Genetic Epidemiology 32: 179–185 (2008)

Genome-Wide Significance for Dense SNP and Resequencing Data

Clive J. Hoggart,[1]* Taane G. Clark,[1,2] Maria De Iorio,[1] John C. Whittaker,[3] and David J. Balding[1]

[1]*Department of Epidemiology and Public Health, Imperial College London, Norfolk Place, London*
[2]*Current affiliation - Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive Oxford and Wellcome Trust Sanger Institute, Hinxton, Cambridge*
[3]*Non-communicable Disease Epidemiology Unit, London School of Hygiene and Tropical Medicine, Keppel Street, London*

The problem of multiple testing is an important aspect of genome-wide association studies, and will become more important as marker densities increase. The problem has been tackled with permutation and false discovery rate procedures and with Bayes factors, but each approach faces difficulties that we briefly review. In the current context of multiple studies on different genotyping platforms, we argue for the use of truly genome-wide significance thresholds, based on all polymorphisms whether or not typed in the study. We approximate genome-wide significance thresholds in contemporary West African, East Asian and European populations by simulating sequence data, based on all polymorphisms as well as for a range of single nucleotide polymorphism (SNP) selection criteria. Overall we find that significance thresholds vary by a factor of >20 over the SNP selection criteria and statistical tests that we consider and can be highly dependent on sample size. We compare our results for sequence data to those derived by the HapMap Consortium and find notable differences which may be due to the small sample sizes used in the HapMap estimate. *Genet. Epidemiol.* 32:179–185, 2008.

Key words: genome-wide association studies; multiple testing; statistical significance

Contract grant sponsor: The UK Medical Research Council.
*Correspondence to: Clive J. Hoggart, Department of Epidemiology and Public Health, Imperial College London, Norfolk Place, London.
E-mail: c.hoggart@imperial.ac.uk
Received 7 August 2007; Accepted 16 October 2007
Published online 28 December 2007 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/gepi.20292

INTRODUCTION

Given the large number of statistical tests (around 10^6) in genome-wide association (GWA) studies, and even larger numbers from future studies using whole-genome resequencing technology, it is important, but difficult, to accurately convey the significance of any reported associations. Frequently this problem is tackled by controlling the family-wise error rate (FWER), which is the probability of one or more significant results under the null hypothesis of no association. For a test applied at each of n single nucleotide polymorphisms (SNPs), the simplest way of approximating the per-test significance level α'' corresponding to a given FWER (α) is to apply a Bonferroni ($\alpha'' = \alpha/n$) or a Šidák ($\alpha'' = 1-(1-\alpha)^{1/n}$) correction [Šidák, 1968, 1971]. If the tests are mutually independent the Šidák correction is exact, but in practice the tests are dependent and both corrections can be very conservative.

More accurate approximations to the genome-wide significance level α'' are difficult to obtain, because the correlations between the tests depend on many factors, such as variations in linkage disequilibrium (LD) and SNP density, as well as the choice of test statistic(s) and sample size. The concept of an effective number of (independent) tests is appealing [Cheverud, 2001; Nyholt, 2004; The International HapMap Consortium, 2005], but this number is closely related to α'' and depends on the same factors [Salyakina et al., 2005; Dudbridge et al., 2006].

False discovery rate procedures [Benjamini and Hochberg, 1995; Storey and Tibshirani, 2003], which control the expected proportion of non-causal SNPs among those declared significant, have been proposed as an alternative to the FWER approach. These have been applied successfully to the analysis of genome-wide gene expression data, for which one typically expects many true positives. However, in the context of GWA studies, the smaller number of positives and the problem of LD between true positives and flanking SNPs, remain problematic [Dudbridge et al., 2006; Yang et al., 2005]. In principle, the use of Bayes factors to measure significance avoids the need for assessing genome-wide error rates. A full Bayesian analysis involves the difficult challenge of identifying a realistic yet tractable alternative model. For example, the alternative model of the Wellcome Trust Case Control

Consortium [2007] assumes the same distribution of effect sizes for all minor allele frequency (MAF) values of the causal variant, which may be considered unrealistic. In practice, Bayes factors are typically applied to one SNP at a time, and are often treated as frequentist test statistics for which error rates must be assessed.

