Published in final edited form as:

Am J Ophthalmol. 2009 August ; 148(2): 183–185. doi:10.1016/j.ajo.2009.02.004.

Statistical Genetic Approaches for Mapping Ophthalmic Trait and Disease Genes

Janet Sinsheimer

Departments of Human Genetics and Biomathematics David Geffen School of Medicine at UCLA
Department of Biostatistics, UCLA School of Public Health, Los Angeles, CA 90095

Keywords

linkage analysis; association; integrative genetics; eye disease; genome wide association study; GWAS; gene network

Some of the finest examples of successful gene mapping have been for ophthalmic traits and diseases^{1,2,3}. In this editorial, I briefly describe statistical approaches used in gene mapping of human ophthalmic traits and diseases and discuss some challenges and promising new approaches. A number of fine statistical genetic software programs are available. Because I use the statistical genetics package Mendel⁴ (www.genetics.ucla.edu/software), I provide the names of the appropriate Mendel options to perform specific gene mapping tests. Readers can get a comprehensive and annotated list of other statistical genetic programs from the Robert S. Boas Center for Genomics and Human Genetics (www.nslj-genetics.org/soft/).

Gene mapping methods fall into two broad statistical approaches, linkage analysis and association. Although they are often considered separately, linkage analysis and association are connected and there are sophisticated statistical methods that allow for joint estimation (see as examples, Pseudomarker⁵ or Mendel's Association_Given_Linkage option⁶). Linkage analysis estimates trait/disease gene positions relative to genes of known location (markers) by using the degree of correlation between trait/disease phenotype inheritance and marker genotype inheritance (co-segregation). When a marker is close to a polymorphism in an autosomal gene that confers risk of a trait/disease, the offspring's chromosomal segments containing both the trait/disease and marker genes will be the same as those on his/her parental chromosomes; *e.g.* they will be inherited without any recombination between the two loci. When the marker is far away from the disease/trait gene (or on another chromosome) then the chance of recombination is 50%.

Linkage analysis is very effective when a single gene is both necessary and sufficient to change the trait value or cause disease (Mendelian trait/disease). In particular, genetic model-based linkage analysis, which specifies the number of genes, their mode of inheritance, the probability of the phenotype given the disease/trait genotype (the penetrance), and estimates genetic distances or recombination fractions, has been used very successfully to map genes for Mendelian diseases/traits^{2,3} in ophthalmic research. The test statistic is often the log-likelihood

© 2009 Elsevier Inc. All rights reserved

janet@mednet.ucla.edu 310-825-8002.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

assuming linkage minus the log-likelihood assuming no linkage. Examples of model-based linkage analysis software are Merlin⁷ and Mendel's Location_Scores option⁴.

Multifactorial or complex traits/diseases, like glaucoma and age related macular degeneration (AMD), are the consequence of multiple, possibly interacting, genetic and environmental factors^{1,2}. Genes that underlie these complex traits/diseases are difficult to map because accurately specifying the genetic model can be almost impossible. A popular approach when studying complex traits/diseases are genetic "model-free" linkage analyses. (*N.B.* These methods are not completely genetic model-free, they make implicit genetic assumptions⁵). Model-free methods look for increased gene sharing (identity by descent) between relatives with similar phenotypes and, sometimes, decreased identity by descent between relatives with dissimilar phenotypes. Wald tests comparing observed and expected identity by descent are popular for affected relative data; variance-component analyses are often used with continuous traits, and programs that implement these methods include Mendel's NPL⁸ and Polygenic_QTL⁹ options, Merlin⁷, SOLAR¹⁰ and Genehunter¹¹.

Linkage analysis relies on the observed recombination between markers and trait/disease genes in families to define the likely regions where the trait/disease genes reside. Therefore, in practical terms, linkage analysis has a resolution limit¹². Association analysis provides finer resolution. Association analysis is based on linkage disequilibrium (LD) where the allele frequencies of two closely situated genes will be highly correlated. If the marker and trait/disease genes are not coincident, then over many generations there will be recombination events that eliminate the correlation and result in linkage equilibrium. LD decreases as (1) the time since the introduction of the polymorphisms into the population increases, (2) the distance between the marker and trait/disease loci increases and (3) the minor allele frequencies decrease. Thus association analysis is most powerful when the trait/disease conferring polymorphism was recently introduced, the risk allele is relatively common and the chromosomal region is densely covered by the markers.

