Published in final edited form as:
Clin Genet. 2009 February ; 75(2): 101–106. doi:10.1111/j.1399-0004.2008.01083.x.

Finding genes underlying human disease

CM Stein and **RC Elston**
Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, OH, USA

Abstract

With the surge in genome-wide association studies (GWAS), many have asked the question 'Are linkage studies dead?' In this article, we survey the approaches used in mapping human disease genes, reviewing the analysis strategies that preceded and laid the groundwork for GWAS. We note that earlier approaches are still useful and the development of new methodology is warranted.

Keywords

complex traits; familial aggregation; genetic epidemiology; gene mapping; statistical genetic modeling; trends in genomics

With the recent development of high-throughput single nucleotide polymorphism genotyping technology, genome-wide association studies (GWAS) have gained a great deal of popularity. While the GWAS approach to discovering human disease genes shows great promise, it is important to understand historical approaches to the identification of disease genes because, as we discuss later, many of them will continue to be important.

Overview of genetic epidemiology: past approaches

Genetic epidemiology is the study of genetic factors that determine the distributions and dynamics of diseases in populations. The study designs and statistical approaches used in this field of research can be grouped into four global categories to show for a trait of interest, in succession, familial aggregation, familial segregation, familial cosegregation with a genetic marker, and association with a genetic variant.

Familial aggregation

Genetic disorders tend to aggregate in families. Familial aggregation is the *sine qua non* for continuing the genetic analysis of a disorder. Familial aggregation can be assessed in several ways. Epidemiological studies may be used to evaluate whether individuals with a family history of a given disease have a greater risk of disease than individuals without such a family history. If a disease has a genetic basis, relatives of individuals with the disease will have greater risk, and that risk will be highest for the identical twins of individuals with the disease. The risk will decrease for individuals with affected siblings, children or parents, more distant affected relatives such as avuncular relationships and cousins, in that order, and will be lowest, equal to the population prevalence, for randomly chosen individuals. This paradigm can be used to estimate the relative recurrence risk (1,2) for a particular disorder, that is, the probability of disease recurrence in a family or its occurrence in a particular type

Corresponding author: Dr Catherine M Stein, Department of Epidemiology and Biostatistics, Case Western Reserve University, Wolstein Research Building, Room 1316, Cleveland, OH 44106, USA. Tel.: (216) 368-5630; fax: (216) 368-4880; cmstein@cwru.edu.

of relative of the proband, divided by the population prevalence. For continuous traits, heritability, defined as the proportion of the total trait variability due to genetic factors, may be estimated under certain assumptions from relative pair correlations (3).

Familial segregation

Segregation analysis is a statistical modeling method used to determine the transmission pattern of a trait within families. Early modeling strategies mapped a set of genotypes into a set of phenotypes (4). Modern segregation analysis has the aim of detecting underlying Mendelian transmission. Several characteristics of the genetic model underlying a phenotype may be gleaned from this analysis, and it is appropriate for both discrete and continuous traits. First, it must be determined whether there is transmission from the parents to the children in a manner consistent with Mendelian laws, so that a parent with genotype AB at a locus has an equal probability of transmitting either the A or B allele to each child. Second, the disease model is estimated. For a discrete trait, dominant and recessive inheritance are considered and, analogously for a continuous trait, we can ask if the trait distribution of the heterozygote AB is identical to that of either of the putative homozygotes AA or BB. Third, and more generally, parameters of a Mendelian model are estimated, including the disease allele frequency and penetrances – the genotype-specific probabilities of disease for a discrete trait or the genotype-specific trait distributions for a continuous trait. Recently, statistical approaches have advanced to evaluate other more complex models, such as multiple genetic loci, multiple alleles at each locus, anticipation, parent-of-origin effects such as imprinting (5), and variable age of onset for discrete disorders. The appropriate segregation model can then be incorporated into linkage analysis, to which we now turn.

Familial cosegregation with a genetic marker

Familial cosegregation of a trait of interest with one or several genetic markers forms the basis of linkage analysis, which seeks to place a disease locus on the genome. Genetic markers and their use have evolved significantly since they were first used for this purpose. In the 1940s, the available genetic markers were blood groups, followed by enzymes in the 1960s. Linkage analysis took a major step forward in the 1980s with the development of restriction fragment length polymorphisms (6) and in the 1990s with the development of short tandem repeats (also known as microsatellites) (7,8), both of which enabled genome-wide searches for disease loci. These genome scans used 200–400 highly polymorphic markers that were mapped fairly evenly across the genome, like a dragnet. The results of a linkage genome scan containing this many markers yields too large a region to pinpoint a specific disease locus, so these regions would be followed by additional genotyping of more densely spaced markers. This approach was most successful for localizing chromosomal regions containing rare Mendelian genes, but regions containing genes for complex traits like breast cancer (BRCA1 and BRCA2) have also been mapped in this manner (9–11). Most recently, single nucleotide polymorphisms (SNPs) were developed in the 2000s and these now comprise the most cost-efficient genetic markers.

