6.047/6.878 Lecture 24: Medical Genetics – The Past to the Present

Guest Lecture by
Mark J. Daly (PhD)
Scribed by Anna Ayuso, Abhishek Sarkar (2011), Joel Brooks (2012)

December 13, 2012

Contents

List of Figures

1 Introduction

Mark J. Daly, Ph.D., is an Associate Professor at the Massachusetts General Hospital/Harvard Medical School and an Associate Member of the Broad Institute. This lecture explains how statistical and computational methods can aid researchers in understanding, diagnosing, and treating disease. The problem of identifying genetic variation which can explain phenotypic variation is known as association mapping. This problem is particularly important to solve for disease phenotypes (e.g., susceptibility). Historically, the method of choice for solving this problem was linkage analysis. However, advances in genotyping technology have allowed for a more powerful method called genome-wide association. More recent advances in genomic data have allowed for novel integrative analyses which can make more powerful predictions about diseases.

2 Goals of investigating the genetic basis of disease

Any discussion about the basis of disease must consider both genetic and environmental effects. However, it is known that many traits, for example those in Figure 1, have significant genetic components. Formally, the *heritability* of a phenotype is the proportion of variation in that phenotype which can be explained by genetic variation. The traits in Figure 1 are all at least 50% heritable.

Accurately estimating heritability involves statistical analyses on samples with highly varied levels of shared genetic variation (e.g., twins, siblings, relatives, and unrelated). Studies on the heritability of Type 2 diabetes, for example, have shown that given you have diabetes, the risk to the person sitting next to you (an unrelated person) increases by 5–10%; the risk to a sibling increases by 30%; and the risk to an identical twin increases by 85%–90%.

Having established that there is a genetic component to disease traits of interest, what are the goals of understanding this component? There are three main goals:

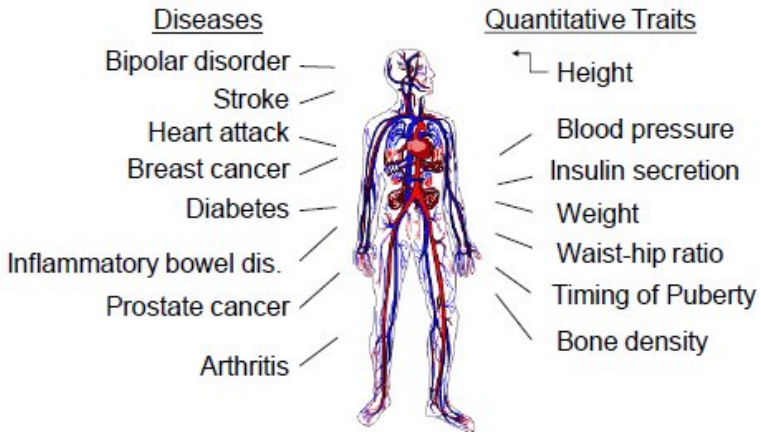

Figure 1: Examples of diseases and quantitative traits which have genetic components

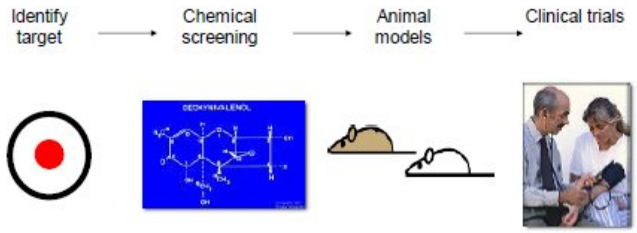

Figure 2: The drug development process

- Directing downstream research in disease
- Potential for improved diagnostics
- Enabling rational drug development

Identifying genetic variants which explain variation in the disease trait obviously contributes to our ability to understand the mechanism (the biochemical pathways, etc.) by which the disease manifests. Moreover, those variants can be used in genetic screens to test for increased risk for the disease trait. But the last goal is of particular interest because strong evidence suggests we do not really know how to develop effective drugs to target particular diseases. For example, in the last 50 years, no truly novel compounds have been developed to treat various psychiatric disorders such as schizophrenia.

