Published in final edited form as:

Hum Genet. 2012 October ; 131(10): 1533–1540. doi:10.1007/s00439-012-1191-1.

The role of phenotype in gene discovery in the whole genome sequencing era

Laura Almasy

Department of Genetics, Texas Biomedical Research Institute, PO Box 760549, San Antonio, TX 78245-0549, USA

Abstract

As whole genome sequence becomes a routine component of gene discovery studies in humans, we will have an exhaustive catalog of genetic variation and the challenge becomes understanding the phenotypic consequences of these variants. Statistical genetic methods and analytical approaches that are concerned with optimizing phenotypes for gene discovery for complex traits offer two general categories of advantages. They may increase power to localize genes of interest and also aid in interpreting associations between genetic variants and disease outcomes by suggesting potential mechanisms and pathways through which genes may affect outcomes. Such phenotype optimization approaches include use of allied phenotypes such as symptoms or ages of onset to reduce genetic heterogeneity within a set of cases, study of quantitative risk factors or endophenotypes, joint analyses of related phenotypes, and derivation of new phenotypes designed to extract independent measures underlying the correlations among a set of related phenotypes through approaches such as principal components. New opportunities are also presented by technological advances that permit efficient collection of hundreds or thousands of phenotypes on an individual, including phenotypes more proximal to the level of gene action such as levels of gene expression, microRNAs, or metabolic and proteomic profiles.

Introduction

In recent years, the field of complex disease genetics has been increasingly focused on maximizing the genetic information in our studies, driven largely by technological advances. We have sought to increase power for genetic studies by better capturing unobserved underlying genetic variation through the development of microsatellite and SNP maps for linkage studies and then by the use of dense SNP panels designed to maximize linkage disequilibrium with common variants for association studies. Similarly, exome sequencing represents an attempt to intelligently sample parts of the genome enriched for functional variants. One indicator of the focus of methods development in complex disease genetics is the Genetic Analysis Workshop (GAW), a forum for development, testing, and comparison of statistical genetic methods held every 2 years. GAW 14 focused on the transition from microsatellites to SNPs for genome scanning (Bailey-Wilson et al. 2005). Genome-wide association was the topic of GAW16 (Cupples et al. 2009) and analysis of exome sequence data was featured at GAW17 (Ghosh et al. 2011). However, this sweep of technological development is reaching its natural conclusion. Costs for whole genome sequencing are rapidly dropping and it will soon become the standard means of genome scanning for human complex disease genetics. We will begin our studies knowing that the functional variants of interest will be among our genotyped markers and increasing power by improving coverage

of the genome will cease to be a key issue. The challenge then for maximizing power to detect effects of interest will be to make better use of this comprehensive data on DNA variation by optimizing other aspects of the study design, such as the choice of individuals and phenotypes to be studied.

In the GWAS era, the hopes for advancement of the field were invested in maximizing the informativeness of the genotype data available for study with phenotype most frequently simplified to a case/control dichotomy. In the post-GWAS era, attention will now turn to the phenotype half of the phenotype–genotype connection. High throughput methods for proteomics, metabolomics, and transcriptomics as well as growing caches of data in electronic medical records are introducing technology-driven enhancements in the availability and comprehensiveness of phenotype information. One might say that this is the start of a “phenome project” that is the natural successor to the human genome project, as having a complete catalog of human genome sequence does not tell us what that sequence does. In essence, we will have the genotypes and will be searching for the phenotypes whose variation they influence. Methodologically, approaches originally developed to deal with genetic heterogeneity or extract additional phenotypic information to improve power in linkage studies will see a resurgence as they are applied to GWAS and whole genome sequencing studies.

Phenotypic complexity: challenges and opportunities

It is generally accepted that complex human phenotypes that are heritable are most likely influenced by multiple genes and their interactions with each other and with the environment. Thus far, the common variants with allele frequencies over 1 % that are well captured through GWAS appear to account for a small portion of the heritability of complex traits. Given this ‘missing heritability’ unaccounted for by GWAS, there is growing realization that many complex human phenotypes are likely influenced by a collection of rare variants of varying effect sizes and that two individuals with the same clinical diagnosis may carry entirely different risk alleles. One implication of this is that what is a single disorder on a clinical level may actually be a heterogeneous group of potentially overlapping disorders on a genetic or etiologic level. Such heterogeneity presents both challenges and opportunities. All methods of gene localization are sensitive to locus heterogeneity in which different genes contribute to phenotypic variation in different subsets of individuals, and association methods are also sensitive to allelic heterogeneity in which different variants in the same locus affect variation. In these cases, heterogeneity reduces power for gene localization.