Before discussing the different strategies for linkage analysis, it is important to mention ascertainment. Because data for genetic studies are not collected randomly, analysis models must account for the manner in which the data were ascertained. In theory, the problem of ascertainment in extended pedigrees has been solved (12,13), although whether this is a problem when the only parameters estimated are recombination fractions is debated (14). There is, however, general agreement that if the linkage analysis jointly estimates any of the parameters related to disease prevalence, allowing for the mode of ascertainment is essential. It is also important here to point out the distinction between finding genes *vs* evaluating their effects. Ascertained samples are effective for detecting and finding disease genes because these genes will likely be more common and are more likely to segregate

within families in which the disease aggregates. However, such samples are not ideal for estimating the magnitude of the genotype effects, these are better estimated in population-based epidemiological studies.

There are two broad approaches to linkage analysis. Model-based linkage analysis involves calculating, for a set of data, the likelihood of a known genetic model, which is proportional to the probability of these data given that genetic model. The ultimate goal is to find the location of a gene underlying the trait relative to one or more marker loci (15). The method is called model-based linkage analysis because it uses information about an assumed genetic model (e.g. dominant, recessive, or additive). The significance of linkage is estimated using the likelihood ratio test, comparing the likelihood of a model under the null hypothesis of no linkage between marker and trait loci with an alternative model where the trait gene location (often expressed as a recombination fraction less than 0.5) has been estimated from the data. The logarithm of the likelihood ratio is called the LOD (often misinterpreted as the logarithm of the odds for linkage). The second approach, model-free linkage analysis, does not require the specification of a disease model and is based on a correlation between similarity in marker allele sharing and similarity in phenotype between pairs of relatives, typically sibling pairs.

There are two broad strategies for the inclusion of genetic markers within linkage analysis, applicable to both model-based and model-free approaches. Two-point linkage (also called single-point marker) analysis assesses linkage between the trait locus and each marker individually, while multipoint analysis considers linkage between the trait locus and a number of marker loci jointly. Multipoint linkage analysis is in principle more powerful because it uses more information, but it is also very sensitive to misspecification of the assumed marker locations. Initial two-point analysis has been recommended, but it can lead to a large loss in power if the markers are SNPs because SNPs are less informative for observing recombination events. Another important point is that most multipoint algorithms assume that markers are in linkage equilibrium (see below), which does not hold for currently available SNP maps. Several approaches to this problem have been proposed, such as analyzing haplotype tagging SNPs (16) or simply restricting the analysis to SNPs in linkage equilibrium (17,18); although appealing, care must be taken when using the former approach to make sure there is no residual linkage disequilibrium (LD) between the markers (16). Others have proposed modeling the LD within the linkage analysis (19,20); this methodology is currently still developing.

Model-based linkage analysis is more powerful than model-free linkage analysis when the genetic model is correctly specified. However, for complex diseases that we are interested in today, the presumed model is almost never correct. Nevertheless, model-based analysis is often validity robust, that is, not prone to increased type 1 error. Linkage analysis may incorporate genetic heterogeneity by assuming, at any genome location, that some families are linked and others are not linked. When performed in the model-based framework, this heterogeneity analysis has been termed the HLOD approach (15). A strength of model-based analysis is its ability to estimate both segregation and linkage parameters; this may gain popularity as computational algorithms develop, provided there is a clearly defined ascertainment procedure.

Model-free linkage analysis approaches have the advantage of not requiring the specification of a disease model. The original model-free linkage method for quantitative traits that used the concept of identity by descent (IBD) was the Haseman–Elston regression model (21). This method simply regressed the squared sibling trait difference on the proportion of marker alleles shared IBD by a full sibling pair. Since then, many extensions have been developed, including pair types other than full sibling pairs, allowance for selected

sampling, age-of-onset models, inclusion of multiple trait loci, and analysis of multiple phenotype measures jointly (22,23). The regression coefficient at the trait locus itself allows one to estimate the component of the trait's variance attributable to segregation at that locus. Other variance component models have been developed for pedigree analysis based on the assumption of multivariate normality. These models have also been extended to the multivariate *t* distribution (24,25). In the simplest situations, the likelihood ratios (reported as LODs) can be converted to p values when the samples are large.