Figure 2 depicts the cycle of drug development. First, researchers hypothesize a possible target of interest that might be related to a disease. They evaluate the biochemistry of this target, test the target in model organisms, and then finally perform clinical trials in humans. However, the vast majority of drugs which make it through this process end up being ineffective in treating the disease for which they were originally designed. This result is mainly a consequence of poor hypotheses about the basis of the disease in question.

Statins are a prominent example of highly effective drugs developed after work on understanding the genetic basis of the disease trait they are targeted at. Dr. Michael Brown and Dr. Joseph Goldstein won the Nobel Prize in Physiology or Medicine in 1985 for their work on the regulation of LDL cholesterol metabolism [5]. They were able to isolate the cause of extreme familial hypercholesterolemia (FH), a Mendelian disorder, to mutations of a single gene encoding an LDL receptor. Moreover, they were able to identify the biochemical pathway which was affected by the mutation to create the disease condition. Statins target that pathway, making them useful not only to individuals suffering from FH, but also as an effective treatment for high LDL cholesterol in the general population.

3 Linkage Analysis

Historically, researchers have used **linkage analysis** to determine genetic variants which explain phenotypic variation. The goal is to determine which variants contribute to the observed pattern of phenotypic variation in a *pedigree*. Figure 3 shows an example pedigree in which squares are male individuals, circles are female individuals, couples and offspring are connected, and individuals in red have the trait of interest.

Linkage analysis relies on the biological insight that genetic variants are not independently inherited (as proposed by Mendel). Instead, meiotic recombination happens a limited number of times (roughly once per chromosome), so many variants *cosegregate* (are inherited together). This phenomenon is known as *linkage disequilibrium* (LD).

As the distance between two variants increases, the probability a recombination occurs between them increases. Thomas Hunt Morgan and Alfred Sturtevant developed this idea to produce **linkage maps** which could not only determine the order of genes on a chromosome, but also their relative distances to each other. The Morgan is the unit of genetic distance they proposed; loci separated by 1 centimorgan (cM) have 1 in 100 chance of being separated by a recombination. Unlinked loci have 50% chance of being separated by a recombination (they are separated if an odd number of recombinations happens between them). Since we

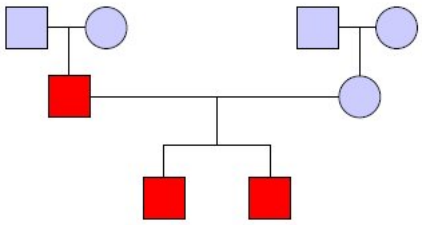

Figure 3: A pedigree which shows the inheritance of some trait

usually do not know *a priori* which variants are causal, we instead use *genetic markers* which capture other variants due to LD. In 1980, David Botstein proposed using *single nucleotide polymorphisms* (SNPs), or mutations of a single base, as genetic markers in humans [4]. If a particular marker is in LD with the actual causal variant, then we will observe its pattern of inheritance contributing to the phenotypic variation in the pedigree and can narrow down our search.

The statistical foundations of linkage analysis were developed in the first part of the 20th century. Ronald Fisher proposed a genetic model which could reconcile Mendelian inheritance with continuous phenotypes such as height [8]. Newton Morton developed a statistical test called the **LOD score** (logarithm of odds) to test the hypothesis that the observed data results from linkage [23]. The null hypothesis of the test is that the *recombination fraction* (the probability a recombination occurs between two adjacent markers) $\theta = 1/2$ (no linkage) while the alternative hypothesis is that it is some smaller quantity. The LOD score is essentially a log-likelihood ratio which captures this statistical test:

$$\text{LOD} = \frac{\log(\text{likelihood of disease given linkage})}{\log(\text{likelihood of disease given no linkage})}$$

The algorithms for linkage analysis were developed in the latter part of the 20th century. There are two main classes of linkage analysis: *parametric* and *nonparametric* [29]. Parametric linkage analysis relies on a model (parameters) of the inheritance, frequencies, and penetrance of a particular variant. Let F be the set of founders (original ancestors) in the pedigree, let g_i be the genotype of individual i, let Φ_i be the phenotype of individual i, and let $f(i)$ and $m(i)$ be the father and mother of individual i. Then, the likelihood of observing the genotypes and phenotypes in the pedigree is:

$$L = \sum_{g_1} \cdots \sum_{g_n} \prod_i \Pr(\Phi_i | g_i) \prod_{f \in F} \Pr(g_f) \prod_{i \notin F} \Pr(g_i | g_{f(i)}, g_{m(i)})$$