One approach to dealing with this potential complexity is to try to use clinical features or quantitative measures to identify etiologically more homogeneous subsets within a group of patients with a given diagnosis. If the number of individuals genotyped or sequenced is a limiting factor, concentrating these scarce resources on a more homogeneous subset of individuals, and thereby limiting both locus and allelic heterogeneity, should increase the power for gene localization. A classical example of the successful application of this approach is the division of affected individuals into early- and late-onset groups, which facilitated the identification of APOE as a major locus for Alzheimer’s disease (St. George-Hyslop et al. 1987) and BRCA1 for breast cancer (Hall et al. 1990). Hall et al. noted that only a portion of their 23 extended families with multiple cases of breast cancer showed evidence of linkage on chromosome 17 and found that the linkage signal was strongest ($\text{LOD} = 5.98$) in the seven families where the mean age at diagnosis was 45 years and linkage to the same location was formally excluded ($\text{LOD} < -2$) in the families with older ages of onset. Use of age at onset to identify potentially more homogeneous subgroups of cases continues to be applied in GWAS studies of traits such as obesity (Meyre et al. 2009),

myocardial infarction (Myocardial Infarction Genetics Consortium et al. 2009), and major depression (Shi et al. 2011). The Myocardial Infarction Consortium began their GWAS with a cohort of cases with early-onset myocardial infarction (MI), defined as age 50 in men or 60 in women, and followed this with replication samples where the cases were chosen to be at or below the mean age of onset for MI. They identified nine loci associated with MI at genome-wide levels of significance. Four of these loci had been reported in a previous MI GWAS study that did not utilize age at onset; two loci had previously been associated with LDL, a quantitative risk factor for MI; and the final three were novel.

For any given complex disease, there may be a variety of potential markers of heterogeneity. Familial aggregation may be useful in selecting among them. Family members are likely to share variants that influence the trait of interest, and in theory phenotypes that index a particular underlying risk allele should also be correlated among those family members. Alternatively, cluster analysis may be used to jointly analyze numerous clinical features to identify subgroups of patients who share a set of characteristics, so that they may be analyzed as a separate group. For example, Gelernter et al. (2006) utilized cluster analysis of symptom variables related to opioid use from the Semi-Structured Assessment for Drug Dependence and Alcoholism to identify five phenotypic subgroups within participants in a linkage study of opioid and cocaine dependence. The highest LOD score observed in their study for a conventionally defined diagnosis of opioid dependence was 2.43. In contrast, two LOD scores >3 were obtained with cluster analysis phenotypes, one with affection status defined by membership in a cluster of heavy opioid users and one based on a cluster of non-opioid users with a high rate of cocaine dependence. A recent variation on this approach utilizes cluster analysis of high-dimensional proteomic, metabolomic, or transcriptomic data to identify subgroups of patients. For example, Harvey et al. (2010) examined gene expression profiles in 207 children with high-risk B-precursor acute lymphoblastic leukemia to identify eight subgroups of patients. Supporting the idea that these mRNA-defined clusters may represent genetically homogeneous groups of patients, they showed that two of the clusters were associated with previously identified translocations.

Attempts to reduce heterogeneity in a sample are, however, not without costs. One that is obvious is that subsetting the sample through cluster analysis or on the basis of some clinical feature necessarily reduces sample size; for continuous variables, it requires the selection of a threshold at which to divide the sample, the choice of which might not always be clear. One method of dealing with this uncertainty is to use ordered subset analysis in which cases or families are ordered on the basis of a quantitative variable (e.g., age at onset) and the analysis is performed multiple times with the threshold for inclusion being systematically increased or decreased (Chung et al. 2008; Qin et al. 2010). This may be used to identify the threshold at which a gene localization signal is maximized; however, it makes interpretation of statistical significance problematic as it inherently increases multiple testing. Typically, permutation of the data set is used to demonstrate that the observed signal is unlikely to occur by chance in random subsets of the data.

Intermediate phenotypes, endophenotypes, and risk factors

Another strategy for dealing with underlying genetic heterogeneity and potentially a complex network of etiologies contributing to variation in a phenotype is to directly study quantitative risk factors that may index different underlying aspects of disease etiology in the hope that these intermediate phenotypes may be genetically less complex with potentially stronger genetic signals. The incorporation of quantitative risk factors into genetic studies of complex disease has varied greatly between different phenotypic areas. Such studies have a long and productive history in cardiovascular disease research, where phenotypes such as cholesterol levels and blood pressure have long been included in genetic

studies. In contrast, in psychiatric genetics, such quantitative risk factors have been named endophenotypes and have been less readily accepted.

One argument used against genetic studies of endophenotypes is that frequently the heritability for the clinical end point, i.e., the psychiatric disorder, is higher than the heritability of a proposed endophenotype. However, as heritability is an aggregate composite of the effects of an unknown number of genetic loci with unknown individual effect sizes, the overall heritability of a trait is not necessarily a useful indicator of the power to detect genes. It is important that non-zero heritability be demonstrated to support gene localization for a trait and the overall heritability certainly places an upper bound on possible effect sizes of individual loci. But a trait with a heritability of 80 % may be more difficult to localize genes for than a trait with a heritability of 40 % if the heritability for the first trait is due to dozens or hundreds of loci of very small effect (e.g., height), and the heritability for the second trait is due to a handful of loci some of which have moderate effect sizes (e.g., levels of gene expression). In general, the closer a phenotype is to the level of gene action, the larger we might expect the genetic effects to be. If a genetic variant alters transcription factor binding, which leads to decreased gene expression that decreases protein levels thus impairing physiological function and increasing risk of disease, the effect size of that variant is largest on transcription factor binding, still quite substantial on gene expression, moderate on physiological function, and smallest on disease risk. As yet, there are relatively few examples directly comparing the effect of the same functional variant in the same sample at these different levels of physiological functioning. Göring et al. (2007) utilized gene expression measures to search for genes involved in regulation of HDL cholesterol and identified a promoter variant affecting a transcription binding site upstream of *VNN1* that showed exceptionally strong association with *VNN1* transcript levels ($p = 5.7 \times 10^{-83}$) and weaker association with HDL cholesterol measures (4.0×10^{-4}), consistent with the notion that effect sizes are highest for phenotypic measures most proximal to the action of the variant. High throughput transcriptomic, metabolomic, and proteomic approaches now provide opportunities to screen thousands of phenotypes close to the level of gene action to identify novel quantitative risk factors.