Several model-free linkage analysis strategies have been developed for discrete traits, specifically the analysis of concordantly affected and discordant pairs. Although originally developed for affected sibling pairs (26), they have been extended for all relative pair types through a conditional logistic model (27) on the assumption of Mendelian segregation (28). This conditional logistic model originally required the estimation of two parameters but, by using a 'minmax' model (29), genetic effects can be estimated using only one free parameter (30). The original Haseman–Elston model may also be employed for binary phenotypes because such a trait can be coded as 0 or 1, thus putting it on a quantitative scale (31). Again, these models have the advantage of not requiring knowledge of the mode of inheritance, so they can be more appropriate than model-based strategies for complex traits. These methods are ideal when the phenotype of interest is truly a binary clinical outcome. For continuous variables, it is more advantageous to use the full continuous scale of the phenotype, as this added information increases the power to detect linkage (32,33).

Association with a genetic variant

Once the approximate location of a disease gene has been found, association studies can be conducted in that region alone to pinpoint the gene and allelic variants involved, as has been performed in mapping the ΔF_{508} deletion associated with cystic fibrosis (34). Genetic association analysis of candidate gene regions without any preceding linkage analysis has a long history of discovering single marker allele associations, for example type 1 diabetes and the human leukocyte antigen (HLA) DR3/DR4 alleles (35). Originally, studies focused simply on known polymorphisms and then, until this decade, on polymorphisms of specific candidate genes hypothesized to functional effects.

Now, large-scale association analyses are possible. These studies take advantage of LD, that is, the fact that throughout the genome alleles at tightly linked loci are associated in the population. The markers analyzed need not be functional, but may simply be in LD with the functional variant. However, the large-scale nature of these studies introduces a multiple testing issue. Specifically, as the number of markers analyzed increases, the probability of detecting a statistically significant association by chance alone also increases. In genome-wide linkage analysis, criteria have been established for declaring statistical significance (36). With these new GWAS, we still need appropriate criteria to protect against type 1 error.

Most association studies have been conducted using case–control data. Case–control studies are considered to be easier to conduct because the entire family need not be identified and enrolled, whereas population controls are easier to identify. However, it is well known that case–control data are not validity robust to population stratification and heterogeneity; even studies within the UK population have shown genetic differences between population subgroups (37). Several strategies have been developed to guard against these problems. Population stratification may be accounted for by the incorporation of genomic control markers (38,39) or by performing a stratified analysis of genetically homogeneous subsets of the data, also called structured association analysis (40,41). However, family studies that include parental genotype data can circumvent these issues of population stratification.

Association analysis with family data has often been performed with the transmission–disequilibrium test (TDT) (42,43). This method was proposed to test linkage in the presence of association in trio (two parents and a child) data and is also valid *for this purpose* when there is more than one offspring. The method has since been extended to test association in sibship, family, and pedigree data by appropriately allowing for familial correlations, and is then validity robust to population stratification and heterogeneity. Since the development of the TDT, several other family-based association models have been developed (44–47). These methods have the added flexibility of including information from all family members, regardless of their positions in the pedigree, but may or may not be valid in the presence of population stratification or general familial correlations.

Genome-wide association: present and future

The ability to conduct GWAS has advanced rapidly with the advent of high-throughput genotyping technology. A few years ago, SNP genotyping platforms could only produce ~10 000 genotypes per individual; that number has increased 100-fold and platforms can now genotype 1 million SNPs per individual. With the advancement of statistical methodology, there are many test statistics available, so many different types of samples (population-based and/or family-based) and different types of traits may be analyzed. The current SNP maps have roughly 85% coverage of the genome in Caucasians, so these approaches have the potential to be very powerful.

However, we may be seeing an embarrassment of riches. Multiple testing is still a major cause for concern; this is an active area of research. The basis of GWAS and the International HapMap project (48) is the common disease common variant hypothesis. The HapMap project identified SNPs in population samples. Because of the restricted number of individuals sequenced (less than 100 per population group), mostly common (frequency greater than 5%) variants would be detected. Thus, rare variants like the NOD2 risk alleles for Crohn's disease (49) may be hard to find in GWAS. Another challenge is that complex diseases are likely influenced by multiple loci, whether unique loci segregate in different populations (heterogeneity) or multiple loci interact to influence disease risk (epistasis). We must still question the reliability of software, both in terms of calling genotypes from the assay and in the validity of the statistical analysis itself. The debate continues regarding the best method of association analysis. How best to choose SNPs? Should haplotype tagging SNPs be used, or SNPs that are evenly spaced throughout the genome, or perhaps only SNPs that may have functional consequence? Alternatively, should haplotypes be analyzed?