Permutation procedures can yield the correct FWER even when the tests are dependent. They are computationally intensive in the setting of GWA studies, which has motivated approximations to reduce the computational effort [Dudbridge and Koeleman, 2004; Seaman and Muller-Myhsok, 2005; Kimmel and Shamir, 2006]. The results of permutation procedures apply only to the current genotyped dataset and must be recomputed when the dataset is altered, therefore they cannot provide truly genome-wide significance thresholds unless the entire genome is sequenced, neither can they be used before genotyping to derive significance thresholds for power calculation to aid the design of association studies.

For linkage analyses, universal thresholds based on the LOD score were established and have proven useful over many years [Lander and Kruglyak, 1995]. For association studies, a genome-wide 5% significance level was estimated [Risch and Merikangas, 1996] as 5×10^{-8}. This result has been influential, but it was derived using a Bonferroni correction assuming 10^6 independent tests (10^5 genes and 10 independent tests in each gene), and its sensitivity to different genomic and study design features has not been explored.

Using SNPs ascertained from the ENCODE Project Consortium [2004], which spanned 10 500-kb regions, the HapMap Consortium estimated the effective number of independent tests per 500 kb to be 350 in the West African population and 150 in European and East Asian populations when testing all SNPs with MAF $>5\%$. Assuming the length of the human genome to be 3,300 Mb this is equivalent to an effective number of tests in the genome of 9.9×10^5 in Europeans and East Asians and 23.1×10^5 in West Africans, which, using a Šidák correction, gives a genome-wide 5% significance threshold of 2.2×10^{-8} for Africans and 5.2×10^{-8} for Europeans and East Asians. These estimates were based on SNPs ascertained by sequencing 48 individuals and genotyped in 209 unrelated individuals. Mock case-control samples were then generated by resampling-phased chromosomes of these individuals. The ENCODE data are not suitable to calculate thresholds for tests of SNPs with MAF $<5\%$ due to the small sample size used to ascertain SNPs. Furthermore, to generate sufficient sizes of mock case-control samples from the ENCODE data, one would have had to resample chromosomes many times, resulting in less genetic diversity than in the actual population.

METHODS

To overcome the limitations of the approaches discussed above, we simulated sequence level data in large populations enabling all polymorphisms to be ascertained, and individuals to be sampled without replacement. The simulation was implemented using the FREGENE software [Hoggart et al., 2007], which simulates large genomic regions in large diploid populations, forward in time, under various demographic scenarios and evolutionary models, including variable recombination rates and gene conversion. In our simulation, we mimicked the modeling assumptions of [Schaffner et al., 2006] in which the demographic and evolutionary model of the simulation was chosen to approximate the history of populations from three continental regions: West Africa, East Asia and Europe. The genetic parameters of the simulation were tuned [Schaffner et al., 2006] to match (1) the allele frequency distribution; (2) the relationship between allele frequency and the probability that an allele is ancestral; (3) F_{st}, and two measures of the extent of LD; (4) the relationship of genetic distance with r^2 and (5) the fraction of pairs of markers with $D' = 1$ in each of the three populations [Schaffner et al., 2006]. Our simulation reproduced the reported results for these five criteria [Schaffner et al., 2006]. Further details of the simulation study are given in the Appendix.

It is infeasible to simulate entire human chromosomes in such large populations, however with our computer memory constraints we were able to simulate regions of 5 Mb and hence we in effect approximated the human genome by 660 chromosomes each of 5 Mb. Genome-wide values were then inferred from the simulated chromosome using Šidák's correction. Our checks indicate that the error in this approximation is negligible (see Appendix for justification). From the three simulated populations, we selected chromosome pairs corresponding to individuals for inclusion in the association studies. We then permuted the case-control labels 5×10^5 times, and recorded the minimum *P*-value in order to approximate the per-SNP significance level corresponding to a FWER of $\alpha = 5$, 10, 20, and 50%. We explored the effect of SNP ascertainment, test statistic and the case/control sample size. Specifically we considered four SNP-ascertainment strategies:

1. SNPs chosen randomly with an average spacing of 5 kb and an approximate uniform MAF distribution in Europeans approximating an

Affymetrics 500 K GeneChip — we denote this as our "standard" study design.
2. All polymorphic sites.
3. All sites with MAF $>0.5\%$ in the relevant population.
4. All sites with MAF $>5\%$ in the relevant population.