For more traditional quantitative risk factors, even if they are not genetically simpler, under most circumstances they offer better power for gene localization than comparable dichotomous phenotypes simply because of the loss of information inherent in dichotomizing what is essentially a continuous process. For an extreme example, direct analyses of body mass index (BMI) are better powered than analyses of the yes/no trait of obesity defined as a threshold on the BMI continuum. In this case, it is obvious that variations in BMIs over the threshold provide meaningful information about phenotype severity among affected individuals and variations in BMIs under the threshold may provide information about greater and lesser degrees of liability among unaffected individuals. Blangero et al. (2003) assessed the loss of information for gene localization incurred in dichotomizing a directly measurable quantitative phenotype and found that the relative efficiency of the discrete categorization, as compared to the original quantitative measure, varied from 0.01 to 0.33 depending on the prevalence of the discrete phenotype. Quantitative ways of measuring the phenotype of interest also better reflect our understanding of the genetics of complex human traits than yes/no dichotomous phenotypes (Plomin et al. 2009). In general, we conceptualize common, complex human diseases as resulting from an accumulation of risk factors under a ‘liability threshold’ model (Falconer 1989). That is, rather than individuals with and without disease being different in kind, they are different in degree. It is likely that many unaffected individuals carry risk alleles, just fewer of them than affected individuals.

To establish a quantitative risk factor or endophenotype, it is necessary to document that the measure differs between affected and unaffected individuals and is heritable (Gottesman and Gould 2003). Further, it needs to vary meaningfully in unaffected individuals such that it is differentiating degrees of risk rather than simply recapitulating the yes/no diagnostic dichotomy. And for use in genetic studies, it should ideally be practically and economically feasible to measure in large numbers of individuals. If it is indexing disease liability, the quantitative risk factor should also differ not only between affected and unaffected individuals but also between unaffected relatives of patients and unaffected individuals with no family history of disease. That phenotypes in unaffected relatives of patients should be intermediate between patients and controls is implied by the existence of pleiotropy between the quantitative risk factor and the disease, which occurs when two phenotypes are influenced by overlapping sets of genes. Although demonstrating intermediate mean trait values in unaffected relatives of patients is more common, a more direct means of establishing the existence of pleiotropy is the genetic correlation between the two traits which quantifies the degree of overlap in the genetic effects on the phenotypes (Falconer 1989).

Screening potential endophenotypes—the endophenotype ranking value

Traditionally, the time required to collect extensive medical history or behavioral data or the biological samples and costs to perform numerous laboratory assays have limited the number of phenotypes collected for a genetic study. However, new technologies are automating phenotyping as well as genotyping, and the quickly expanding opportunities to collect large numbers of potential endophenotypes, clinical measures, and quantitative risk factors correlated with disease present new challenges. With the advent of electronic medical records, it may be possible to access a wide variety of ancillary information on study participants at little or no cost to a research study, and new genetic studies are being developed from clinical populations in which patients have consented to use of their medical records for genetic research (e.g., Bielinski et al. 2011). In addition to more classical biomedical, clinical, and neuropsychological measures, high throughput “omics” technologies now provide potentially tens of thousand of quantitative risk factors in the form of measures of gene transcription, microRNA levels, proteomics, and metabolomics. These measures are likely close to the level of gene action, improving power to detect genetic effects, but the sheer number of phenotypes available compounds the multiple testing problems inherent in genome-wide screening if such screens are conducted for potentially hundreds or thousands of phenotypes. The endophenotype ranking value (ERV; Glahn et al. 2012) represents a simple mechanism for objectively ranking quantitative risk factors to identify those that may provide the best power to localize genes with pleiotropic effects on a related human disease. The ERV is based on the standardized genetic covariance between the risk factor and the disease and is a simple composite of the heritability of the endophenotype (h_{endo}^2), the heritability of the disease (h_{dis}^2), and the genetic correlation between them (ρ_G):

$$\text{ERV} = \text{sqrt}(h_{\text{endo}}^2) \text{sqrt}(h_{\text{dis}}^2) \rho_G$$

Varying between zero and one, with higher values representing a stronger endophenotype, the ERV identifies phenotypes with a balance between a strong genetic signal (the heritability of the endophenotype) and pleiotropy between the endophenotype and the disease of interest (the genetic correlation). An ideal quantitative risk factor has both a high heritability and a high genetic correlation with disease liability. Glahn et al. (2012) screened 11,000-plus measures of RNA transcription in lymphocytes and identified eight genes

whose expression levels were as strong or stronger endophenotypes for recurrent major depression as the Beck Depression Inventory (BDI).

The “reverse endophenotype” approach

In addition to their value for gene finding, it has been suggested that endophenotypes or quantitative risk factors may also facilitate the understanding and interpretation of loci identified through genome scans of a disease end point. By their nature, complex human diseases likely involve multiple potential biological pathways through which risk alleles may affect liability to disease. Demonstrating an association between a particular genetic marker and disease risk is only a first step to understanding how a gene and its variants may affect disease and, therefore, what kind of interventions on the underlying biological systems may treat or prevent illness.