Association analysis can be conducted with families or unrelated individuals. Cases (affected individuals) and controls (randomly selected or unaffected individuals) provide a particularly simple study design. The underlying assumption of this analysis is that cases have the same risk conferring alleles at the trait/disease genes (common disease – common variant hypothesis, CDCV). Because the risk conferring genes are in LD with nearby markers, cases have marker alleles in common. So when a marker is close to the trait/disease gene, cases have different marker genotype frequencies than the controls. Thus simple tests of association are contingency table analyses or likelihood ratio tests that compare the marker genotype frequencies of cases and controls. Examples of statistical packages are Mendel's Allele_Frequencies and Cases_And_Controls options¹³ and PLINK¹⁴.

Because the markers need to be quite close to the trait/disease genes in order to have sufficient power to detect LD, association studies were once limited to refining the chromosomal regions first found through linkage analysis or to a small number of candidate genes. In general, these early studies were unsuccessful. However, with the current ability to genotype 100,000 to 1,000,000 single nucleotide polymorphisms (SNPs) in thousands of individuals, genome wide association studies (GWAS) have replaced linkage analysis as the preferred gene mapping approach. The first successful GWAS is the association of a common variant of the complement factor H gene (CFH) with AMD¹⁵. Although it is tempting to conclude that many genes for ophthalmic traits/diseases can as easily be mapped with GWAS designs, it is important to remember that the CFH-AMD association represents a particularly strong effect that was also found using the traditional approach of linkage analysis followed by association fine mapping^{16,17}. It is also important to remember that not all trait/disease genes will conform to

the CDCV hypothesis. A number of rare variants on different chromosomal backgrounds will exist in those genes, making their detection in a GWAS highly unlikely¹⁸.

Currently most association studies start by testing each SNP separately. However, the products of the approximately 25,000 genes in the human genome must interact and so this one gene at a time approach to gene mapping may fail. The development of efficient and powerful statistical methods to uncover gene networks that determine clinical trait values is active area of research^{19,20}. A key assumption of these integrative genetic approaches is that the polymorphisms that ultimately affect clinically observable traits act by perturbing molecular networks. One approach is to construct gene co-expression network modules and determine which, if any, of these modules are correlated with the clinical traits^{19,20}.

The massive amount of data now available present computational and statistical challenges that are active research areas^{14,21,22}. Storage and manipulation of so much data is cumbersome and methods have recently been developed to compress datasets and to efficiently extract relevant subsets¹⁴. Implementing diagnostic procedures and interpreting the results also present challenges^{14,21}. Association testing with 100,000+ SNPs and multiple trait/disease phenotypes leads to a serious problem of how to limit the number of false positive results without substantially raising the false negative rate. Sequence data will only increase the problem²¹. The massive amount of data also provides opportunities, however. As an example, association studies can lead to incorrect results if there is population stratification but with large amounts of SNP data ancestry can be accurately inferred and researchers can control for population stratification or exploit it in gene mapping²³.

Statistical genetics plays an important role in ophthalmic research and will continue to play an important role as the amount of relevant data increases. I close with a few comments for those researchers wanting a few simple rules of thumb to determine the optimal statistical approach. Unfortunately, no one study design or statistical approach to gene mapping will be optimal for every ophthalmic trait/disease. The best approach will depend on a number of factors including the prevalence of the trait/disease, the age of onset, the underlying genetic and environmental determinants, resources, and the means of recruiting individuals or families. I highly recommend that, early in the study design phase, researchers collaborate with a statistician to determine the approach most likely to succeed in their case.

Acknowledgements

The author is supported in part by United States Public Health Service grants MH59490 and GM53275. I thank Prof. Paivi Pajukanta and an anonymous reviewer for their comments on an early draft of this manuscript.

