Estimation of the Multiple Testing Burden for Genomewide Association Studies of Common Variants

Itsik Pe'er[1], Roman Yelensky[2,3,8], David Altshuler[2,3,4,5,7], Mark J. Daly[2,5,6,%]

[1]Department of Computer Science, Columbia University, New York, NY; [2]Center for Human Genetic Research, [3]Department of Molecular Biology and [4]Diabetes Unit, Massachusetts General Hospital, Boston, MA; [5]Broad Institute of M.I.T. and Harvard, Cambridge, MA; Departments of [6]Medicine and [7]Genetics, Harvard Medical School, Boston; [8]Harvard-M.I.T. Division of Health Sciences and Technology, Cambridge, MA

[%] To whom correspondence should be addressed:
Mark J. Daly
Massachusetts General Hospital
185 Cambridge Street
CPZN-6818
Boston, MA 02114-2790
Phone: (617) 643-3290
Fax: (617) 643-3293
Email: mjdaly@chgr.mgh.harvard.edu

Running title: Testing burden in genetic association studies

Keywords: human genetics, association studies, testing burden

Whole Genome Association Studies (WGASs) offer a systematic strategy to assess the influence of common (minor allele frequency ≥ 5%) genetic variants on phenotypes (Risch and Merikangas 1996). Most variants tested will not be associated to any particular phenotype, but may produce false positive association signals, masking potential true positives. Forecasting these null-distribution of false-positives is important as a practical guideline for interpreting genomewide association scans, akin to classical work (Lander and Kruglyak 1995) directing linkage analysis. The concrete question is, given an association signal of a certain nominal p-value, how unlikely is it in a WGAS?

Naïve, Bonferroni (Sidak 1967) corrections for standard testing of multiple, independent hypotheses are overconservative in this context: local correlation among these tests means that effectively there are considerably less independent tests than Single Nucleotide Polymorphisms (SNPs) examined. Theoretical (Tavare et al. 1997) and simulation studies (Lin et al. 2004) relate the number of such tests to the number of historical recombinations, estimated to be much smaller. Yet, no previous systematic evaluation of the testing burden.is available.

Such an evaluation is particularly critical to the standard, two step design for WGASs (Thomas et al. 2004; Skol et al. 2006) that does not lend itself to significance evaluation by permuting phenotypic labels. In this design common variation is first screened for association signals using cost-effective typing of hundreds of thousands of SNPs (Barrett and Cardon 2006; Pe'er et al. 2006). Next, regions of potentially positive signals are followed-up with denser, saturated SNP sets, in order to validate, refine and strengthen the associations. As well worked out in linkage analysis (Kruglyak and Daly 1998), this directed increase in marker density around positives alters the null signal distribution with the practical effect of mimicking a WGAS of all 6-7 million common SNPs. Hence, permuting data with only the smaller, typed set of SNPs underestimates expected false positives.

The testing burden associated with examining all common alleles does lend itself to empirical evaluation from data, thanks to the Human Haplotype Map (HapMap)

ENCODE regions (Altshuler et al. 2005). These regions offer near-complete description of common SNPs (Pe'er et al. 2006), and allow simulating association studies with no true signal (de Bakker et al. 2005). More specifically, we generate the genetic data for a simulated (case or control) individual at an ENCODE region by randomly pairing two of the phased chromosomes available from HapMap trios for that region. We repeat this to obtain 2000 individuals randomly labeled cases or controls, mimicking a null study. The maximal-scoring difference in allele frequencies between "cases" and "controls" across all SNPs in such a region is evaluated for significance, and the p-value distribution is estimated by repeating the simulation 10^7 times. This distribution observes more significant p-values then theoretically distributed p-values for a single test statistic due to multiple testing. We repeat this evaluation procedure for the trio-base HapMap populations (CEU and YRI), for all ENCODE regions, and for different cohort sizes. The per-region testing burden is the factor by which significance is exaggerated. As ENCODE regions represent the genomewide average recombination and mutation rates, we can extrapolate to estimate the genomewide testing burden in such an association study.

Figure 1a reports the extrapolated number of independent tests required to mimic the expectation of the best p-value in a WGAS, i.e. the empirical testing burden. For all common SNPs, we find the testing burden to be considerably lower than available bounds[1], at half million tests in the HapMap European (CEU) samples. This means, for instance, that the probability of a WGAS in a European population that examines all common alleles to exhibit, by random chance alone (no true genetic effect), a result with p-value$<10^{-7}$ is smaller than 0.05. In the HapMap African (YRI) samples, that have more SNPs, and less linkage disequilibrium, testing burden is higher at one million. Since

---

[1] The formula $\mathrm{Log}(k)\times N_e\times R$ in

Tavare S, Balding DJ, Griffiths RC, Donnelly P (1997) Inferring coalescence times from DNA sequence data. Genetics 145(2): 505-518. esti,mates 1.1million common recombinations in Europeans, where:

- k is the number of coalescence branches considered the reciprocal of the minor allele frequency threshold for sites considered, i.e. $k = 20$ for common SNPs.
- N_e is the effective population size , ~10,000 in Europeans
- R is the average number of recombination events per meiosis, 36 [8]

ENCODE data are still incomplete w.r.t. rare variants, they provide only a lower bound on their associated testing burden, showing it to be more than 2-fold higher than for common alleles.