The ‘reverse endophenotype’ approach advocates using clinical features or quantitative measures of intermediate phenotypes not for gene finding, but for understanding the action of previously identified loci. Consider a locus associated with risk of heart disease. It might act through effects on blood pressure homeostasis, lipid metabolism, inflammation, or any of a variety of other mechanisms. Identifying which of these quantitative clinical markers is also associated with the locus helps to clarify how variation in the gene affects disease risk, particularly in cases where the function of the associated gene was previously unknown or in cases where a gene of known function could affect one or more of a number of underlying pathways such as is the case for transcription factors. This may help to inform future functional studies of the gene. For example, Denny et al. (2011) screened electronic medical records of 13,617 individuals to identify additional phenotypes also associated with an SNP near *FOXE1* that was previously identified through case/control analyses of hypothyroidism. They found that frequency of thyroiditis, goiters, and thyrotoxicosis differed by SNP genotype, but that Graves disease did not. A closely related concept is the phenome-wide association study (Pendergrass et al. 2011), in which a large number of phenotypes are screened and the patterns of results examined for additional insight into potential pleiotropic effects.

Other techniques discussed above may also be applied with the goal of interpreting existing signals, rather than increasing the power for localization. For example, ordered subset analysis may be used to identify clinical features associated with a previously detected locus or to select a subset of cases with the strongest evidence for localization for sequencing or other follow-up studies.

Multiple testing

Even with the application of ranking systems such as the ERV, it is likely that a genome-wide screen will utilize multiple phenotypes. While the cost of whole genome sequencing is rapidly decreasing, it is still not trivial and investigators will naturally want to extract the maximum possible scientific value from these expensive and ambitious studies by studying as many relevant phenotypes for which they have adequate power to detect genetic signals. Such phenotype screening compounds the already formidable multiple testing issues inherent in genome-wide sequence screens. This issue is not unique to whole genome sequence, of course, and has been raised repeatedly in the context of linkage and genome-wide association screens. Essentially, if many phenotypes are tested, but the usual genome-wide p value threshold is used, the results are not as protected against false positives as that p value would suggest. The greater the number of phenotypes screened, the greater are the chances of exceeding that genome-wide p value threshold by chance alone.

Compounding the problem is the fact that the phenotypes being screened will typically be correlated with each other, as they were likely selected for their common relation to a disease of interest. This correlation among phenotypes makes a Bonferroni correction overly conservative. There are a number of potential strategies for dealing with multiple testing issues raised by the use of multiple phenotypes. From an analytical standpoint, the simplest is to reimagine the study as hypothesis generating, rather than hypothesis testing, and rely on replication of a small number of specific phenotype/genotype hypotheses in an independent sample to establish the significance threshold. Of course, this may not always be possible, particularly for phenotypes that are not commonly measured or are expensive to obtain and therefore may not be available in sufficient numbers of individuals in more than one genetic study. An alternative is to utilize the correlational structure between the phenotypes to estimate an effective number of independent phenotypes, which can be done easily using methods originally developed for estimating the number of independent SNPs given the pattern of linkage disequilibrium between them (e.g., Moskva and Schmidt 2008). A Bonferroni correction to the desired p value threshold can then be made based on the effective number of independent phenotypes, taking into account the correlations between them. Stein et al. (2012) utilized this approach in a meta-analysis of GWAS of hippocampal, intracranial, and total brain volumes. They examined associations for eight measures of brain volume and used a matrix of cross-correlations among phenotypes that was calculated from the t statistics for associations with each of these phenotypes with each GWAS SNP. The eigenvalues of the correlation matrix were then used to determine that these eight measures were equivalent to testing of four independent phenotypes, and the desired significance threshold for their GWAS was Bonferroni corrected for the effective number of independent phenotypes tested.

Joint analyses of multiple phenotypes

Many studies of quantitative risk factors screen the genome with these measures and then test whether associated markers also show a signal with a disease of interest. Particularly in the case of studies of normal variation, the genome screening results for multiple traits that are known to be related may be examined for concordance of findings across phenotypes. An alternative to these sequential sorts of approaches is joint analysis of multiple phenotypes. Methods for joint analyses of multiple traits have been developed for both linkage (e.g., Almasy et al. 1997) and association (e.g., Saint-Pierre et al. 2011). While joint analyses reduce multiple testing and may increase power, they generally come at the cost of additional degrees of freedom in the test statistic. Additionally, although in theory these multivariate analysis methods can be expanded to include any number of phenotypes, in practice, they can become computationally prohibitive as the number of phenotypes increases and it is uncommon to see more than two phenotypes in an analysis and quite rare to find more than three phenotypes analyzed jointly.

Theoretical and simulation studies suggest that multitrait analyses increase power to localize genes influencing a phenotype of interest under a variety of genetic models, but that power is maximized when there are multiple loci with genetic correlations of different directions jointly influencing two phenotypes such that the correlation due to the locus of interest and the correlation due to residual genetic effects of other loci are in opposite directions (Amos et al. 2001; Saint-Pierre et al. 2011). In addition to simulations documenting the improvement in power of their bivariate association approach, Saint-Pierre et al. illustrated the potential results in a GWAS of bone mineral density measured at the lumbar spine and femoral neck. All of the top association signals from univariate GWAS of either trait alone were also picked up in the bivariate GWAS. In contrast, of the top 100 associations from the bivariate GWAS, 52 SNPs also had strong signals from at least one of the univariate scans and many markers had weaker signals, but 16 SNPs with strong signals in the bivariate

GWAS failed to reach even nominal significance ($p < 0.05$) in either of the univariate GWAS, suggesting that the bivariate approach may pick up associations that would otherwise be missed.