We considered three sample sizes: 100, 1,000 and 5,000 cases, with equal numbers of controls. In all analyses, a Cochran-Armitage trend test (sensitive to additive effects only) was applied. In addition Pearson's 2 degree of freedom (df) test and separate tests for dominant and recessive effects were applied to the standard SNP ascertainment with 1,000 cases and controls. Furthermore, when these additional tests were applied the additive, dominant and recessive tests were combined at each SNP by taking the maximum of the three test statistics [MAX test; Freidlin et al., 2002]. The additive and the 2 df tests were combined similarly. See the Appendix for full details of the calculation of genome-wide significance thresholds.

RESULTS

Tables I, II, III show our estimates of the per-SNP significance level (α'') for the corresponding FWER (α) of 5, 20 and 50% for the simulated West African, East Asian and European populations. The results for $\alpha = 10\%$ are not shown; they can be approximated by a linear interpolation between the $\alpha = 5$ and 20% values. The effective number of tests was calculated as the number of independent tests that would give the calculated threshold α'' at the significance level $\alpha = 20\%$ using the Šidák correction.

For all of the simulated GWA studies, our estimate of α'' is larger than would be estimated from a genome-wide Šidák correction, because the effective number of tests is always less than the actual number of tests. For our standard GWA the lowest significance thresholds were observed in Europeans. This may appear surprising because lower LD in the West African population leads us to expect less dependence of the tests and so lower thresholds than among Europeans. However, the SNPs were ascertained on the basis of European allele frequencies which resulted in 145 and 258 of these SNPs being monomorphic in the West African and East Asian populations respectively, whereas all were polymorphic in Europeans. More generally, SNPs typically have a higher MAF in Europeans than in the other two populations and SNPs with low MAF are less likely to be significant. This is because for rare variants the null distribution of *P*-values is non-uniform, having less weight at small *P*-values.

If the chromosomes are sequenced and all polymorphic sites tested, the ascertainment issue vanishes and, as expected, α'' is lowest in the West African population. With 5,000 cases and controls the ratios of α'' for sequence data to its value under the standard SNP map for a genome-wide significance level of 20% are 23.8, 24.2 and 12.6 in the West African, East Asian, and European populations respectively. The threshold α'' increases as the SNP density decreases from MAF >0.5 to $>5\%$, but as before α'' is lowest in the African population.

With 1,000 cases and controls, for all SNP ascertainments other than the standard, α'' is roughly twice as large in East Asian and European populations in comparison to West Africans, but the difference is less for 5,000 cases and controls when testing all polymorphisms and for MAF $>0.5\%$. The

TABLE I. West African population: significance thresholds and effective number of tests (each with 95% confidence interval)

	Per-test siginificance level $\alpha'' \times 10^{-8}$ for family-wise error rate $\alpha =$			No. of tests in genome $\times 10^6$		
	5%	20%	50%	Effective	Actual	Ratio
Standard						
100 Cases/controls	34 (23–41)	130 (110–150)	390 (350–430)	0.17 (0.15–0.2)	0.66	0.26
1,000 Cases/controls	14 (7.2–18)	61 (52–70)	210 (180–220)	0.37 (0.32–0.43)		0.56
5,000 Cases/controls	16 (10–19)	62 (55–71)	180 (170–190)	0.36 (0.31–0.4)		0.55
MAF $>5\%$						
1,000 Cases/controls	1.6 (1–2)	6.7 (5.9–7.3)	21 (19–22)	3.3 (3–3.8)	9.8	0.34
5,000 Cases/controls	1.5 (1–1.9)	6.0 (4.8–6.8)	18 (16–20)	3.7 (3.3–4.6)		0.38
MAF $>0.5\%$						
1,000 Cases/controls	0.98 (0.69–1.6)	4.9 (4.2–5.7)	15 (13–16)	4.5 (3.9–5.3)	21.8	0.21
5,000 Cases/controls	0.66 (0.51–0.85)	2.6 (2.3–3)	8.0 (7.4–8.7)	8.7 (7.4–9.9)		0.40
All polymorphisms						
1,000 Cases/controls	0.92 (0.68–1.4)	4.0 (3.4–4.7)	13 (12–14)	5.6 (4.7–6.6)	120	0.047
5,000 Cases/controls	0.65 (0.32–0.88)	2.6 (2.4–3)	7.4 (6.8–8)	8.5 (7.5–9.5)		0.071

MAF, minor allele frequency.

