Brief Report

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

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

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

Genomewide association studies are an exciting strategy in genetics, recently becoming feasible and harvesting many novel genes linked to multiple phenotypes. Determining the significance of results in the face of testing a genomewide set of multiple hypotheses, most of which are producing noisy, null-distributed association signals, presents a challenge to the wide community of association researchers. Rather than each study engaging in independent evaluation of significance standards, we have undertaken the task of developing such standards for genomewide significance, based on data collected by the International Haplotype Map Consortium. We report an estimated testing burden of a million independent tests genomewide in Europeans, and twice that number in Africans. We further identify the sensitivity of the testing burden to the required significance level, with implications to staged design of association studies. *Genet. Epidemiol.* 2008. © 2008 Wiley-Liss, Inc.

Key words: human genetics; association studies; testing burden

The Supplementary Material described in this article can be found at http://www.interscience.wiley.com/jpages/0741-0395/suppmat.
Contract grant sponsor: National Center for the Multiscale Analysis of Genomic and Cellular Networks (MAGNet); Contract grant sponsor: NIH; Contract grant number: 5 U54 CA121852.
*Correspondence to: Mark J. Daly, Massachusetts General Hospital, 185 Cambridge Street, CPZN-6818, Boston, MA 02114-2790.
E-mail: mjdaly@chgr.mgh.harvard.edu
Received 5 July 2007; Revised 12 October 2007; Accepted 6 December 2007
Published online in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/gepi.20303

Whole Genome Association Studies (WGASs) are examinations of a dense set of single nucleotide polymorphisms (SNPs) across essentially all available regions of the genome to survey much of common genetic variation for a role in heritable disease traits. WGASs [Hirschhorn and Daly, 2005] offer a systematic strategy to assess the influence of common (minor allele frequency $\geq 5\%$) genetic variants on phenotypes [Risch and Merikangas, 1996]. Although the number of SNPs typed in such a study may vary, typically between 10^5 and 10^6 SNPs, statistical analysis often involves additional testing, so that the number of added tests dominates the number of typed-SNPs tested. This additional testing may involve consideration of combinations of typed, promising SNPs that predict nearby alleles in the original samples [Klein et al., 2005; Wellcome Trust Case Control Consortium, 2007], of experimental, second stage typing of such alleles in additional samples [Arking et al., 2006] or of additional sets of SNPs typed in another study for joint analysis [Saxena et al., 2007; Scott et al., 2007; Zeggini et al., 2007]. In all these scenarios, the WGAS aspires to test association to more variants than physically typed, ideally testing all common variants in the genome. Most variants tested will not be associated to any particular phenotype, but may produce false-positive association signals, masking potential true positives. Forecasting the null distribution of these false-positives is important as a practical guideline for interpreting genomewide association scans, akin to classical work [Lander and Kruglyak, 1995] directing genomewide linkage analysis of indirectly typed variants. The concrete question is, given an association signal of a certain nominal *P*-value, how unlikely is it in a WGAS that attempts to examine all common variants?

The number of SNPs on the array may guide multiple testing correction if only these SNPs are tested for genetic association. In contrast, we focus on testing not only typed SNPs but also most other common variants in the genome. 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 (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 on a dense data set that can mimic fine mapping on a near-complete scan of variation, such as the second stage in a multi-staged design.

Such an evaluation is particularly critical to study designs that include a second stage of additional genotyping [Thomas et al., 2004; Skol et al., 2006] or analysis [Klein et al., 2005] around putative causal SNPs that are proposed by the first-stage analysis, as these designs do not trivially lend themselves to significance evaluation by permuting phenotypic labels. For 2-stage genotyping designs, 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., 2006a,b]. Next, regions of potentially positive signals are followed up with denser, saturated SNP sets, 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 first-stage data with only the smaller, typed set of SNPs underestimates expected false positives. Permuting the second stage data is possible only for the regions that were followed up; therefore, impossible to implement in a nested fashion for every permutation run of the first stage.

Implementation of a permutation procedure for study designs with a second stage of analysis in promising regions requires rigorous, automatic criteria for such follow-up. As second stage analysis may be based on post hoc review of the associated region, pinning down the desired follow-up criteria in an objective fashion is challenging.

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 [The International HapMap Consortium, 2005]. These regions offer near-complete description of common SNPs [Pe'er et al., 2006a,b] across 1/600 of both the physical and the genetic length of the genome. The demonstrated ability of these regions to represent linkage disequilibrium among common variants across the genome [Pe'er et al., 2006a,b; The International HapMap Consortium, 2005] allows their use for 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 trio parents for that region. We repeat this to obtain 2,000 individuals randomly labeled cases or controls, mimicking a null study. The maximal Z-score 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 $N = 10^7$ times. This distribution observes more significant P-values than 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 propose ENCODE-based extrapolation to estimate the genomewide testing burden in such an association study.