Deriving a new phenotype that captures commonalities among a set of phenotypes

An alternate approach to jointly analyzing a set of related traits is to attempt to derive a new phenotype or phenotypes that represent one or more underlying common factors shared among the set of traits. Ideally, these new phenotypes are orthogonal to each other and uncorrelated in the hope that each of them represents a different underlying pathway or domain of genetic influence. This is commonly done by factor analysis and particularly by principal components analysis. If the set of related phenotypes is particularly large, this approach may greatly reduce the multiple testing that would be incurred in conducting separate genome screens for each phenotype and is computationally more tractable than full multivariate analyses. A variant of this strategy is the “Principal Components of Heritability” (PCH) method in which rather than maximizing the overall phenotypic variance explained by the largest derived components, the approach maximizes the heritability of the main components, i.e., the genetic variance explained by the component (Ott and Rabinowitz 1999). While there are numerous examples of linkage studies conducted on principal components derived from suites of related phenotypes (e.g., Dick et al. 2002; Arya et al. 2002), this approach has not yet been widely applied in GWAS.

The cluster analysis approach described above to identify more homogenous subgroups of patients may also be used as a means of extracting new phenotypes from a set of clinical features. In this application, rather than assigning patients to a particular subgroup, the probability of group membership obtained from the cluster analysis is analyzed as a quantitative trait for each individual for each subgroup. For example, Wilcox et al. (2009) utilized this approach in a GWAS of phenotypes related to cardiovascular disease in the Framingham Heart Study. They identified five clusters among participants in the Framingham Offspring Cohort, including one characterized by high rates of obesity and one by features of the metabolic syndrome. A case/control GWAS comparing cases in these clusters to controls in two other clusters resulted in one genome-wide significant SNP on chromosome 4. GWAS using probability of membership in the obesity and metabolic syndrome clusters also identified the region on chromosome 4 with stronger p values.

Conclusions

Methods and study designs that take advantage of phenotype complexity, intermediate phenotypes or endophenotypes, and multiple related phenotypes offer two primary advantages in the age of genome-wide sequencing studies. They can aid in interpreting the results of our gene localization and identification studies, helping to understand the biological pathways through which DNA variants have their effects on human variation and allowing us to form more detailed hypotheses for follow-up functional studies. They may also improve power for gene localization and identification studies by reducing or explicitly modeling genetic heterogeneity, by utilizing phenotypes more proximal to gene action, or by exploiting correlational structures to identify genes jointly influencing suites of related phenotypes. Many of these approaches were developed in the linkage era and have, as yet, not been widely applied in GWAS or sequencing studies. Simulations to assess power and false positive rates in these contexts and additional examples of practical applications of these strategies will be needed to establish their potential utility. On the phenotyping end, continued technological developments in high-throughput gene expression and metabolomic and proteomic techniques have made possible the collection of thousands of measures on

each individual that will present analytical challenges as well as opportunities, and continued development of statistical genetic methods to take advantage of richer and more detailed phenome data will be required.

Acknowledgments

This work was supported in part by R01 MH59490 from the National Institute of Mental Health and R01 GM31575 from the National Institute of General Medical Sciences.

