Discovering Genes Involved in Disease and the Mystery of Missing Heritability

Appendix

Eleazar Eskin
Department of Computer Science
Department of Human Genetics
University of California, Los Angeles
eeskin@cs.ucla.edu

Appendix 1: GWAS Statistics and Power

In GWAS, the effect of each variant on the trait is examined independently of the other variants. When analyzing variant k, the following model for the effect of variant k on the phenotype is utilized

$$y_j = \mu^* + \beta_k x_{kj} + e_j^* \quad (1)$$

and in vector notation

$$y = \mu^* \mathbf{1} + \beta_k X_k + \mathbf{e}^* \quad (2)$$

where X_k is a column vector of normalized genotypes for variant k and $\mathbf{e}^* \sim \mathcal{N}(0, \sigma^2_{e^*}\mathbf{I})$ We note that equation (1) above and equation (1) in the main text differ by the omission of the terms $\sum_{i \neq k} \beta_i x_{ij}$ which results in the values of these terms being absorbed in the mean and residual which is why we use the notation μ^* and e_j^* instead of μ and e_j. We describe the implications of this omission in more detail in the main text.

Using equation (1), we can use the observed data to obtain an estimate of β_k. This reduces to a simple regression problem where the resulting estimates are $\hat{\mu} = \frac{1}{N}\mathbf{1}^T y$, $\hat{\beta}_k = (X_k^T X_k)^{-1} X_k^T y = \frac{X_k^T y}{N}$ since X_k is normalized so $X_k^T X_k = N$. The estimated residuals $\hat{\mathbf{e}} = y - \hat{\mu}\mathbf{1} - \hat{\beta}_k X_k$ can be used to estimate the standard error $\hat{\sigma} = \sqrt{\frac{\hat{\mathbf{e}}^T \hat{\mathbf{e}}}{n-2}}$. Since the studies are large, the association statistic

$$S_k = \frac{\hat{\beta}_k}{\hat{\sigma}}\sqrt{N} \sim \mathcal{N}\left(\frac{\beta_k}{\sigma_{e^*}}\sqrt{N}, 1\right) \quad (3)$$

will approximately follow the standard normal distribution under the null hypothesis of no association and can be used to determine whether or not the statistic is significant.

Since GWASes collect many markers, the significance threshold is adjusted for multiple hypothesis testing. The community has settled on using the significance threshold of $\alpha_s = 5 \times 10^{-8}$ as the genome-wide significance threshold which takes into account the large number of SNPs in the human genome. Since our statistics, S_k, are normally distributed, a GWAS simply computes S_k for each variant and then checks to see if $\Phi(S_k) < \alpha_s/2$ or $\Phi(S_k) > 1 - \alpha_s/2$ in which case the variant is associated. $\Phi(x)$ computes the cumulative standard normal distribution. Figure 1 shows an example of checking for significance of a variant using the normal distribution.

The statistical power measures the probability of detecting an association under the assumption that an association is present with a certain effect size. Intuitively, the power measures the probability that the truly associated variants will be discovered. Since statistical power depends on both the effect size and the number of individuals in the study, statistical power can be used to guide the choice of study size as well as provide expectations on what effect sizes can and can not be discovered in association studies.

Using the distributions above, we can also estimate the statistical power of an association study. We assume that the effect size at variant i is β_i and the residual variance from equation (1) is $\sigma^2_{e^*}$. Since we know that the statistic S_k follows the distribution in equation (3), the question is how often this statistic is significant (i.e. $\Phi(S_k) < \alpha_s/2$ or $\Phi(S_k) > 1 - \alpha_s/2$). This probability can be estimated using the following

$$P(\alpha_s, \beta, \sigma, N) = \Phi(-\Phi^{-1}(\alpha_s/2) + \frac{\beta}{\sigma}\sqrt{N}) + 1 - \Phi(\Phi^{-1}(\alpha_s/2) + \frac{\beta}{\sigma}\sqrt{N}) \quad (4)$$

Note that the power depends on what is referred to as the non-centrality parameter (in our case $\frac{\beta}{\sigma}\sqrt{N}$) which