We now outline a formal procedure for estimating testing burden. Suppose the simulation considers a region that spans a fraction $g = 1/600$ of the genome (all ENCODE regions totaling 5 Mb). For a nominal P-value, p that is computed from the theoretical distribution of the association statistic, we tally $n(p)$, the number of studies out of N simulated, at which the best regionwide nominal P-value reaches or exceeds p. $n(p)/N$ is therefore an estimator of the permutation-based P-value regionwide. The expected number H of hits—regions that have a SNP whose score exceeds p—across the genome is therefore $H(p) = n(p)/gN$. Testing burden is defined to be the ratio between the nominal and permutation-based P-values: $n(p)/pN$ regionwide or $n(p)/pgN$ genomewide. This can be estimated for every p. Choosing p such that $H(p) = 0.05$ would be relevant for the genomewide significance threshold in the initial cohort, whereas $H(p) > 1$ would be relevant to a 2-stage design that carries over $H(p)$ false-discovery loci to be typed in additional samples. We chose the middleground, focusing on the value p relevant for a single null hit genomewide. This is motivated by two potential practical outcomes of a study. If a study includes several positive findings, false-discovery rates will be much lower than one even when $H(p) = 1$, motivating interest in SNPs at that significance level. Alternatively, even in studies consistent with the null hypothesis of no association, this significance level is interesting, as it is approached or attained by the top SNPs that are the most suggestive candidates such a study may propose for additional investigation. We note that this threshold does not formally control familywise error rate, nor false-discovery rate, and is intended to provide practical guidelines, rather than be taken literally.

We observe that when $H(p) = 1$, the genomewide burden is simply $1/p$. Putting this observation to practical use, we sort the N respective top single-hits in each of the simulations from the smallest (most significant) to the largest. We choose p to be the gN th value up this list, and report the reciprocal as the testing burden. We note that for a single ENCODE region, the expected number of runs achieving such a P-value among $N = 10^7$ simulations is $gN = 500\,\text{kb}/3\,\text{Gb} \times 10^7 \approx 1,700$, and the standard deviation of this number is $\sigma = \sqrt{g(1-g)N} \approx 40$. This provides a practical way to estimate confidence in estimating the gNth order statistic due to the number of simulations being finite by considering ($gN-2\sigma$)th and ($gN+2\sigma$)th order statistics. Another source of sampling error has to do with the small fraction of the genome being analyzed. The differences in estimation across ENCODE regions can guide us with respect to this sampling error.

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 ENCODE SNPs, we find the testing burden to be around one million tests in the