TABLE II. East Asian population: significance thresholds and effective number of tests (each with 95% confidence interval)

	Per-test siginificance level $\alpha'' \times 10^{-8}$ for family-wise error rate $\alpha =$			No. of tests in genome $\times 10^6$		
	5%	20%	50%	Effective	Actual	Ratio
Standard						
100 Cases/controls	41 (26–58)	160 (140–180)	450 (420–480)	0.14 (0.12–0.16)	0.66	0.21
1,000 Cases/controls	17 (14–23)	82 (69–95)	280 (250–310)	0.27 (0.23–0.33)		0.41
5,000 Cases/controls	19 (14–24)	85 (71–97)	240 (220–260)	0.26 (0.23–0.32)		0.39
MAF > 5%						
1,000 Cases/controls	3.0 (2.1–4.3)	13 (10–14)	37 (33–41)	1.8 (1.6–2.1)	6.7	0.27
5,000 Cases/controls	2.7 (2.1–3.6)	12 (9.5–13)	37 (33–41)	1.9 (1.7–2.3)		0.28
MAF > 0.5%						
1,000 Cases/controls	2.5 (1.8–3.1)	9.5 (8.2–10)	25 (22–26)	2.4 (2.1–2.7)	13.7	0.18
5,000 Cases/controls	1.3 (0.79–1.9)	5.2 (4.4–5.9)	14 (13–15)	4.3 (3.8–5)		0.31
All polymorphisms						
1,000 Cases/controls	2.1 (1.4–3.2)	9.6 (8.5–11)	28 (26–31)	2.3 (2–2.6)	116	0.02
5,000 Cases/controls	0.86 (0.6–1.3)	3.5 (3.1–4.1)	10 (9.2–12)	6.4 (5.4–7.2)		0.055

MAF, minor allele frequency.

TABLE III. European population: significance thresholds and effective number of tests (each with 95% confidence interval)

	Per-test siginificance level $\alpha'' \times 10^{-8}$ for family-wise error rate $\alpha =$			No. of tests in genome $\times 10^6$		
	5%	20%	50%	Effective	Actual	Ratio
Standard						
100 Cases/controls	29 (19–41)	120 (110–140)	330 (290–360)	0.19 (0.17–0.21)	0.66	0.29
1,000 Cases/controls	15 (12–18)	62 (52–74)	180 (170–200)	0.36 (0.3–0.43)		0.55
5,000 Cases/controls	11 (7.2–14)	44 (38–52)	150 (130–160)	0.51 (0.43–0.59)		0.77
MAF > 5%						
1,000 Cases/controls	2.1 (1.4–3.4)	11 (9.2–12)	34 (32–38)	2.1 (1.9–2.4)	7.4	0.28
5,000 Cases/controls	3.1 (2.2–3.9)	10 (9–12)	31 (29–34)	2.2 (1.9–2.5)		0.3
MAF > 0.5%						
1,000 Cases/controls	1.4 (1.1–1.9)	7.4 (6.2–8.3)	23 (20–24)	3.0 (2.7–3.6)	14.2	0.21
5,000 Cases/controls	1.3 (0.64–1.7)	5.2 (4.6–5.8)	14 (13–15)	4.3 (3.8–4.9)		0.3
All polymorphisms						
1,000 Cases/controls	1.8 (1.3–2.4)	7.1 (6–8.4)	25 (22–26)	3.1 (2.6–3.7)	116	0.027
5,000 Cases/controls	0.69 (0.41–0.99)	3.5 (2.8–4)	9.8 (8.8–11)	6.5 (5.6–7.9)		0.056

MAF, minor allele frequency.

smaller α'' in the West African population can be explained by two factors arising from its larger effective population size: lower average LD between SNPs, and hence tests at neighboring SNPs are less dependent, and a greater number of polymorphisms. We can see the effect of the greater LD in the East Asian and European populations in comparison with the West African population by the lower ratio of effective number of tests to actual number of tests.

Typically α'' decreases as the sample size increases, for example an increase in sample size from 1,000 to 5,000 cases and controls for MAF $>0.5\%$ and all polymorphic sites ascertainment schemes, resulted in the required α'' decreasing by a factor of two to three and the effective number of tests approximately doubling. This is because increased sample size changes the distribution of the test statistic at rare variants to allow smaller *P*-values. The exception is with SNP ascertainment MAF $>5\%$ when increasing the sample size from 1,000 to 5,000; this has no effect on α'' because of the absence of rare alleles.