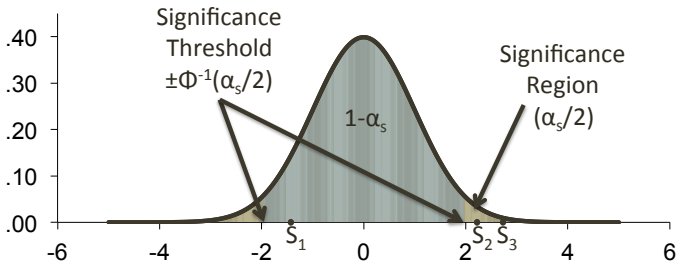

Figure 1: Significance testing in association studies. The null distribution is shown which is the standard normal distribution and the expected distribution of the association statistics under the assumption that the effect size is 0. For each variant, that association statistic in equation (3) is computed and its significance is evaluated using the null distribution. If the statistic falls in the significance region of the distribution, the variant is declared associated. In this example, S_1 is not significant and S_2 and S_3 are significant. The exact location of the threshold is defined as the location on the x axis where the tail probability area equals the significance threshold (α_s). This is denoted using the quantile of the standard normal $\pm\Phi^{-1}(\alpha_s/2)$.

is the mean of the distribution of the statistic under the assumption of the genetic effect. A visualization of estimating the power is shown in Figure 2.

Figure 3(a) and 3(b) show the effect of minor allele frequency and study size on the power of discovering associations. As can be shown in the figure, for small minor allele frequencies, even very large studies have very low power to detect associations.

Appendix 2: Computational Problems in Genetics

- how to efficiently estimate mixed model parameters, which has been an active area of research for several years[15, 13, 20, 30]
- how to efficiently identify pairs of segments in individuals which were inherited from a recent common ancestry[4, 5, 9]
- predicting haplotypes or the sequence of alleles on a chromosome from the genotype information which mixes the two chromosomes[6, 2, 7, 1, 10, 26]
- identifying the population origin of each region of an individual's genome for individuals who are admixed or a mixture of multiple ancestral populations[24, 23].
- identifying the geographical origin of an individual[22, 27, 3]
- identifying pairs of genetic variants which have a larger effect on the trait than each individually[29, 28]
- inferring the genetic relationships between individuals[17, 19, 11]
- inferring the genetic history of a region of the genome in the population which is referred to as an ancestral recombination graph[25]
- predicting missing genotype data in an association study[21, 14, 12]
- correcting for multiple testing in genome-wide association studies[16, 8]
- efficiently computing association statistics[18]

1. REFERENCES

[1] D. Aguiar and S. Istrail. Haplotype assembly in polyploid genomes and identical by descent shared tracts. *Bioinformatics*, 29(13):i352–i360, 7 2013.

[2] V. Bafna, D. Gusfield, S. Hannenhalli, and S. Yooseph. A note on efficient computation of haplotypes via perfect phylogeny. *J Comput Biol*, 11(5):858–66, 2004.

[3] Y. Baran, I. Quintela, A. Carracedo, B. Pasaniuc, and E. Halperin. Enhanced localization of genetic samples through linkage-disequilibrium correction. *Am J Hum Genet*, 5 2013.

[4] B. L. Browning and S. R. Browning. Improving the accuracy and efficiency of identity by descent detection in population data. *Genetics*, 194(2):459–71, 3 2013.

[5] A. Gusev, J. K. Lowe, M. Stoffel, M. J. Daly, D. Altshuler, J. L. Breslow, J. M. Friedman, and I. Pe'er. Whole population, genome-wide mapping of hidden relatedness. *Genome Res*, 19(2):318–26, 2 2009.

[6] D. Gusfield. Haplotyping as perfect phylogeny: Conceptual framework and efficient solutions. In *Research in Computational Molecular Biology*, pages 166–175. ACM, 2002.

[7] E. Halperin and E. Eskin. Haplotype reconstruction from genotype data using imperfect phylogeny. *Bioinformatics*, 20(12):1842–9, 8 2004.