ENCODE regions deliberately represent a variety of genomic characteistics (The International HapMap Consortium 2003), and also testing burden varies greatly from region to region. Yet, testing burden is not strongly correlated with neither the actual number of common SNPs in the particular region, nor the regionwide recombination rate (Fig 1b).

It is important to realize that testing burden is not constant across p-values: association signals with more extreme p-values involve more burden (Fig 1c). This is because the power of such signals to distinguish better between partially correlated tests, resulting in more testing burden. With weaker signals, correlated tests are undistinguishable, hence testing burden is reduced. This means the best practice for correcting a nominal p-value for the entire genome is to use a lookup-table, rather than a fixed correction factor. Fortunately, the first stage of a WGAS is designed for a true positive to reach only a moderate p-value, expected to be achieved be numerous sites (Skol et al. 2006). Such a stage would require less correction for multiple testing than the final stage aiming at genomewide significance. Finally, studies of larger size show more burden of multiple testing (Supplementary fig 1). We hypothesize that this effect is also related to the increased power of larger studies to distinguish highly- (but not perfectly-) correlated causal variants.

These and other results offer considrable understanding of the distribution of null signals in idealized association studies. Practical association studies may exhibit more extreme p-values then predicted by our study even without real effects due to demographical and genotyping technology differences between cases and controls that create artifactual hits. Furthermore only the accumulating experience in such studies will reveal more about the complementary parameters describing the alternative hypothesis, which speak to the number and strength of true signals. Together, the distribution of null and true signals will enable rigorous decision whether a given result indicates true association.

Figure 1 legend: A. The empirical testing burden (y-axis) for all common SNPs in different ENCODE regions in the HapMap panels of Yorubans from Ibadan, Nigeria (YRI; green) and CEPH individuals of European ancestry fro Utah (CEU; orange). Testing burden is evaluated as the reciprocal of the best nominal p-value expected in a null study of 1000 cases, 1000 controls extrapolate to the entire genome, as extrapolated from ENCODE. **B**. The testing burden (y-axis) of each region as a function of the region's length in centiMorgans (x-axis, left) or of the number of SNPs tested (x,axix, right) **C**. The testing burden (y-axis) of all (smooth) or common (tick-marked) SNPs in a typical ENCODE region (ENr213), as a function of the empirically evaluated p-value (x-axis).

Supplementary Figure 1 legend:
Testing burden (y-axis) extrapolated to the entire genome from simulated sturies different numbers of cases/controls (x-axis) in YRI (green) and CEU (orange) data from ENCODE, averaged across all regions.

Refenrences

Altshuler D, Brooks LD, Chakravarti A, Collins FS, Daly MJ et al. (2005) A haplotype map of the human genome. Nature 437(7063): 1299-1320.

Barrett J, Cardon L (2006) Study Design Issues in Whole Genome Association Studies. Nat Genet Submitted.

de Bakker PI, Yelensky R, Pe'er I, Gabriel SB, Daly MJ et al. (2005) Efficiency and power in genetic association studies. Nat Genet 37(11): 1217-1223.

Kruglyak L, Daly MJ (1998) Linkage thresholds for two-stage genome scans. Am J Hum Genet 62(4): 994-997.

Lander E, Kruglyak L (1995) Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. Nat Genet 11(3): 241-247.

Lin S, Chakravarti A, Cutler DJ (2004) Exhaustive allelic transmission disequilibrium tests as a new approach to genome-wide association studies. Nat Genet 36(11): 1181-1188.

Pe'er I, Chretien Y, PIW PdB, Barrett J, Daly M et al. (2006) Biases and reconciliation in estimations of linkage disequilibriumin the human genome. American journal of human genetics 73(4).

Pe'er I, de Bakker PI, Maller J, Yelensky R, Altshuler D et al. (2006) Evaluating and Improving Power in Whole Genome Association Studies using Fixed Marker Sets. Nat Genet 38(6).

Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. Science 273(5281): 1516-1517.

Sidak Z (1967) Rectangular Confidence Regions for the Means of Multivariate Normal Distributions. Journal of the American Statistical Association 62: 626-633.

Skol AD, Scott LJ, Abecasis GR, Boehnke M (2006) Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. Nat Genet 38(2): 209-213.

Tavare S, Balding DJ, Griffiths RC, Donnelly P (1997) Inferring coalescence times from DNA sequence data. Genetics 145(2): 505-518.

The International HapMap Consortium (2003) The International HapMap Project. Nature 426(6968): 789-796.

Thomas D, Xie R, Gebregziabher M (2004) Two-Stage sampling designs for gene association studies. Genet Epidemiol 27(4): 401-414.