Table IV compares the estimates of the per-SNP significance level (α'') for the corresponding FWER for a range of single SNP tests. We see that testing for a dominant effect gives approximately the same threshold as for an additive effect. However, when testing for recessive effects the threshold is higher,

TABLE IV. European population with 1,000 cases and controls: significance thresholds and effective number of tests for different test statistics (each with 95% confidence interval)

	Per-test siginificance level $\alpha'' \times 10^{-8}$ for family-wise error rate $\alpha =$			No. of tests in genome $\times 10^6$		
	5%	20%	50%	Effective	Actual	Ratio
Standard						
Additive test	15 (12–18)	62 (52–74)	180 (170–200)	0.36 (0.3–0.43)	0.66	0.55
Dominant test	12 (9.1–19)	61 (51–70)	180 (160–200)	0.37 (0.32–0.44)	0.66	0.56
Recessive test	21 (14–29)	93 (80–100)	280 (260–300)	0.24 (0.21–0.28)	0.66	0.36
MAX test	7.1 (4.4–8.2)	27 (23–32)	87 (79–93)	0.83 (0.69–0.99)	2	0.42
2 df test	18 (12–25)	75 (63–89)	220 (210–240)	0.30 (0.25–0.35)	0.66	0.45
Both additive and 2 df tests	8.1 (5.1–12)	40 (34–47)	120 (110–130)	0.55 (0.48–0.66)	1.3	0.42
MAF > 20%						
Additive test	4.6 (2.5–6.5)	21 (19–24)	66 (61–72)	1 (0.92–1.1)	3.8	0.26
Dominant test	3.3 (2.1–4.6)	18 (15–21)	57 (52–63)	1.2 (1–1.5)	3.8	0.32
Recessive test	4.7 (3.4–5.9)	21 (18–24)	65 (60–68)	1 (0.93–1.3)	3.8	0.26
MAX test	0.98 (0.51–1.7)	6.6 (5.6–7.5)	22 (21–25)	3.4 (3–4)	12	0.28
2 df test	4.7 (3.4–6)	18 (16–24)	62 (56–67)	1.2 (0.93–1.4)	3.8	0.32
Both additive and 2 df tests	2.5 (1.7–3.4)	12 (10–14)	35 (33–39)	1.9 (1.6–2.1)	7.7	0.25

df, degree of freedom; MAF, minor allele frequency.

because the counts of minor-allele homozygotes in cases and controls will be small for rare alleles. Thus, the resulting *P*-value cannot be as small as for the tests of trend or dominance. To check this argument, we evaluated the required thresholds for the three tests applied at all SNPs with MAF > 20%; they were approximately the same. These results indicate that the required threshold for a given FWER is dependent on the test employed only through the minimum genotype frequency. Thus, if the MAF and sample size are large enough such that genotype counts in cases and controls are sufficiently large (as a rule of thumb > 5) different tests do not give rise to different thresholds.

The effective number of tests for the MAX statistic is 8.3×10^5, which is only slightly less than the sum of the effective number of tests when the three statistics are tested individually (9.7×10^5). Although the three-test statistics are highly correlated, the genomic locations at which their minima occur are close to independent. Similarly, the additive and 2 df tests behave as if nearly independent when combined, even though the two test statistics are highly correlated. Thus, there is a substantial penalty in terms of type 1 error for applying multiple tests at a SNP.

DISCUSSION

We provide approximate significance thresholds for whole-genome association studies for West African, East Asian and European populations with various study designs and single-SNP analyses. These results provide experimenters with guidelines for declaring genome-wide significance in present studies and future studies with sequence data. With the standard SNP ascertainment and 1,000 cases and controls the effective number of tests is about half the actual number of tests, thus the Šidák correction would underestimate the required threshold by one half, with 5,000 cases and controls the Šidák correction is less conservative. These significance thresholds are also relevant to meta-analyses in which test statistics are derived by combining results from independent studies.

SNP ascertainment has the largest effect among the factors considered. With 5,000 cases and controls results vary by a factor of up to 24 between testing all polymorphisms and our standard analysis which has an average SNP spacing of 5 kb. Increasing sample size increases the effective number of tests, and hence reduces the required significance threshold, when many rare variants are tested, because of increased ability to detect rare variants with larger sample sizes.

