007;316:820–822. [PubMed: 17495150]
40. Birney E, Stamatoyannopoulos J, Dutta A, et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 2007;447:799–816. [PubMed: 17571346]
41. Chen J, Yu K, Hsing A, Therneau TM. A partially linear tree-based regression model for assessing complex joint gene–gene and gene–environment effects. Genet. Epidemiol 2007;31:238–251. [PubMed: 17266115]
42. Li L, Cheng AS, Jin VX, et al. A mixture model-based discriminate analysis for identifying ordered transcription factor binding site pairs in gene promoters directly regulated by estrogen receptor-α. Bioinformatics 2006;22:2210–2216. [PubMed: 16809387]
43. Franke L, van Bakel H, Fokkens L, de Jong ED, Egmont-Petersen E, Wijmenga C. Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. Am. J. Hum. Genet 2006;78:1011–1025. [PubMed: 16685651]
44. Lesnick TG, Papapetropoulos S, Mash DC, et al. A genomic pathway approach to a complex disease: axon guidance and Parkinson disease. PLoS Genet 2007;3:E98. [PubMed: 17571925]
45. Conti DV, Cortessis V, Molitor J, Thomas DC. Bayesian modeling of complex metabolic pathways. Hum. Hered 2003;56:83–93. [PubMed: 14614242]
46. Conti DV, Gauderman WJ. SNPs, haplotypes, and model selection in a candidate gene region: the SIMPle analysis for multilocus data. Genet. Epidemiol 2004;27:429–441. [PubMed: 15543635]
47. Jansen R, Yu H, Greenbaum D, et al. A Bayesian networks approach for predicting protein–protein interactions from genomic data. Science 2003;302:449–453. [PubMed: 14564010]
48. Srinivasan BS, Shah NH, Flannick JA, Abeliuk E, Novak AF, Batzoglou S. Current progress in network research: toward reference networks for key model organisms. Brief. Bioinformatics 2007;8:318–332. [PubMed: 17728341]

Table 1

Important quality control considerations.

Variable	Comments
Genotyping efficiency	Low efficiency often correlates with error. Some low efficiency SNPs or samples may still be good.
Genotyping quality	Worse quality score (GenCall) correlates strongly with error rate.
Gender	Check expectations for X marker heterozygosity and Y marker-positive results. Can estimate error rate.
Mendelian inheritance errors	For trio/family data, can identify problem samples and families. Can estimate error rate.
Sample mix-ups	Check for sample duplication, contamination, switches by comparing genotypes across all samples.
Population stratification	Check for population substructure using the genome-wide data.
Linkage disequilibrium	Use the redundant data (correlated SNPs) to test for genotyping error.
Hardy–Weinberg equilibrium	Violation across all sample groups may indicate error, but can also be a good test of association.
Copy-number variants	Can create apparent genotyping error. Can check statistically from genotypes or from raw image intensities.
Platform specific problems	Affymetrix 500K has difficulty detecting rare allele homozygotes.