The time required to compute this likelihood is exponential in both the number of markers being considered and the number of individuals in the pedigree. However, Elston and Stewart gave an algorithm for more efficiently computing it assuming no inbreeding in the pedigree [7]. Their insight was that conditioned on parental genotypes, offspring are conditionally independent. In other words, we can treat the pedigree as a Bayesian network to more efficiently compute the joint probability distribution. Their algorithm scales linearly in the size of the pedigree, but exponentially in the number of markers.

There are several issues with parametric linkage analysis. First, individual markers may not be *informative* (give unambiguous information about inheritance). For example, homozygous parents or genotyping error could lead to uninformative markers. To get around this, we could type more markers, but the algorithm does not scale well with the number of markers. Second, coming up with model parameters for a Mendelian disorder is straightforward. However, doing the same for non-Mendelian disorders is non-trivial. Finally, estimates of LD between markers are not inherently supported.

Nonparametric linkage analysis does not require a genetic model. Instead, we first infer the inheritance pattern given the genotypes and the pedigree. We then determine whether the inheritance pattern can explain the phenotypic variation in the pedigree.

Lander and Green formulated an HMM to perform the first part of this analysis [17]. The states of this HMM are *inheritance vectors* which specify the result every meiosis in the pedigree. Each individual is represented by 2 bits (one for each parent). The value of each bit is 0 or 1 depending on which of the

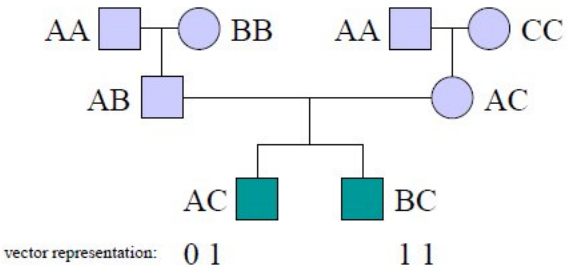

Figure 4: Representing a particular pattern of inheritance as an inheritance vector

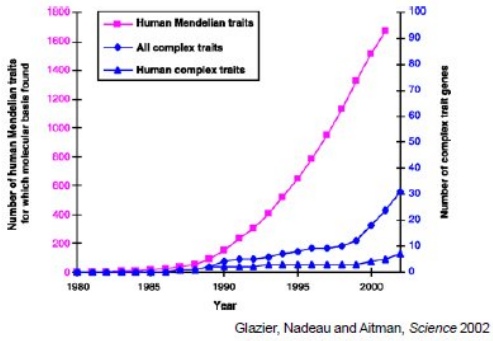

Figure 5: Discovery of genes for different disease types versus time

grand-parental alleles is inherited. Figure 4 shows an example of the representation of two individuals in an inheritance vector.

Each step of the HMM corresponds to a marker; a transition in the HMM corresponds to some bits of the inheritance vector changing. This means the allele inherited from some meiosis changed, i.e. that a recombination occurred. The transition probabilities in the HMM are then a function of the recombination fraction between adjacent markers and the Hamming distance (the number of bits which differ, or the number of recombinations) between the two states. We can use the forward-backward algorithm to compute posterior probabilities on this HMM and infer the probability of every inheritance pattern for every marker.

This algorithm scales linearly in the number of markers, but exponentially in the size of the pedigree. The number of states in the HMM is exponential in the length of the inheritance vector, which is linear in the size of the pedigree. In general, the problem is known to be NP-hard (to the best of our knowledge, we cannot do better than an algorithm which scales exponentially in the input) [25]. However, the problem is important not only in this context, but also in the contexts of *haplotype inference* or *phasing* (assigning alleles to homologous chromosomes) and *genotype imputation* (inferring missing genotypes based on known genotypes). There have been many optimizations to make this analysis more tractable in practice [1, 9, 10, 12–15, 18, 20].

Linkage analysis identifies a broad genomic region which correlates with the trait of interest. To narrow down the region, we can use fine-resolution genetic maps of recombination breakpoints. We can then identify the affected gene and causal mutation by sequencing the region and testing for altered function.