Given how far we have advanced with these SNP technologies, should GWAS be preceded by linkage analysis? Linkage analysis may more efficiently detect rare variants. This idea is well illustrated by Figure 2 in Ardlie et al. (50); rare disease alleles with moderate to strong effects are best detected by linkage analysis, while common alleles with possibly weaker effects are best detected by association analysis. However, genes influencing rare diseases will be hard to identify, regardless of the underlying allele frequency, except by linkage analysis. But how do we perform multipoint linkage analysis when markers are in LD? Again, some models have been proposed to answer this question, but this is still an area of active research.

The future may present even more challenges and questions. Do we study genetics, that is transmission from *gener*ation to *gene*ration, or DNA – the molecule that mediates how what is transmitted eventually results in the phenotype? Parent-of-origin and imprinting effects are known to influence a variety of complex traits, but often parental genotypes are unavailable for analysis; thus, new methods using sibling and other data warrant development (51). Recent advances indicate that copy number variants may have an impact

on disease (52). Analysis of gene expression level may elucidate the action of DNA on disease, suggesting that gene regulation may be more complicated than originally thought (53). As our knowledge of the human genome increases, our definition of 'complex disease' grows increasingly more complex, and the genetics of disease may involve much more than simply a single gene variant producing a single rogue protein to produce disease risk. In most cases, the best strategy to identify disease genes may be to take multiple strategies. Candidate gene studies should be motivated by pathway models related to disease pathogenesis. Linkage analysis can offer an affordable first-pass to identify regions of the genome where rare susceptibility variants are located. GWAS may help elucidate those elusive variants with smaller effects (54), but samples must be very large to be powered enough to attain the significance levels that allow for the multiple testing. In the future, in addition to the availability of 1 million SNP variants being assayed on a chip, full sequence variation will also be available. The 1000 genomes project (www.1000genomes.org) (55) seeks to fully sequence samples from 1000 individuals across the globe, with the goal of cataloguing variants that occur with 1% frequency or less. This vast database will better enable the identification of elusive rare disease-causing variants. But the use of such data will require the development of appropriate statistical techniques, or we may continue to face an embarrassment of riches.

Concluding remarks

A well-designed optimal two-stage association study (56), possibly pooling DNA at stage 1, is now possible for GWAS. Three-stage association designs are also promising, but need further study. Replication and resequencing are important sampling design stages and essential to establish the involvement of a given gene in disease pathogenesis; these stages are not trivial. Collecting a sample of families for a linkage study before conducting a genome-wide association study will still be useful (57).

Acknowledgments

This work was supported in part by grants from the US Public Health Service: National Center for Research Resources Resource Grant RR03655 and Multidisciplinary Clinical Research Career Development Programs Grant KL2RR024990; National Institute of General Medical Sciences Research Grant GM-28356; and National Cancer Institute Cancer Center Support Grant P30CAD43703.

References

1. Olson JM, Cordell HJ. Ascertainment bias in the estimation of sibling genetic risk parameters. Genet Epidemiol 2000;18:217–235. [PubMed: 10723107]
2. Risch N. Linkage strategies for genetically complex traits. I. Multilocus models. Am J Hum Genet 1990;46:222–228. [PubMed: 2301392]
3. Falconer, D.; Mackay, T. Introduction to quantitative genetics. 4. Harlow, England: Prentice Hall; 1996.
4. Hartl DL, Maruyama T. Phenogram enumeration: the number of regular genotype-phenotype correspondences in genetic systems. J Theor Biol 1968;20:129–163. [PubMed: 5727236]
5. Shete S, Elston RC, Lu Y. A novel approach to detect parent-of-origin effects from pedigree data with application to Beckwith-Wiedemann syndrome. Ann Hum Genet 2007;71:804–814. [PubMed: 17578507]
6. Botstein D, White RL, Skolnick M, Davis RW. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am J Hum Genet 1980;32:314–331. [PubMed: 6247908]
7. Litt M, Luty JA. A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene. Am J Hum Genet 1989;44:397–401. [PubMed: 2563634]