References

- Almasy L, Dyer TD, Blangero J. Bivariate quantitative trait linkage analysis: pleiotropy versus coincident linkages. *Genet Epidemiol.* 1997; 14:953–958. [PubMed: 9433606]
- Amos CI, de Andrade M, Zhu DK. Comparison of multivariate tests for genetic linkage. *Hum Hered.* 2001; 51:133–144. [PubMed: 11173964]
- Arya R, Blangero J, Williams K, Almasy L, Dyer TD, Leach RJ, O'Connell P, Stern MP, Duggirala R. Factors of insulin resistance syndrome-related phenotypes are linked to genetic locations on chromosomes 6 and 7 in nondiabetic Mexican-Americans. *Diabetes.* 2002; 51:841–847. [PubMed: 11872689]
- Bailey-Wilson JE, Almasy L, de Andrade M, Bailey J, Bickeböllner H, Cordell H, Daw W, Goldin L, Goode E, Gray-Mcguire C, Hening W, Jarvik G, Maher B, Mendell N, Paterson A, Rice J, Satten G, Suarez B, Vieland V, Wilcox M, Zhang H, Ziegler A, MacCluer JW. Genetic Analysis Workshop 14: Microsatellite and SNP marker loci for genome-wide scans. *BMC Genet.* 2005; 6:S1. [PubMed: 16451554]
- Bielinski SJ, Chai HS, Pathak J, Talwalkar JA, Limburg PJ, Gullerud RE, Sicotte H, Klee EW, Ross JL, Kocher JP, Kullo IJ, Heit JA, Petersen GM, de Andrade M, Chute CG. Mayo Genome Consortia: a genotype–phenotype resource for genome-wide association studies with an application to the analysis of circulating bilirubin levels. *Mayo Clin Proc.* 2011; 86:606–614. [PubMed: 21646302]
- Blangero J, Williams JT, Almasy L. Novel family-based approaches to genetic risk in thrombosis. *J Thromb Haemost.* 2003; 1:1391–1397. [PubMed: 12871272]
- Chung RH, Schmidt S, Martin ER, Hauser ER. Ordered-subset analysis (OSA) for family-based association mapping of complex traits. *Genet Epidemiol.* 2008; 32:627–637. [PubMed: 18473393]
- Cupples LA, Beyene J, Bickeboeller H, Daw EW, Fallin MD, Gauderman WJ, Ghosh S, Goode EL, Hauser ER, Hinrichs A, Kent JW Jr, Martin LJ, Martinez M, Neuman RJ, Province M, Szymczak S, Wilcox MA, Ziegler A, MacCluer JW, Almasy L. Genetic Analysis Workshop 16: strategies for genome-wide association study analyses. *BMC Proc.* 2009; 3(Suppl 7):S1. [PubMed: 20017962]
- Denny JC, Crawford DC, Ritchie MD, Bielinski SJ, Basford MA, Bradford Y, Chai HS, Bastarache L, Zuvich R, Peissig P, Carrell D, Ramirez AH, Pathak J, Wilke RA, Rasmussen L, Wang X, Pacheco JA, Kho AN, Hayes MG, Weston N, Matsumoto M, Kopp PA, Newton KM, Jarvik GP, Li R, Manolio TA, Kullo IJ, Chute CG, Chisholm RL, Larson EB, McCarty CA, Masys DR, Roden DM, de Andrade M. Variants near FOXE1 are associated with hypothyroidism and other thyroid conditions: using electronic medical records for genome- and phenome-wide studies. *Am J Hum Genet.* 2011; 89:529–542. [PubMed: 21981779]
- Dick DM, Nurnberger J Jr, Edenberg HJ, Goate A, Crowe R, Rice J, Bucholz KK, Kramer J, Schuckit MA, Smith TL, Porjesz B, Begleiter H, Hesselbrock V, Foroud T. Suggestive linkage on chromosome 1 for a quantitative alcohol-related phenotype. *Alcohol Clin Exp Res.* 2002; 26:1453–1460. [PubMed: 12394277]
- Falconer, DS. Quantitative genetics. 3rd edn. Wiley; New York: 1989.
- Gelernter J, Panhysen C, Wilcox M, Hesselbrock V, Rounsaville B, Poling J, Weiss R, Sonne S, Zhao H, Farrer L, Kranzler HR. Genome wide linkage scan for opioid dependence and related traits. *Am J Hum Genet.* 2006; 78:759–769. [PubMed: 16642432]
- Ghosh S, Bickeboeller H, Bailey J, Bailey-Wilson J, Cantor R, Culverhouse R, Daw W, DeStefano A, Engelman C, Hemmelman C, Hinrichs A, Houwing-Duistermaat J, Koenig I, Kent J Jr, Pankratz N, Paterson A, Pugh E, Suarez B, Sun Y, Thomas A, Tintle N, Zhu X, MacCluer J, Almasy L.