4 Genome-wide Association Studies

Linkage analysis has proven to be highly effective in studying the genetic basis of Mendelian (single gene) diseases. In the past three decades, thousands of genes have been identified as contributing to Mendelian diseases. Figure 5 shows this explosion in published associations. We have identified the genetic basis of disease such as sickle cell anemia, cystic fibrosis, muscular dystrophy, and severe forms of common diseases such as diabetes and hypertension. For these diseases, mutations are severe and obvious; the environment, behavior, and chance have little effect.

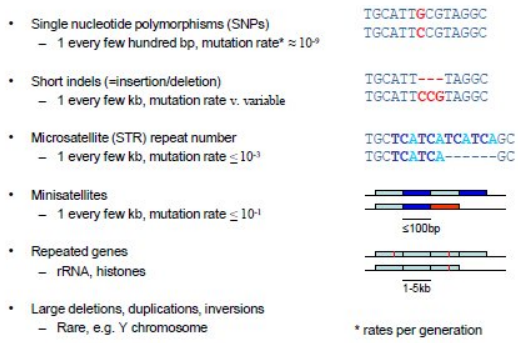

Figure 6: Different types of genetic variation

However, most diseases (and many other traits of interest) are not Mendelian. These **complex traits** arise from the interactions of many genes and possibly the environment and behavior. A canonical complex trait is human height: it is highly heritable, but environmental factors can affect it. Recently, researchers have identified hundreds of variants which are associated with height [2, 22].

Linkage analysis is not a viable approach to find these variants. In the 1990s, researchers proposed a methodology called **genome-wide association** to systematically correlate markers with traits. These studies sample large pools of cases and controls, measure their genotypes at on the order of one million markers, and try to correlate variation in their genotypes with their variation in phenotype.

Genome-wide association studies (GWASs) are possible due to three advances. First, advances in our understanding of the genome and the creation of genomic resources. The key biological insight is the fact that humans are one of the least genetically diverse species. On the order of tens of millions of SNPs are shared between different human subpopulations. For any particular region of the genome, we observe only a limited number of **haplotypes** (allele combinations which are inherited together). Because of this high redundancy, we only need to measure a fraction of all the variants in the human genome in order to capture them all with LD. We can adapt the algorithms for inferring inheritance patterns in linkage analysis to impute genotypes for the markers which we did not genotype.

Genome resources allow us to carefully choose markers to measure and to make predictions based on markers which show statistically significant association. We now have the reference sequence of the human genome (allowing for alignments, genotype and SNP calling) and HapMap, a comprehensive catalog of SNPs in humans. We also have genome-wide annotations of genes and regulatory elements.

Second, advances in genotyping technology such as microarrays and high-throughput sequencing. Although there are many types of variation in the human genome (Figure 6 shows some examples), SNPs are the vast majority. They are also the easiest and cheapest to measure using these technologies. However, we still need to account for the other types of variants. Recently developed DNA microarrays can detect copy-number variation in addition to SNPs.

The third advance is a new expectation of collaboration between researchers. GWASs rely on large sample sizes to increase the *power* (probability of a true positive) of statistical tests. The explosion in the number of published GWASs has allowed for a new type of **meta-analysis** which combines the results of several GWASs for the same phenotype to make more powerful associations. Meta-analysis accounts for various technical and population-genetic biases in individual studies. Researchers who conduct GWASs are expected to collaborate with others who have conducted GWASs on the same trait in order to show replicability of results. By pooling together the data, we also have more confidence in the reported associations.