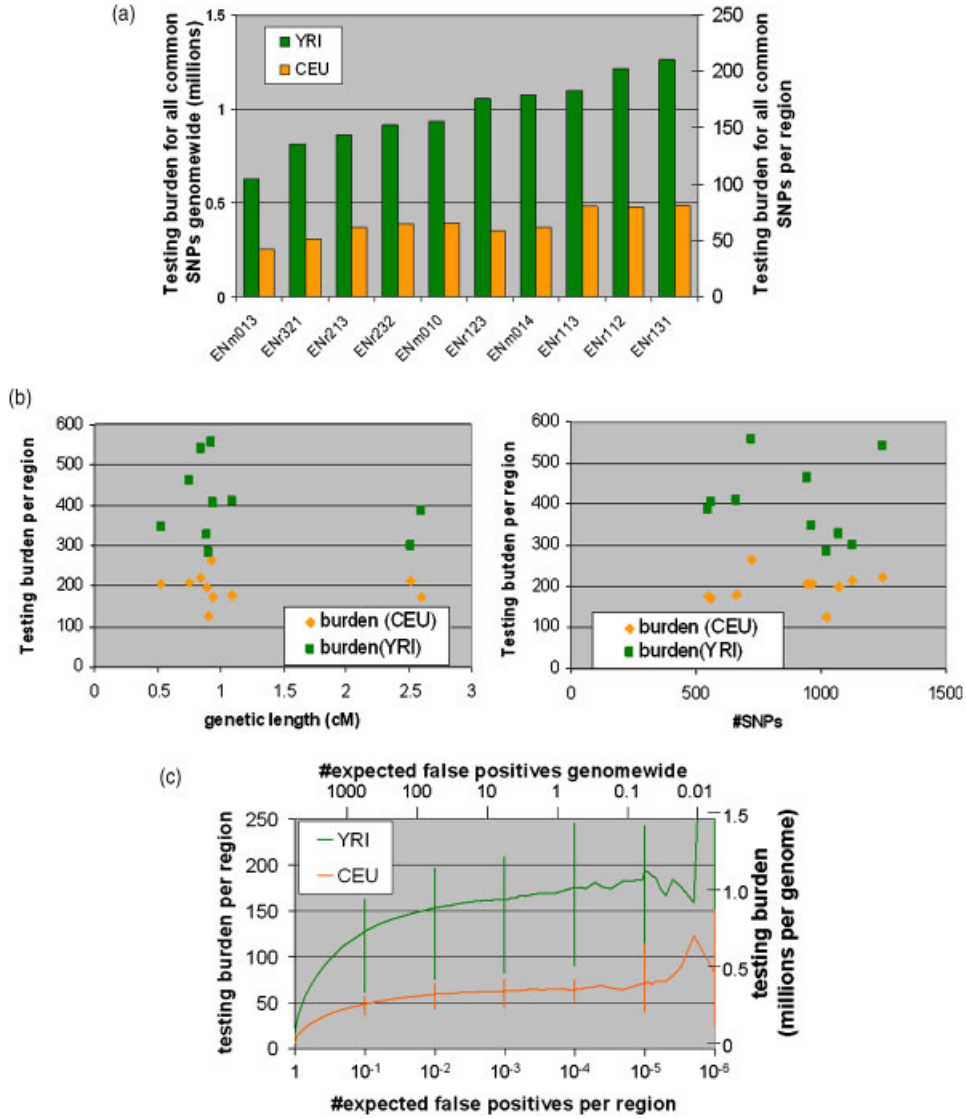

Fig. 1. A. The empirical testing burden (y-axis) for all common SNPs in different ENCODE regions in the HapMap panels of Yorubans from Ibadan, Nigeria (YRI; green) and CEPH individuals of European ancestry from Utah (CEU; orange). Testing burden is estimated from simulated null studies of 1,000 cases, 1,000 controls extrapolated 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-axis, 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). SNPs, single nucleotide polymorphisms. [Color figure can be viewed in the online issue which is available at www.interscience.wiley.com]

HapMap European (CEU) samples, and for all common SNPs, we find the testing burden to be roughly half million tests in the same population: considerably lower than available boundsthat prove the number of edges in the ancestral recombination graph to exceed the number of independent tests in a data set [Lin et al., 2004]. As such edges can be attributed to either splits or recombinations, their number depends on the sample size (negligible in the context of the entire genome) and ancestral recombination events. The formula $\log(k) \times N_e \times R$ in [Tavare et al., 1997] estimates 1.1 million 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.

A practical, first-cut guideline for correcting nominal P-values may be multiplying them by this genomewide testing burden. 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 ENCODE data are still incomplete with respect to rare variants, they provide only a lower bound on their associated testing burden, showing it to be more than two-fold higher than for common alleles.

Testing burden varies across the different ENCODE regions, which may be expected given that ENCODE regions deliberately represent a variety of genomic characteristics [The International HapMap Consortium, 2003]. Empirical standard deviation across the 10 regions amounts to 19.6% of the testing burden, in both YRI and CEU populations (Fig. 1A), suggesting a standard error of 6.2% in estimating average testing burden from 10 regions. We have evaluated the sampling error due to finite number of simulation by considering different order statistics as described above, and showed it to be smaller than 0.2%. We therefore ascribe most of the observed variation in estimates to sampling different regions. Yet, the process of selecting of ENCODE regions made sure their average GC content, gene content, recombination rate, etc. were similar to the genomewide average [The International HapMap Consortium, 2003]. Although we offer no genomewide evidence that ENCODE is the representative of the genome in terms of other measures such as testing burden examined here, this premise, adopted by others using ENCODE data, is used as a standard benchmark for estimating frequencies of genomewide phenomenon in a wide domain of applications [Birney et al., 2007]. We note that testing burden is not strongly correlated neither with the actual number of common SNPs in the particular region ($R^2 < 0.03$) nor with the regionwide recombination rate ($R^2 < 0.01$; see Fig. 1B). In retrospect, this justifies extrapolation of our measurements from ENCODE to the entire genome by physical span.

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 means that accurately correcting statistical tests by a constant factor is impossible. Our simulations validate the formal analysis of modeling multiple genetic tests [Dudbridge and Koeleman, 2004; Hirschhorn and Daly, 2005] in pointing out that restriction of such modeling to a constant testing burden does not sufficiently capture the full correlation structure between tests. There is no genomewide testing burden to fit all significance levels, but rather one can correct for such multiple testing by a burden function, which depends on the significance level of interest. This means that 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.