8. Weber JL, May PE. Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. Am J Hum Genet 1989;44:388–396. [PubMed: 2916582]
9. Hall JM, Friedman L, Guenther C, et al. Closing in on a breast cancer gene on chromosome 17q. Am J Hum Genet 1992;50:1235–1242. [PubMed: 1598904]
10. Hall JM, Lee MK, Newman B, et al. Linkage of early-onset familial breast cancer to chromosome 17q21. Science 1990;250:1684–1689. [PubMed: 2270482]
11. Wooster R, Bignell G, Lancaster J, et al. Identification of the breast cancer susceptibility gene BRCA2. Nature 1995;378:789–792. [PubMed: 8524414]
12. Ginsberg, E.; Malkin, I.; Elston, RC. Theoretical aspects of pedigree analysis. Tel Aviv: Ramot Publishing House; 2006.
13. Ginsburg E, Malkin I, Elston RC. Sampling correction in pedigree analysis. Stat Appl Genet Mol Biol 2003;2 Article 2. Epub 18 March 2003.
14. Vieland VJ, Hodge SE. The problem of ascertainment for linkage analysis. Am J Hum Genet 1996;58:1072–1084. [PubMed: 8651268]
15. Ott, J. Analysis of human genetic linkage. 3. Baltimore: Johns Hopkins Press; 1999.
16. Duggal P, Gillanders EM, Mathias RA, et al. Identification of tag single-nucleotide polymorphisms in regions with varying linkage disequilibrium. BMC Genet 2005;6 (Suppl 1):S73. [PubMed: 16451687]
17. Bacanu SA. Multipoint linkage analysis for a very dense set of markers. Genet Epidemiol 2005;29:195–203. [PubMed: 16094613]
18. Goode EL, Jarvik GP. Assessment and implications of linkage disequilibrium in genome-wide single-nucleotide polymorphism and microsatellite panels. Genet Epidemiol 2005;29 (Suppl 1):S72–S76. [PubMed: 16342185]
19. Abecasis GR, Wigginton JE. Handling marker-marker linkage disequilibrium: pedigree analysis with clustered markers. Am J Hum Genet 2005;77:754–767. [PubMed: 16252236]
20. Kurbasic A, Hossjer O. A general method for linkage disequilibrium correction for multipoint linkage and association. Genet Epidemiol. 2008 (in press).
21. Haseman J, Elston R. The investigation of linkage between a quantitative trait and a marker locus. Behav Genet 1972;2:3–19. [PubMed: 4157472]
22. Wang T, Elston RC. Two-level Haseman-Elston regression for general pedigree data analysis. Genet Epidemiol 2005;29:12–22. [PubMed: 15838848]
23. Wang T, Elston RC. Regression-based multivariate linkage analysis with an application to blood pressure and body mass index. Ann Hum Genet 2007;71:96–106. [PubMed: 17227480]
24. Almasy L, Blangero J. Multipoint quantitative-trait linkage analysis in general pedigrees. Am J Hum Genet 1998;62:1198–1211. [PubMed: 9545414]
25. Amos CI, Elston RC. Robust methods for the detection of genetic linkage for quantitative data from pedigrees. Genet Epidemiol 1989;6:349–360. [PubMed: 2721929]
26. Risch N. Linkage strategies for genetically complex traits. II. The power of affected relative pairs. Am J Hum Genet 1990;46:229. [PubMed: 2301393]
27. Olson J. A general conditional-logistic model for affected-relative-pair linkage studies. Am J Hum Genet 1999;65:1760–1769. [PubMed: 10577930]
28. Holmans P. Asymptotic properties of affected-sib-pair linkage. Am J Hum Genet 1993;52:362–374. [PubMed: 8430697]
29. Whittemore AS, Tu I-P. Simple, robust linkage tests for affected sibs. Am J Hum Genet 1998;62:1228–1242. [PubMed: 9599188]
30. Goddard K, Witte J, Suarez B, Catalona W, Olson J. Model-free linkage analysis with covariates confirms linkage of prostate cancer to chromosomes 1 and 4. Am J Hum Genet 2001;68:1197–1206. [PubMed: 11309685]
31. Elston RC, Song D, Iyengar SK. Mathematical assumptions versus biological reality: myths in affected sib pair linkage analysis. Am J Hum Genet 2005;76:152–156. [PubMed: 15540158]
32. Duggirala R, Williams J, Williams-Blangero S, Blangero J. A variance components approach to dichotomous trait linkage using a threshold model. Genet Epidemiol 1997;14:987–992. [PubMed: 9433612]