[8] B. Han, H. M. Kang, and E. Eskin. Rapid and accurate multiple testing correction and power estimation for millions of correlated markers. *PLoS Genet*, 5(4):e1000456, 4 2009.

[9] D. He. IBD-Groupon: an efficient method for detecting group-wise identity-by-descent regions simultaneously in multiple individuals based on pairwise IBD relationships. *Bioinformatics*, 29(13):i162–i170, 7 2013.

[10] D. He, B. Han, and E. Eskin. Hap-seq: An optimal algorithm for haplotype phasing with imputation using sequencing data. *J Comput Biol*, 20(2):80–92, 2 2013.

[11] D. He, Z. Wang, B. Han, L. Parida, and E. Eskin. Iped: Inheritance path-based pedigree reconstruction algorithm using genotype data. *J Comput Biol*, 20(10):780–91, 10 2013.

[12] B. Howie, C. Fuchsberger, M. Stephens, J. Marchini, and G. R. Abecasis. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat Genet*, 44(8):955–9, 2012.

[13] H. M. Kang, J. H. Sul, S. K. Service, N. A. Zaitlen, S.-Y. Y. Kong, N. B. Freimer, C. Sabatti, and E. Eskin. Variance component model to account for sample structure in genome-wide association studies. *Nat Genet*, 42(4):348–54, 4 2010.

[14] H. M. Kang, N. A. Zaitlen, and E. Eskin.

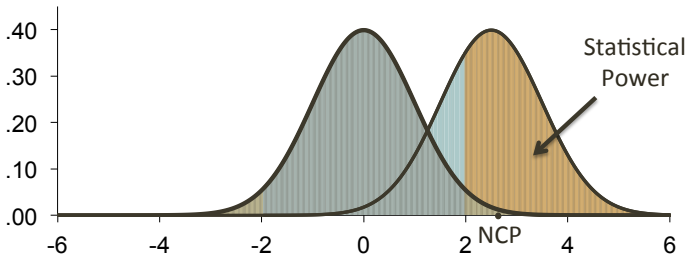

Figure 2: Power of association studies. The power is defined by considering the expected distribution of the association statistics assuming a specific effect size which is referred to as the alternative distribution. This effect size as well as the number of individuals in the study define the non-centrality parameter (NCP) of the alternative distribution. The area of the alternative (noncentral) distribution outside the significance threshold defined by the null distribution is the probability that an observation under the assumption of the specific effect size will be declared significant. This probability is referred to as the power.

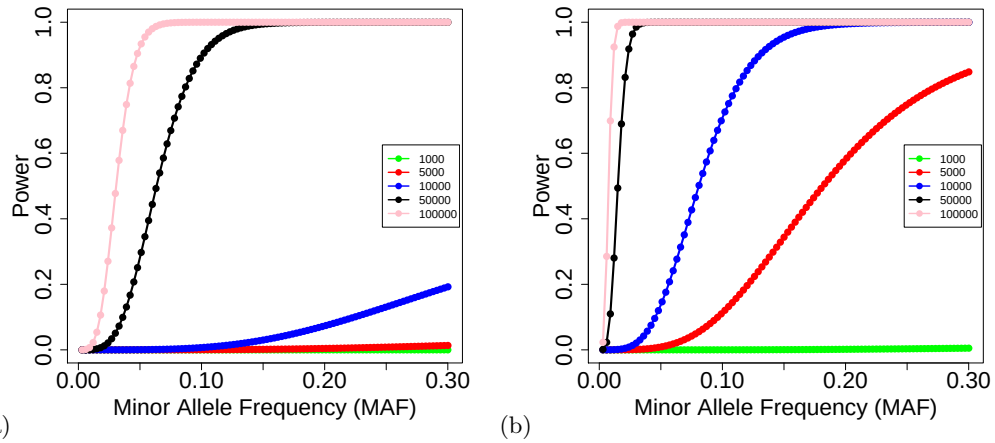

Figure 3: The effect of minor allele frequency and study size on the statistical power of a GWAS. Power is shown for studies of size 1000, 5000, 10,000, 50,000 and 100,000 as a function of minor allele frequency for (a) an effect equal to 10% of a standard deviation of the phenotype and (b) a larger effect equal to 20% of a standard deviation of the phenotype.