The main problem in conducting GWASs is eliminating confounding factors. First, genotyping error, which is common enough to require special treatment regardless of which technology we use. To account for such errors, we use thresholds on metrics like minor allele frequency and deviation from Hardy–Weinberg equilibrium and throw out SNPs which do not meet the criteria. Second, systematic genetic differences between human subpopulations. There are several methods to account for this **population substructure** such as genomic control [6], structured association [27], and principal component analysis [24, 26]. Third, covariates such as environmental and behavioral effects. We can account for these by including them in our

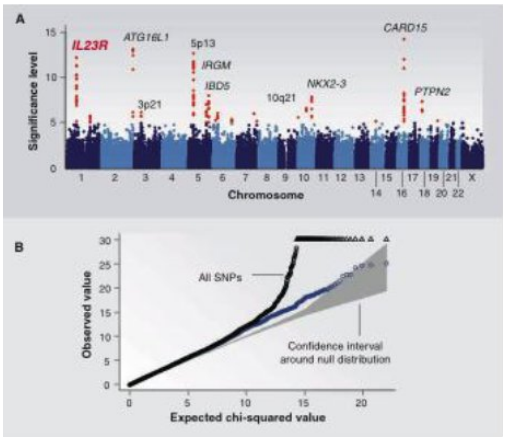

Figure 7: (A) Manhattan plot and (B) Q-Q plot for GWAS of Crohn's disease

statistical model.

The statistical analysis involved in GWAS is fairly straightforward. We assume the effect of each SNP is independent and additive to make the analysis tractable. For each SNP, we perform a hypothesis test whose null hypothesis is that the observed variation in the genotype at that SNP across the subjects does not correlate with the observed variation in the phenotype across the subjects. Because we perform one test for each SNP, we need to deal with the **multiple testing problem**. Each test has some probability of giving a false positive result, and as we increase the number of tests, the probability of getting a false positive in any of them increases. There are several methods to account for multiple testing such as Bonferroni correction and measures such as the false discovery rate [3] and the irreproducible discovery rate [19].

In addition to reporting SNPs which show the strongest associations, we typically also use *Manhattan plots* to show where these SNPs are located in the genome and *quantile-quantile (Q-Q) plots* to detect biases which have not been properly accounted for. A Manhattan plot is a scatter plot of log-transformed p-values against genomic position (concatenating the chromosomes). In Figure 7A, the points in red are those which meet the significance threshold. They are labeled with candidate genes which are close by.

A Q-Q plot is a scatter plot of log-transformed observed p-values against log-transformed expected p-values. We use uniform quantiles as the expected p-values: assuming there is no association, we expect p-values to be uniformly distributed. Deviation from the diagonal suggests p-values are more significant than would be expected. However, early and consistent deviation from the diagonal suggests too many p-values are too significant, i.e. there is some bias which is confounding the test. In Figure 7B, the plot shows observed test statistic against expected test statistic (which is equivalent). Considering all markers includes the Major Histocompatability Complex (MHC), which is the region associated with immune response. This region has a unique LD structure which confounds the statistical analysis, as is clear from the deviation of the black points from the diagonal (the gray area). Throwing out the MHC removes much of this bias from the results (the blue points).

GWAS identifies markers which correlate with the trait of interest. However, each marker captures a neighborhood of SNPs which it is in LD with, making the problem of identifying the causal variant harder. Typically, the candidate gene for a marker is the one which is closest to it. From here, we have to do further study to identify the relevance of the variants which we identify. However, this remains a challenging problem for a few reasons:

- Regions of interest identified by association often implicate multiple genes
- Some of these associations are nowhere near any protein coding segments
- Linking these regions to underlying biological pathways is difficult

Our primary goal is to use these found associations to understand the biology of disease in an actionable manner, as this will help guide therapies in order to treat these diseases. This can be approached in one of

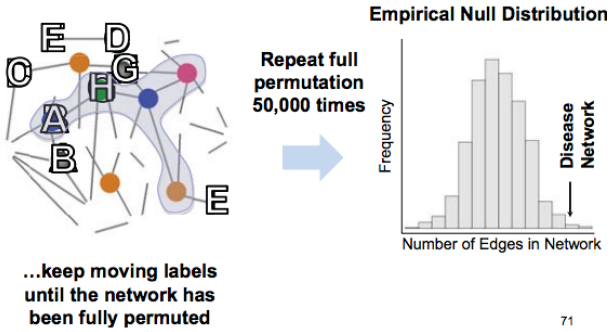

Figure 8: Evaluating Disease Network Significance

two manners: the *bottom-up* approach, or the *top-down* approach.