33. Wijsman EM, Amos CI. Genetic analysis of simulated oligogenic traits in nuclear and extended pedigrees: summary of GAW10 contributions. Genet Epidemiol 1997;14:719–735. [PubMed: 9433569]
34. Kerem B, Rommens JM, Buchanan JA, et al. Identification of the cystic fibrosis gene: genetic analysis. Science 1989;245:1073–1080. [PubMed: 2570460]
35. Svejgaard, A.; Ryder, LP. Association between HLA and disease. In: Dausset, J.; Svejgaard, A., editors. HLA and disease. Copenhagen: Munksgaard; 1977. p. 46-71.
36. Lander E, Kruglyak L. Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. Nat Genet 1995;11:241–247. [PubMed: 7581446]
37. 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. [PubMed: 17554300]
38. Devlin B, Roeder K. Genomic control for association studies. Biometrics 1999;55:997–1004. [PubMed: 11315092]
39. Zhu X, Li S, Cooper RS, Elston RC. A unified association analysis approach for family and unrelated samples correcting for stratification. Am J Hum Genet 2008;82:352–365. [PubMed: 18252216]
40. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics 2000;155:945–959. [PubMed: 10835412]
41. Satten GA, Flanders WD, Yang Q. Accounting for unmeasured population substructure in case-control studies of genetic association using a novel latent-class model. Am J Hum Genet 2001;68:466–477. [PubMed: 11170894]
42. Ewens WJ, Spielman RS. What is the significance of a significant TDT? Hum Hered 2005;60:206–210. [PubMed: 16391488]
43. Spielman RS, McGinnis RE, Ewens WJ. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet 1993;52:506–516. [PubMed: 8447318]
44. Abecasis GR, Cookson WOC, Cardon LR. Pedigree tests of transmission disequilibrium. Eur J Hum Genet 2000;8:545–551. [PubMed: 10909856]
45. George V, Tiwari HK, Zhu X, Elston RC. A test of transmission/disequilibrium for quantitative traits in pedigree data, by multiple regression. Am J Hum Genet 1999;65:236–245. [PubMed: 10364537]
46. George V, Elston RC. Testing the association between polymorphic markers and quantitative traits in pedigrees. Genet Epidemiol 1987;4:193–202. [PubMed: 3609719]
47. Lange C, Demeo DL, Laird NM. Power and design considerations for a general class of family-based association tests: quantitative traits. Am J Hum Genet 2002;71:1330–1341. [PubMed: 12454799]
48. The International HapMap Consortium. The international HapMap project. Nature 2003;426:789–796. [PubMed: 14685227]
49. Lesage S, Zouali H, Cezard JP, et al. CARD15/NOD2 mutational analysis and genotype-phenotype correlation in 612 patients with inflammatory bowel disease. Am J Hum Genet 2002;70:845–857. [PubMed: 11875755]
50. Ardlie KG, Kruglyak L, Seielstad M. Patterns of linkage disequilibrium in the human genome. Nat Rev Genet 2002;3:299–309. [PubMed: 11967554]
51. Rampersaud E, Morris RW, Weinberg CR, Speer MC, Martin ER. Power calculations for likelihood ratio tests for offspring genotype risks, maternal effects, and parent-of-origin (POO) effects in the presence of missing parental genotypes when unaffected siblings are available. Genet Epidemiol 2007;31:18–30. [PubMed: 17096358]
52. McCarroll SA, Altshuler DM. Copy-number variation and association studies of human disease. Nat Genet 2007;39:S37–S42. [PubMed: 17597780]
53. Morley M, Molony CM, Weber TM, et al. Genetic analysis of genome-wide variation in human gene expression. Nature 2004;430:743–747. [PubMed: 15269782]
54. Risch N, Merikangas K. The future of genetic studies of complex human diseases. Science 1996;273:1516–1517. [PubMed: 8801636]

55. Siva N. 1000 genomes project. Nat Biotechnol 2008;26:256. [PubMed: 18327223]

56. Elston RC, Lin D, Zheng G. Multi-stage sampling for genetic studies. Annu Rev Genomics Hum Genet 2007;8:327–342. [PubMed: 17506660]

57. Clerget-Darpoux F, Elston RC. Are linkage analysis and the collection of family data dead? Prospects for family studies in the age of genome-wide association. Hum Hered 2007;64:91–96. [PubMed: 17476108]