References

1. Swaroop A, Branham KEH, Chen W, Abecasis G. Genetic susceptibility to age-related macular degeneration: a paradigm for dissecting complex disease traits. *Hum Mole Genet* 2007;16:R174–R182. (Review 2)
2. Iyengar SK. The quest for genes causing complex traits in ocular medicine. *Arch. Ophthalmol* 2007;125:11–18. [PubMed: 17210847]
3. Goodwin P. Hereditary retinal disease. *Curr. Opin. Ophthalmol* 2008;19:255–262. [PubMed: 18408503]
4. Lange K, Cantor R, Horvath S, Perola M, Sabatti C, Sinsheimer J, Sobel E. Mendel version 4.0: A complete package for the exact genetic analysis of discrete traits in pedigree and population data sets. *Amer J Hum Genetics* 2001;69(supplement):504. [PubMed: 11462172]
5. Göring HHH, Terwilliger JD. Linkage Analysis in the Presence of Errors IV: Joint Pseudomarker Analysis of Linkage and/or Linkage Disequilibrium on a Mixture of Pedigrees and Singletons When

the Mode of Inheritance Cannot Be Accurately Specified. *Amer J Hum Genet* 2000;66:1310–1327. [PubMed: 10731466]

6. Cantor RM, Chen GK, Pajukanta P, Lange K. Association testing in a linked region using large pedigrees. *Amer J Hum Genet* 2005;76:538–542. [PubMed: 15657872]
7. Abecasis GR, Cherny SS, Cookson WO, Cardon LR. “Merlin-rapid analysis of dense genetic maps using sparse gene flow trees”. *Nature Genetics* 2002;30:97–101. [PubMed: 11731797]
8. Lange EM, Lange K. Powerful allele sharing statistics for nonparametric linkage analysis. *Hum Hered* 2004;57:49–58. [PubMed: 15133312]
9. Bauman L, Almasy L, Blangero J, Duggirala R, Sinsheimer JS, Lange K. Fishing for pleiotropic QTLs in a polygenic sea. *Ann Hum Genet* 2005;69:590–611. [PubMed: 16138917]
10. Almasy L, Blangero J. Multipoint quantitative-trait linkage analysis in general pedigrees. *Amer J Hum Genet* 1998;62:1198–1211. [PubMed: 9545414]
11. Kruglyak L, Daly MJ, ReeveDaly MP, Lander ES. Parametric and nonparametric linkage analysis: a unified multipoint approach. *Amer J Hum Genet* 1996;58:1347–1363. [PubMed: 8651312]
12. Boehnke B. Limits of resolution of genetic linkage studies: implications for the positional cloning of human disease genes. *Amer J Hum Genet* 1994;55:379–390. [PubMed: 8037215]
13. Lange K, Sinsheimer JS, Sobel E. Association testing with Mendel. *Genet Epidemiol* 2005;29:36–50.
14. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC. PLINK: a toolset for whole-genome association and population-based linkage analysis. *Amer J Hum Genet* 2007;81:559–575. [PubMed: 17701901]
15. Klein RJ, Zeiss C, Chew EY, Tsai J-Y, Sackler RS, Haynes C, Henning AK, SanGiovanni JP, Mane SM, Mayne ST, Bracken MB, Ferris FL, Ott J, Barnstable C, Hoh J. Complement Factor H polymorphism in age-related macular degeneration. *Science* 2005;308:385–389. [PubMed: 15761122]
16. Haines JL, Hauser MA, Schmidt S, Scott WK, Olson LM, Gallins P, Spencer KL, Kwan SY, Noureddine M, Gilbert JR, Schnetz-Boutaud N, Agarwal A, Postel EA, Pericak-Vance MA. Complement Factor H variant increases the risk of age-related macular degeneration. *Science* 2005;308:419–412. [PubMed: 15761120]
17. 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]
18. Weiss KM, Terwilliger JD. How many diseases does it take to map a gene with SNPs? *Nat Genet* 2000;26:151–157. [PubMed: 11017069]
19. Lusis AJ, Attie AD, Reue K. Metabolic syndrome: from epidemiology to systems biology. *Nat Rev Genet* 2008;9:819–830. [PubMed: 18852695]
20. Chen Y, Zhu J, Lum PY, Yang X, Pinto S, MacNeil DJ, Zhang C, Lamb J, Edwards S, Sieberts SK, Leonardson A, Castellini LW, Wang S, Champy M-F, Zhang B, Emilsson V, Doss S, Ghazalpour A, Horvath S, Drake TA, Lusis AJ, Schadt EE. Variation in DNA elucidate molecular networks that cause disease. *Nature* 2008;452:429–435. [PubMed: 18344982]
21. Wu TT, Lange K. Coordinate descent algorithms for lasso penalized regression. *Annals Appl. Stat* 2008;2:224–244.
22. Elston RC, Spence MA. Advances in statistical human genetics over the last 25 years. *Stat. Med* 2006;25:3049–80. [PubMed: 16900576]
23. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet* 2006;38:904–909. [PubMed: 16862161]