EMINIM: an adaptive and memory-efficient algorithm for genotype imputation. *J Comput Biol*, 17(3):547–60, 3 2010.

[15] H. M. Kang, N. A. Zaitlen, C. M. Wade, A. Kirby, D. Heckerman, M. J. Daly, and E. Eskin. Efficient control of population structure in model organism association mapping. *Genetics*, 178(3):1709–23, 3 2008.

[16] G. Kimmel, M. I. Jordan, E. Halperin, R. Shamir, and R. M. Karp. A randomization test for controlling population stratification in whole-genome association studies. *Am J Hum Genet*, 81(5):895–905, 11 2007.

[17] B. Kirkpatrick, S. Li, R. Karp, and E. Halperin. Pedigree reconstruction using identity by descent. In V. Bafna and S. Sahinalp, editors, *Research in Computational Molecular Biology*, pages 136–152. Springer, 2011.

[18] E. Kostem and E. Eskin. Efficiently identifying significant associations in genome-wide association studies. In *Research in Computational Molecular Biology*, pages 118–131. Springer, 2013.

[19] S. Kyriazopoulou-Panagiotopoulou, D. Kashef Haghighi, S. J. Aerni, A. Sundquist, S. Bercovici, and S. Batzoglou. Reconstruction of genealogical relationships with applications to phase III of HapMap. *Bioinformatics*, 27(13):i333–41, 7 2011.

[20] C. Lippert, J. Listgarten, Y. Liu, C. M. Kadie, R. I. Davidson, and D. Heckerman. Fast linear mixed models for genome-wide association studies. *Nat Methods*, 8(10):833–5, 2011.

[21] J. Marchini, B. Howie, S. Myers, G. McVean, and P. Donnelly. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet*, 39(7):906–13, 7 2007.

[22] J. Novembre, T. Johnson, K. Bryc, Z. Kutalik, A. R. Boyko, A. Auton, A. Indap, K. S. King, S. Bergmann, M. R. Nelson, M. Stephens, and C. D. Bustamante. Genes mirror geography within europe. *Nature*, 456(7218):98–101, 11 2008.

[23] S. Sankararaman, S. Sridhar, G. Kimmel, and E. Halperin. Estimating local ancestry in admixed populations. *Am J Hum Genet*, 82(2):290–303, 2 2008.

[24] A. Sundquist, E. Fratkin, C. B. Do, and S. Batzoglou. Effect of genetic divergence in identifying ancestral origin using hapaa. *Genome Res*, 18(4):676–82, 4 2008.

[25] Y. Wu. Association mapping of complex diseases with ancestral recombination graphs: Models and efficient algorithms. In *Research in Computational Molecular Biology*, pages 488–502. Springer, 2007.

[26] W.-Y. Y. Yang, F. Hormozdiari, Z. Wang, D. He, B. Pasaniuc, and E. Eskin. Leveraging multi-SNP reads from sequencing data for haplotype inference. *Bioinformatics*, 7 2013.

[27] W.-Y. Y. Yang, J. Novembre, E. Eskin, and E. Halperin. A model-based approach for analysis of spatial structure in genetic data. *Nat Genet*,

44(6):725–31, 2012.

[28] X. Zhang, S. Huang, F. Zou, and W. Wang. Team: efficient two-locus epistasis tests in human genome-wide association study. *Bioinformatics*, 26(12):i217–27, 6 2010.

[29] X. Zhang, F. Pan, Y. Xie, F. Zou, and W. Wang. COE: A general approach for efficient genome-wide two-locus epistasis test in disease association study. In *Research in Computational Molecular Biology*, pages 253–269. Springer, 2009.

[30] X. Zhou and M. Stephens. Genome-wide efficient mixed-model analysis for association studies. *Nat Genet*, 44(7):821–4, 2012.