In order to better understand the intuition behind this variable testing burden, we recall that a constant testing burden arises in the context of independent multiple statistical tests. In contrast, dense SNPs along the genome are partially and locally correlated to varying extents. Formally, the pair (Z_a, Z_b) of Z score statistics of two correlated alleles of different, nearby SNPs, a and b, respectively, will have a bivariate normal distribution, with mean (0,0) and covariance matrix $\begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix}$. If the allele a is significantly associated, showing a standard normal score $Z_a = z_0$, then given this association, the allele b will have a nonzero expected standard score, with the conditional distribution being $(Z_b | Z_a = z_0) \sim N(rz_0, 1 - r^2)$. The chance of b to achieve as significance level is $\Pr(N(rz_0, 1 - r^2) > z_0) = \phi(-z_0\sqrt{(1-r)/(1+r)})$. The events X_b and X_b of a and b achieving this significance level, respectively, thus have correlation

$$\begin{aligned}\rho(X_a, X_b) &= \frac{\mathrm{Cov}(X_a, X_b)}{\sqrt{\mathrm{Var}(X_a)\mathrm{Var}(X_b)}} \\ &= \frac{\phi(-z_0)\phi\left(-z_0\sqrt{\frac{1-r}{1+r}}\right) - \phi(-z_0)^2}{\phi(z_0)\phi(-z_0)} \\ &= 1 - \frac{\phi\left(z_0\sqrt{\frac{1-r}{1+r}}\right)}{\phi(z_0)}\end{aligned}$$

which is decreasing with Z_0. This means that the more significant the P-value, the lesser the correlation coefficient, or in other words, the lower the

significance the less correction for multiple testing the correlated tests require.

Fortunately, in a 2-stage design of a WGAS, the first stage is designed for a true positive to reach only a moderate *P*-value, expected to be achieved by 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. An alternative explanation is that this observed effect is an artifact of our oversampling design: Rather than simulating data by a true bootstrap procedure that samples real data without replacement for each simulated data set. We are simulating data sets of thousands of individuals based on 120 chromosomes only. We note that a similar result was not observed in a similar set of oversampling analyses [Dudbridge, 2006], suggesting attribution of this increased burden to the density and redundancy of ENCODE data we use.

These and other results offer considerable understanding of the distribution of null signals in idealized association studies. Practical association studies may exhibit more extreme *P*-values than 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 of 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.

ACKNOWLEDGMENTS

We thank Duncan Thomas for comments on an early draft of this manuscript. We further thank the anonymous reviewers of this manuscript for their insightful comments. I. P. was funded in part by National Center for the Multiscale Analysis of Genomic and Cellular Networks (MAGNet), NIH grant number 5 U54 CA121852.

REFERENCES

Arking DE, Pfeufer A, Post W, Kao WH, Newton-Cheh C, et al. 2006. A common genetic variant in the NOS1 regulator NOS1AP modulates cardiac repolarization. Nat Genet 38: 644–651.

Barrett J, Cardon L. 2006. Evaluating coverage of genome-wide association studies. Nat Genet 38:659–662.

Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR, et al. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447:799–816.

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

Dudbridge F. 2006. A note on permutation tests in multistage association scans. Am J Hum Genet 78:1094–1095; author reply 1096.

Dudbridge F, Koeleman BP. 2004. Efficient computation of significance levels for multiple associations in large studies of correlated data, including genomewide association studies. Am J Hum Genet 75:424–435.

Hirschhorn JN, Daly MJ. 2005. Genome-wide association studies for common diseases and complex traits. Nat Rev 6: 95–108.

Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, et al. 2005. Complement factor H polymorphism in age-related macular degeneration. Science 308:385–389.

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

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

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

Pe'er I, Chretien Y, PIW PdB, Barrett J, Daly M, et al. 2006a. Biases and reconciliation in estimations of linkage disequilibriumin the human genome. Am J Hum Genet 73.

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

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

Saxena R, Voight BF, Lyssenko V, Burtt NP, de Bakker PI, et al. 2007. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. Science 316: 1331–1336.

Scott LJ, Mohlke KL, Bonnycastle LL, Willer CJ, Li Y, et al. 2007. A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. Science 316:1341–1345.

Sidak Z. 1967. Rectangular confidence regions for the means of multivariate normal distributions. J Am Stat Assoc 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 38:209–213.

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

The International HapMap Consortium. 2003. The International HapMap Project. Nature 426:789–796.

The International HapMap Consortium. 2005. A haplotype map of the human genome. Nature 437:1299–1320.

Thomas D, Xie R, Gebregziabher M. 2004. Two-stage sampling designs for gene association studies. Genet Epidemiol 27: 401–414.

Wellcome Trust Case Control Consortium. 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447:661–678.

Zeggini E, Weedon MN, Lindgren CM, Frayling TM, Elliott KS, et al. 2007. Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. Science 316:1336–1341.