The **bottom-up** approach is to investigate a particular gene that has a known association with a disease, and investigate it's biological importance within a cell. Kuballa et al.[16] were able to use this bottom-up approach to learn that a particular risk variant associated with Crohn's Disease leads to impairment of autophagy of certain pathogens. Furthermore, the authors were able to create a mouse model of the same risk variant found in humans. Identifying biological implications of risk variants at the cellular level and creating these models is invaluable as the models can be directly used to test new potential treatment compounds.

In contrast, the **top-down** approach involves looking at *all* known associations, and tries to link them to shared biological processes. This approach is based on the idea that many of the associated genes with a disease share relevant biological pathways. This is commonly done by taking existing networks like protein-protein interaction networks, and layering the associated genes on top of them. However, these resulting disease networks may not be significant due to bias in both the discovery of associations and the experimental bias of the data that the associations are being integrated with. This significance can be estimated by permuting the labels for the nodes in the network many times, and then computing how rare the level of connectivity is for the given disease network. This process is illustrated in Figure 8. As genes connected in the network should be co-expressed, it has been shown that these disease networks can be further validated from gene-expression profiling[11].

It is important to note GWAS captures more variants than linkage analysis. Linkage analysis identifies rare variants which have negative effects. GWAS can also identify these variants, but in addition can identify rare variants which have protective effects. Linkage analysis cannot identify these variants because they are anti-correlated with disease status. More importantly, GWAS can identify common variants with smaller effect sizes. Linkage analysis relies on the assumption that a single variant explains the disease. But this assumption does not hold for complex traits such as disease. Instead, we need to consider many markers in order to explain the genetic basis of these traits.

We have learned several lessons from GWAS. First, fewer than one-third of reported associations are coding or obviously functional variants. Second, only some fraction of associated non-coding variants are significantly associated to expression level of a nearby gene. Third, many are associated to regions with no nearby coding gene. Finally, the majority of reported variants are associated to multiple autoimmune or inflammatory diseases. These revelations indicate that there are still many mysteries lurking in the genome waiting to be discovered.

5 Current Research Directions

GWAS helps to identify loci that are associated with a particular disease, but they do not point out specific causal alleles related to these diseases of interest. Advances in **next gen sequencing** (NGS) have made sequencing an individuals genome a much less costly and time consuming task. However, getting from the genome to any casual alleles can be a difficult task. Any given individual may have hundreds of rare variants

in their genome that may be causing loss of function in a gene[21], however, locating this variants and associating them with a particular disease is challenging. However, by performing base pair level sequencing of GWAS confirmed loci, specific alleles relating to disease have emerged[28].

6 Further Reading

7 Tools and Techniques

8 What Have We Learned?

In the past several decades, we have made huge advances in developing techniques to investigate the genetic basis of disease. Historically, we have used linkage analysis to find causal variants for Mendelian disease with great success. More recently, we have used genome-wide association studies to begin investigating more complex traits with some success. However, more work is needed in developing methods to interpret these GWAS and identifying causal variants and their role in disease mechanism. Improving our understanding of the genetic basis of disease will us develop more effective diagnoses and treatments.

References

[1] G.R. Abeçasis, S.S. Cherny, W.O. Cookson, and L.R. Cardon. Merlin—rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genetics*, 30(1):97–101, 2002.

[2] H.L. Allen et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*, 467(7317):832–838, 2010.

[3] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, 57:289–300, 1995.

[4] D. Botstein, R.L. White, M. Skolnick, and R.W. Davis. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American Journal of Human Genetics*, 32:314–331, 1980.

[5] M.S. Brown and J.L. Goldstein. A receptor-mediated pathway for cholesterol homeostasis. *Science*, 232(4746):34–47, 1986.

[6] B. Devlin and K. Roeder. Genomic control for association studies. *Biometrics*, 55:997–1004, 1999.

[7] R.C. Elston and J. Stewart. A general model for the genetic analysis of pedigree data. *Human Heredity*, 21:"523–542", 1971.

[8] Sir R.A. Fisher. The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh*, 52:399–433, 1918.

[9] D.F. Gudbjartsson, K. Jonasson, M.L. Frigge, and A. Kong. Allegro, a new computer program for multipoint linkage analysis. *Nature Genetics*, 25(1):12–13, 2000.