The HapMap Consortium estimated the effective number of tests in the genome if all SNPs with MAF > 5% were tested to be 9.9×10^5 in Europeans and East Asians and 23×10^5 in West Africans, but the sample size was not stated. Our equivalent estimates are 21×10^5 in Europeans, 18×10^5 in East Asians and 33×10^5 in West Africans, for a sample size of 1,000 cases and 1,000 controls. No standard deviation was reported for the HapMap estimates, and so we cannot assess whether our higher estimates are in fact significantly different, but if so the discrepancy could be due to the difference in numbers of cases and controls, or polymorphisms being missed by

ENCODE due to the small sample sizes used to ascertain SNPs.

If several experimenters perform analyses for the same phenotype on different genotyping platforms, each may choose to control their FWER separately, taking into account only the SNPs tested. However, for the wider public the relevant FWER is for the combined datasets of all relevant studies and not the number of tests performed by any one research group. Similarly, the Wellcome Trust Case Control Consortium (WTCCC) argue for controlling the type 1 error rate, but note that ''one should not correct significance levels for the number of tests performed''. We propose universal genome-wide significance thresholds based on all polymorphisms, whether or not typed, to achieve these goals without the challenging task of specifying a tractable yet realistic alternative model. The WTCCC used a significance threshold of 5×10^{-7}, which from Table 3 corresponds to a genome-wide false positive rate of above 50%. If only common variants are tested the threshold calculated for MAF $>5\%$ could be justified. This threshold can also be thought of as a lower bound when SNPs have been ascertained by tagging using HapMap data because there is little power with these SNPs to detect variants with MAF $<5\%$ [Zeggini et al., 2005]. However the threshold for all polymorphic sites would provide a more stringent threshold that will become increasingly appropriate as SNP densities increase and more low MAF SNPs are tested.

ACKNOWLEDGMENTS

C.H. and T.C. are funded by a grant from the UK Medical Research Council under the Link Applied Genomics scheme. The authors thank Toby Andrew, Frank Dudbridge and Paul O'Reilly for helpful comments.

REFERENCES

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate – a practical and powerful approach to multiple testing. J Roy Stat Soc Ser B 57:289–300.

Cheverud JM. 2001. A simple correction for multiple comparisons in interval mapping genome scans. Heredity 87:52–58.

Dudbridge F, Koeleman BPC. 2004. Efficient Computation of Significance Levels for Multiple Associations in Large Studies of Correlated Data, Including Genome-wide Association Studies. Am J Hum Genet 75:424–435.

Dudbridge F, Gusnanto A, Koeleman BPC. 2006. Detecting Multiple Associations in Genome-wide Studies. Hum Genom 2:310–317.

Freidlin B, Zheng G, Li ZH, Gastwirth JL. 2002. Trend tests for case-control studies of genetic markers: Power, sample size and robustness. Hum Hered 53:146–152.

Hoggart CJ, Chadeau-Hyam M, Clark TG, Lampariello R, Whittaker JC, De Iorio M, Balding DJ. 2007. Sequence-level population simulations over large genomic regions. Genetics: in press; doi: 10.1534/genetics.106.069088.

Kimmel G, Shamir R. 2006. A fast method for computing high-significance disease association in large population-based studies. Am J Hum Genet 79:481–492.

Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson B, Sigurdardottir S, Barnard J, Hallbeck B, Masson G, Schlein A, Palsson ST, Frigge ML, Thorgeirsson TE, Gulcher JR, Stefansson K. 2002. A high-resolution recombination map of the human genome. Nat Genet 31:241–247.

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

Nyholt DR. 2004. A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. Am J Hum Genet 74:765–769.

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

Salyakina D, Seaman SR, Browning BL, Dudbridge F, Muller-Myhsok B. 2005. Evaluation of Nyholt's procedure for multiple testing correction. Hum Hered 60:19–25.

Seaman SR, Muller-Myhsok B. 2005. Rapid simulation of *P* values for product methods and multiple-testing adjustment in association studies. Am J Hum Genet 6:399–408.

Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, Altshuler D. 2006. Calibrating a coalescent simulation of human genome sequence variation. Genome Res 15:1576–1583.

Storey JD, Tibshirani R. 2003. Statistical significance for genome-wide studies. Proc Natl Acad Sci USA 100:9440–9445.

Šidák Z. 1968. On multivariate normal probabilities of rectangles: their dependence on correlations. Ann Math Statist 39: 1425–1434.

Šidák Z. 1971. On probabilities of rectangle in multivariate normal Student distributions: their dependence on correlations. Ann Math Statist 41:169–175.