Figures

Fig 1a

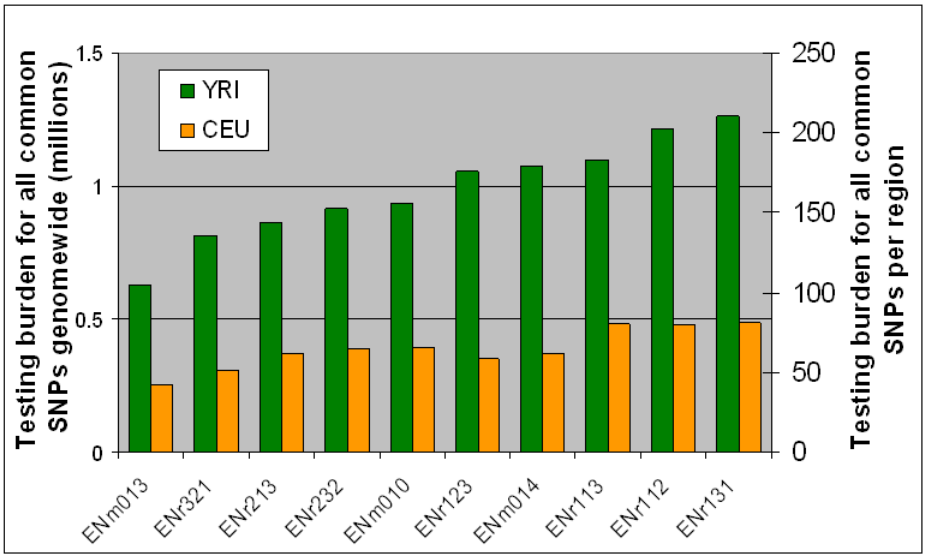


Fig 1b

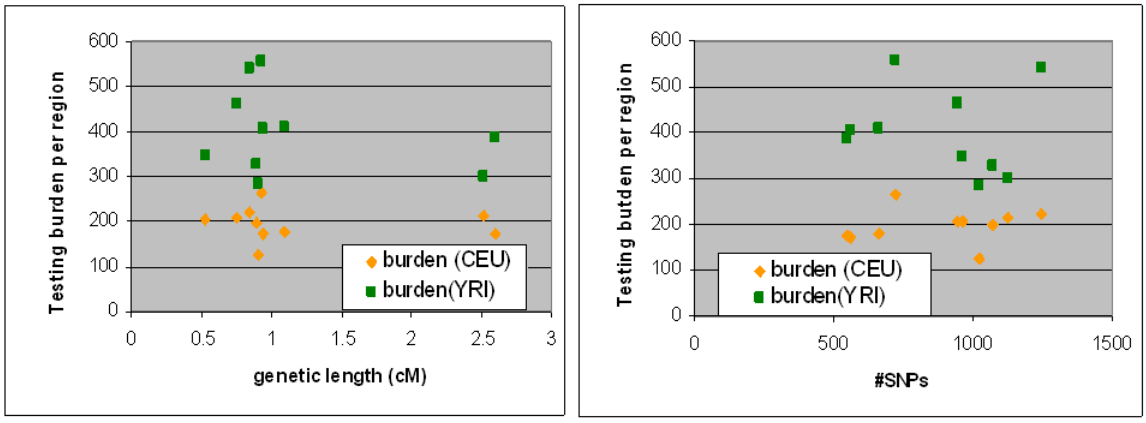
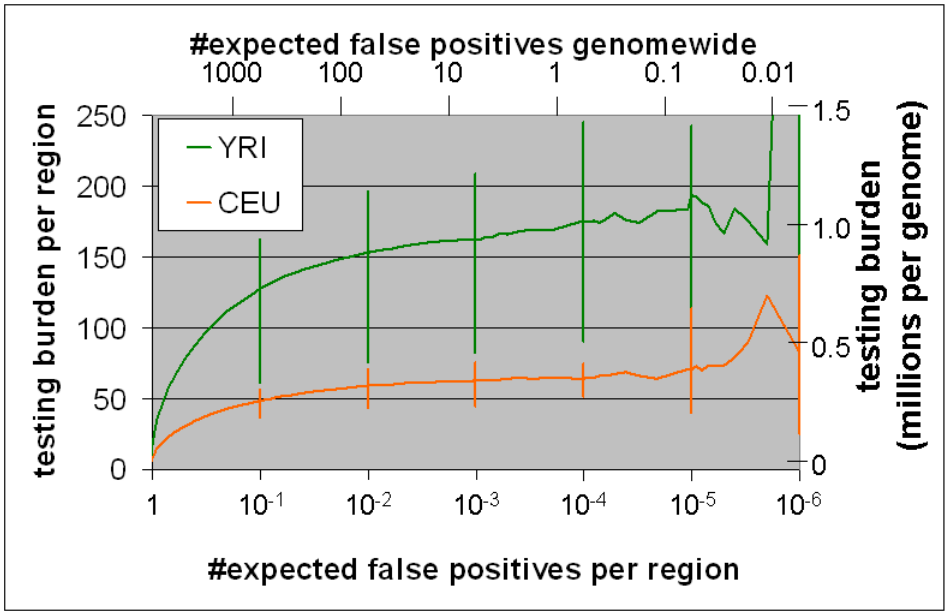


Fig 1c



Supp Fig 1