[10] D.F Gudbjartsson, T. Thorvaldsson, A. Kong, G. Gunnarsson, and A. Ingolfsdottir. Allegro version 2. *Nature Genetics*, 37(10):1015–1016, 2005.

[11] X. Hu, H. Kim, E. Stahl, R. Plenge, M. Daly, and S. Raychaudhuri. Integrating autoimmune risk loci with gene-expression data identifies specific pathogenic immune cell subsets. *The American Journal of Human Genetics*, 89(4):496–506, 2011.

[12] R.M. Idury and R.C. Elston. A faster and more general hidden markov model algorithm for multipoint likelihood calculations. *Human Heredity*, 47:197–202, 1997.

[13] A. Ingolfsdottir and D. Gudbjartsson. Genetic linkage analysis algorithms and their implementation. In Corrado Priami, Emanuela Merelli, Pablo Gonzalez, and Andrea Omicini, editors, *Transactions on Computational Systems Biology III*, volume 3737 of *Lecture Notes in Computer Science*, pages 123–144. Springer Berlin / Heidelberg, 2005.

[14] L. Kruglyak, M.J. Daly, M.P. Reeve-Daly, and E.S. Lander. Parametric and nonparametric linkage analysis: a unified multipoint approach. *American Journal of Human Genetics*, 58:1347–1363, 1996.

[15] L. Kruglyak and E.S. Lander. Faster multipoint linkage analysis using fourier transforms. *Journal of Computational Biology*, 5:1–7, 1998.

[16] P. Kuballa, A. Huett, J.D. Rioux, M.J. Daly, and R.J. Xavier. Impaired autophagy of an intracellular pathogen induced by a crohn's disease associated atg16l1 variant. *PLoS One*, 3(10):e3391, 2008.

[17] E.S. Lander and P. Green. Construction of multilocus genetic linkage maps in humans. *Proceedings of the National Academy of Sciences*, 84(8):2363–2367, 1987.

[18] E.S. Lander, P. Green, J. Abrahamson, A. Barlow, M.J. Daly, S.E. Lincoln, and L. Newburg. Mapmaker: An interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. *Genomics*, 1(2):174–181, 1987.

[19] Q. Li, J.B. Brown, H. Huang, and P.J. Bickel. Measuring reproducibility of high-throughput experiments. *Annals of Applied Statistics*, 5:1752–1797, 2011.

[20] E.Y. Liu, Q. Zhang, L. McMillan, F.P. de Villena, and W. Wang. Efficient genome ancestry inference in complex pedigrees with inbreeding. *Bioinformatics*, 26(12):i199–i207, 2010.

[21] D.G. MacArthur, S. Balasubramanian, A. Frankish, N. Huang, J. Morris, K. Walter, L. Jostins, L. Habegger, J.K. Pickrell, S.B. Montgomery, et al. A systematic survey of loss-of-function variants in human protein-coding genes. *Science*, 335(6070):823–828, 2012.

[22] B.P. McEvoy and P.M. Visscher. Genetics of human height. *Economics & Human Biology*, 7(3):294 – 306, 2009.

[23] N.E. Morton. Sequential tests for the detection of linkage. *The American Journal of Human Genetics*, 7(3):277–318, 1955.

[24] N. Patterson, A. Price, and D. Reich. Population structure and eigenanalysis. *PLoS Genetics*, 2:e190, 2006.

[25] A. Piccolboni and D. Gusfield. On the complexity of fundamental computational problems in pedigree analysis. *Journal of Computational Biology*, 10:763–773, October 2003.

[26] A. Price et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38:904–909, 2006.

[27] J. Pritchard, M. Stephens, N. Rosenberg, and P. Donnelly. Association mapping in structured populations. *American Journal of Human Genetics*, 67:170–181, 2000.

[28] M.A. Rivas, M. Beaudoin, A. Gardet, C. Stevens, Y. Sharma, C.K. Zhang, G. Boucher, S. Ripke, D. Ellinghaus, N. Burtt, et al. Deep resequencing of gwas loci identifies independent rare variants associated with inflammatory bowel disease. *Nature genetics*, 2011.

[29] T. Strachan and A.P. Read. *Human Molecular Genetics*. Wiley-Liss, New York, 2 edition, 1999.