The ENCODE Project Consortium. 2004. The ENCODE (Encyclopedia Of DNA Elements) Project. Science 306: 636–640.

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

The 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.

Yang Q, Cui J, Chazaro I, Cupples LA, Demissie S. 2005. Power and type I error rate of false discovery rate approaches in genome-wide association studies. BMC Genet 6:(Suppl 1):S134.

Zeggini E, Rayner W, Morris AP, Hattersley AT, Walker M, Hitman GA, Deloukas P, Cardon LR, McCarthy MI. 2005. An evaluation of HapMap sample size and tagging SNP performance in large-scale empirical and simulated data sets. Nat Genet 37:1320–1322.

APPENDIX

SIMULATION STUDIES

Here we describe in more detail the simulation model implemented which was formulated by Schaffner et al. [2006]. In the simulation, a homogeneous population evolved before splitting in two, one of which represents a small population that migrated out of Africa. The two populations then continued to evolve for 3,500 generations before the

non-African population split into two mimicking ancestral Asian and European populations. The three populations continued to evolve for a further 2,000 generations before undergoing a final expansion. Population bottlenecks were implemented immediately after the population splits in all subpopulations. Migration occurred between the African and European populations and the African and Asian populations. A constant mutation rate of 1.5×10^{-8} per site and per generation and a constant gene conversion rate of 4.5×10^{-9} per site and per generation with a tract length of 500 bases were used. The crossover rate was variable and the model for the variation in rates was hierarchical. An average regional rate was set based on the deCODE genetic map [Kong et al., 2002], local rates were then set stochastically and finally hotspots were sampled conditional on the local rate. The intensities and spacing of the hotspots were both stochastic, for further details of the genetic and demographic model see Schaffner et al. [2006].

CALCULATING GENOME-WIDE SIGNIFICANCE THRESHOLDS

Under the null hypothesis that no SNP is associated with case-control status, the per-SNP significance level α'' that corresponds to an FWER of α satisfies $\alpha = \Pr(\min\{p_i\} < \alpha'')$, where p_i denotes the *P*-value from the *i*th SNP. Computing α'' directly is difficult because of the large number of SNPs and their complex correlation structure. Instead, we approximate α'' via simulation of 5-Mb chromosomes and assume the 3.3-Gb genome is made up of 660 independent 5-Mb chromosomes. Thus, for a FWER of $\alpha = 5\%$ for a genome-wide study, a Šidák correction, with 660 independent tests, requires a significance level of $\alpha' = 0.0078\%$ within each 5-Mb chromosome. Thus, we estimate α'' by the 0.0078% point of the distribution of the minimum *P*-value in a 5-Mb interval, using n permutations of the case-control labels (here $n = 5 \times 10^5$). In other words, we use $qn\alpha'$ as an estimate of α'' where qi denotes the ith smallest *P*-value arising from the n permutations of case-control labels. An approximate confidence interval for α'' can be obtained by treating the true position of the α'-percentile of q as a Binomial random variable with parameters n and α'. Then, using the normal approximation to the Binomial, we obtain the 95% confidence limits $q\{n\alpha' - 1.96\sqrt{n\alpha'(1-\alpha')}\}$ and $q\{n\alpha' + 1.96\sqrt{n\alpha'(1-\alpha')}\}$.

A genome of 660 5-Mb chromosomes will have overall less LD than the actual human genome, which may bias upwards the estimate of the number of independent tests. To investigate the effect of this assumption, we simulated a 20-Mb region using the same evolutionary parameters but a simpler demographic model in which there was no migration but a single homogeneous population of 10,000 individuals. We compared the estimated genome-wide significance level for this population using the entire 20-Mb region and also using 10-Mb and 5-Mb subintervals of the 20-Mb region. We found that while the point estimates derived from the 5-Mb regions were slightly lower than those derived from the 20-Mb regions, the 95% confidence intervals overlapped considerably. Furthermore, the 95% confidence intervals based on 10 and 20-Mb regions were practically indistinguishable suggesting that effect of region size has ''flattened'' off by this point.

For computational efficiency, for each simulation model we permuted the case-control labels from just one sample from the population, rather than repeatedly re-sampling cases and controls from the population. The lack of sample replication is not a problem for regions as large as 5-Mb, because there is in effect internal replication within the region. To verify this, we took 10 independent case-control samples from the standard simulation. We found that estimates of α'' from each were not significantly different at the 5% level.