- Identifying rare variants from exome scans: The GAW17 experience. *BMC Proc.* 2011; 5(Suppl 9):S1–1. [PubMed: 22373325]
- Glahn DC, Curran JE, Winkler AM, Carless MA, Kent JW Jr, Charlesworth JC, Johnson MP, Göring HH, Cole SA, Dyer TD, Moses EK, Olvera RL, Kochunov P, Duggirala R, Fox PT, Almasy L, Blangero J. High dimensional endophenotype ranking in the search for major depression risk genes. *Biol Psychiatry.* 2012; 71:6–14. [PubMed: 21982424]
- Göring HH, Curran JE, Johnson MP, Dyer TD, Charlesworth J, Cole SA, Jowett JB, Abraham LJ, Rainwater DL, Comuzzie AG, Mahaney MC, Almasy L, MacCluer JW, Kissebah AH, Collier GR, Moses EK, Blangero J. Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. *Nat Genet.* 2007; 39:1208–1216. [PubMed: 17873875]
- Gottesman II, Gould TD. The endophenotype concept in psychiatry: etymology and strategic intentions. *Am J Psychiatry.* 2003; 160:636–645. [PubMed: 12668349]
- Hall JM, Lee MK, Newman B, Morrow JE, Anderson LA, Huey B, King MC. Linkage of early-onset familial breast cancer to chromosome 17q21. *Science.* 1990; 250:1684–1689. [PubMed: 2270482]
- Harvey RC, Mullighan CG, Wang X, Dobbin KK, Davidson GS, Bedrick EJ, Chen IM, Atlast SR, Kang H, Ar K, Wilson CS, Wharton W, Murphy M, Devidas M, Carroll AJ, Borowitz MJ, Bowman WP, Downing JR, Relling M, Yang J, Bhojwani D, Carroll WL, Camitta B, Reaman GH, Smith M, Hunger SP, Willman CL. Identification of novel cluster groups in pediatric high-risk B-precursor acute lymphoblastic leukemia with gene expression profiling: correlation with genome-wide DNA copy number alterations, clinical characteristics, and outcome. *Blood.* 2010; 116:4874–4884. [PubMed: 20699438]
- Meyre D, Delplanque J, Chèvre JC, Lecoœur C, Lobbens S, Gallina S, Durand E, Vatin V, Degraeve F, Proença C, Gaget S, Körner A, Kovacs P, Kiess W, Tichet J, Marre M, Hartikainen AL, Horber F, Potoczna N, Hercberg S, Levy-Marchal C, Pattou F, Heude B, Tauber M, McCarthy MI, Blakemore AI, Montpetit A, Polychronakos C, Weill J, Coin LJ, Asher J, Elliott P, Järvelin MR, Visvikis-Siest S, Balkau B, Sladek R, Balding D, Walley A, Dina C, Froguel P. Genome-wide association study for early-onset and morbid adult obesity identifies three new risk loci in European populations. *Nat Genet.* 2009; 41:157–159. [PubMed: 19151714]
- Moskvina V, Schmidt KM. On multiple-testing correction in genome-wide association studies. *Genet Epidemiol.* 2008; 32:567–573. [PubMed: 18425821]
- Myocardial Infarction Genetics Consortium; Kathiresan S, Voight BF, Purcell S, Musunuru K, Ardisson D, Mannucci PM, Anand S, Engert JC, Samani NJ, Schunkert H, Erdmann J, Reilly MP, Rader DJ, Morgan T, Spertus JA, Stoll M, Girelli D, McKeown PP, Patterson CC, Siscovick DS, O'Donnell CJ, Elosua R, Peltonen L, Salomaa V, Schwartz SM, Melander O, Altshuler D, Ardisson D, Merlini PA, Berzuini C, Bernardinelli L, Peyvandi F, Tubaro M, Celli P, Ferrario M, Fève R, Marziliano N, Casari G, Galli M, Ribichini F, Rossi M, Bernardi F, Zonin P, Piazza A, Mannucci PM, Schwartz SM, Siscovick DS, Yee J, Friedlander Y, Elosua R, Marrugat J, Lucas G, Subirana I, Sala J, Ramos R, Kathiresan S, Meigs JB, Williams G, Nathan DM, MacRae CA, O'Donnell CJ, Salomaa V, Havulinna AS, Peltonen L, Melander O, Berglund G, Voight BF, Kathiresan S, Hirschhorn JN, Asselta R, Duga S, Sreafico M, Musunuru K, Daly MJ, Purcell S, Voight BF, Purcell S, Nemesh J, Korn JM, McCarroll SA, Schwartz SM, Yee J, Kathiresan S, Lucas G, Subirana I, Elosua R, Surti A, Guiducci C, Gianniny L, Mirel D, Parkin M, Burt N, Gabriel SB, Samani NJ, Thompson JR, Braund PS, Wright BJ, Balmforth AJ, Ball SG, Hall AS, Wellcome Trust Case Control Consortium. Schunkert H, Erdmann J, Linsel-Nitschke P, Lieb W, Ziegler A, König I, Hengstenberg C, Fischer M, Stark K, Grosshennig A, Preuss M, Wichmann HE, Schreiber S, Schunkert H, Samani NJ, Erdmann J, Ouwehand W, Hengstenberg C, Deloukas P, Scholz M, Cambien F, Reilly MP, Li M, Chen Z, Wilensky R, Matthai W, Qasim A, Hakonarson HH, Devaney J, Burnett MS, Pichard AD, Kent KM, Satler L, Lindsay JM, Waksman R, Knouff CW, Waterworth DM, Walker MC, Mooser V, Epstein SE, Rader DJ, Scheffold T, Berger K, Stoll M, Hogue A, Girelli D, Martinelli N, Olivieri O, Corrocher R, Morgan T, Spertus JA, McKeown P, Patterson CC, Schunkert H, Erdmann E, Linsel-Nitschke P, Lieb W, Ziegler A, König IR, Hengstenberg C, Fischer M, Stark K, Grosshennig A, Preuss M, Wichmann HE, Schreiber S, Hólm H, Thorleifsson G, Thorsteinsdóttir U, Stefansson K, Engert JC, Do R, Xie C, Anand S, Kathiresan S, Ardisson D, Mannucci PM, Siscovick D, O'Donnell CJ, Samani NJ, Melander O, Elosua R, Peltonen L, Salomaa V, Schwartz SM, Altshuler D. Genome-wide

- association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants. *Nat Genet.* 2009; 41:334–41. [PubMed: 19198609]
- Ott J, Rabinowitz D. A principal-components approach based on heritability for combining phenotype information. *Hum Hered.* 1999; 49:106–111. [PubMed: 10077732]
- Pendergrass SA, Brown-Gentry K, Dudek SM, Torstenson ES, Ambite JL, Avery CL, Buyske S, Cai C, Fesinmeyer MD, Haiman C, Heiss G, Hindorff LA, Hsu CN, Jackson RD, Kooperberg C, Le Marchand L, Lin Y, Matise TC, Moreland L, Monroe K, Reiner AP, Wallace R, Wilkens LR, Crawford DC, Ritchie MD. The use of phenome-wide association studies (PheWAS) for exploration of novel genotype–phenotype relationships and pleiotropy discovery. *Genet Epidemiol.* 2011; 35:410–422. [PubMed: 21594894]
- Plomin R, Haworth CM, Davis OS. Common disorders are quantitative traits. *Nat Rev Genet.* 2009; 10:872–878. [PubMed: 19859063]
- Qin X, Hauser ER, Schmidt S. Ordered subset analysis for case–control studies. *Genet Epidemiol.* 2010; 34:407–417. [PubMed: 20568256]
- Saint-Pierre A, Kaufman JM, Ostertag A, Cohen-Solal M, Boland A, Toye K, Zelenika D, Lathrop M, de Vernejoul MC, Martinez M. Bivariate association analysis in selected samples: application to a GWAS of two bone mineral density phenotypes in males with high or low BMC. *Eu J Hum Genet.* 2011; 19:710–716.
- Shi J, Potash JB, Knowles JA, Weissman MM, Coryell W, Scheftner WA, Lawson WB, DePaulo JR Jr, Gejman PV, Sanders AR, Johnson JK, Adams P, Chaudhury S, Jancic D, Evgrafov O, Zvinyatskovskiy A, Ertman N, Gladis M, Neimanas K, Goodell M, Hale N, Ney N, Verma R, Mirel D, Holmans P, Levinson DF. Genome-wide association study of recurrent early-onset major depressive disorder. *Mol Psychiatry.* 2011; 16:193–201. [PubMed: 20125088]
- St. George-Hyslop PH, Tanzi RE, Polinsky RJ, Haines JL, Nee L, Watkins PC, Myers RH, Feldman RG, Pollen D, Drachman D. The genetic defect causing familial Alzheimer's disease maps on chromosome 21. *Science.* 1987; 235:885–890. [PubMed: 2880399]
- Stein JL, Medland SE, Vasquez AA, Hibar DP, Senstad RE, Winkler AM, Toro R, Appel K, Barteczek R, Bergmann O, Bernard M, Brown AA, Cannon DM, Chakravarty MM, Christoforou A, Domin M, Grimm O, Hollinshead M, Holmes AJ, Homuth G, Hottenga JJ, Langan C, Lopez LM, Hansell NK, Hwang KS, Kim S, Laje G, Lee PH, Liu X, Loth E, Lourdasamy A, Mattingdal M, Mohnke S, Maniega SM, Nho K, Nugent AC, O'Brien C, Papmeyer M, Pütz B, Ramasamy A, Rasmussen J, Rijpkema M, Risacher SL, Roddey JC, Rose EJ, Ryten M, Shen L, Sprooten E, Strengman E, Teumer A, Trabzuni D, Turner J, van Eijk K, van Erp TG, van Tol MJ, Wittfeld K, Wolf C, Woudstra S, Aleman A, Alhusaini S, Almasy L, Binder EB, Brohawn DG, Cantor RM, Carless MA, Corvin A, Czisch M, Curran JE, Davies G, de Almeida MA, Delanty N, Depondt C, Duggirala R, Dyer TD, Erk S, Fagerness J, Fox PT, Freimer NB, Gill M, Göring HH, Hagler DJ, Hoehn D, Holsboer F, Hoogman M, Hosten N, Jahanshad N, Johnson MP, Kasperaviciute D, Kent JW Jr, Kochunov P, Lancaster JL, Lawrie SM, Liewald DC, Mandl R, Matarin M, Mattheisen M, Meisenzahl E, Melle I, Moses EK, Mühleisen TW, Nauck M, Nöthen MM, Olvera RL, Pandolfo M, Pike GB, Puls R, Reinvang I, Rentería ME, Rietschel M, Roffman JL, Royle NA, Rujescu D, Savitz J, Schnack HG, Schnell K, Seiferth N, Smith C, Steen VM, Valdés Hernández MC, Van den Heuvel M, van der Wee NJ, Van Haren NE, Veltman JA, Völzke H, Walker R, Westlye LT, Whelan CD, Agartz I, Boomsma DI, Cavalleri GL, Dale AM, Djurovic S, Drevets WC, Hagoort P, Hall J, Heinz A, Jack CR Jr, Foroud TM, Le Hellard S, Maciardi F, Montgomery GW, Poline JB, Porteous DJ, Sisodiya SM, Starr JM, Sussmann J, Toga AW, Veltman DJ, Walter H, Weiner MW, the Alzheimer's Disease Neuroimaging Initiative (ADNI); EPIGEN Consortium; IMA-GEN Consortium; Saguenay Youth Study Group (SYS); Bis JC, Ikram MA, Smith AV, Gudnason V, Tzourio C, Vernooij MW, Launer LJ, Decarli C, Seshadri S, Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium; the Enhancing Neuro Imaging Genetics through Meta-Analysis (ENIGMA) Consortium. Andreassen OA, Apostolova LG, Bastin ME, Blangero J, Brunner HG, Buckner RL, Cichon S, Coppola G, de Zubicaray GI, Deary IJ, Donohoe G, de Geus EJ, Espeseth T, Fernández G, Glahn DC, Grabe HJ, Hardy J, Hulshoff Pol HE, Jenkinson M, Kahn RS, McDonald C, McIntosh AM, McMahon FJ, McMahon KL, Meyer-Lindenberg A, Morris DW, Müller-Myhsok B, Nichols TE, Ophoff RA, Paus T, Pausova Z, Penninx BW, Potkin SG, Sämann PG, Saykin AJ, Schumann G, Smoller JW, Wardlaw JM, Weale ME, Martin NG, Franke B, Wright MJ, Thompson PM. Identification of common variants

associated with human hippocampal and intracranial volumes. *Nat Genet.* 2012; 44:552–561. [PubMed: 22504417]

Wilcox M, Li Q, Sun Y, Stang P, Berlin J, Wang D. Genome-wide association study for empirically derived metabolic phenotypes in the Framingham Heart Study offspring cohort. *BMC Proc.* 2009; 3(Suppl 7):S53. [PubMed: 20018046]

\$watermark-text

\$watermark-text

\$watermark-text