Hum Genet (2012) 131:747–756
DOI 10.1007/s00439-011-1118-2

ORIGINAL INVESTIGATION

Evaluating the effective numbers of independent tests and significant *p*-value thresholds in commercial genotyping arrays and public imputation reference datasets

Miao-Xin Li · Juilian M. Y. Yeung · Stacey S. Cherny · Pak C. Sham

Received: 23 August 2011 / Accepted: 13 November 2011 / Published online: 6 December 2011

Abstract Current genome-wide association studies (GWAS) use commercial genotyping microarrays that can assay over a million single nucleotide polymorphisms (SNPs). The number of SNPs is further boosted by advanced statistical genotype-imputation algorithms and large SNP databases for reference human populations. The testing of a huge number of SNPs needs to be taken into account in the interpretation of statistical significance in such genome-wide studies, but this is complicated by the non-independence of SNPs because of linkage disequilibrium (LD). Several previous groups have proposed the use of the effective number of independent markers (M_e) for the adjustment of multiple testing, but current methods of calculation for M_e are limited in accuracy or computational speed. Here, we report a more robust and fast method to calculate M_e. Applying this efficient method [implemented in a free software tool named Genetic type 1 error calculator (GEC)], we systematically examined the M_e, and the corresponding *p*-value thresholds required to control the genome-wide type 1 error rate at 0.05, for 13 Illumina or Affymetrix genotyping arrays, as well as for HapMap Project and 1000 Genomes Project datasets which are widely used in genotype imputation as reference panels. Our results suggested the use of a *p*-value threshold of $\sim 10^{-7}$ as the criterion for genome-wide significance for early commercial genotyping arrays, but slightly more stringent *p*-value thresholds $\sim 5 \times 10^{-8}$ for current or merged commercial genotyping arrays, $\sim 10^{-8}$ for all common SNPs in the 1000 Genomes Project dataset and $\sim 5 \times 10^{-8}$ for the common SNPs only within genes.

Electronic supplementary material The online version of this article (doi:10.1007/s00439-011-1118-2) contains supplementary material, which is available to authorized users.

M.-X. Li · J. M. Y. Yeung · S. S. Cherny · P. C. Sham
Department of Psychiatry, The University of Hong Kong, Pokfulam, Hong Kong

M.-X. Li · P. C. Sham
The Centre for Reproduction, Development and Growth, The University of Hong Kong, Pokfulam, Hong Kong

M.-X. Li · P. C. Sham
Genome Research Centre, The University of Hong Kong, Pokfulam, Hong Kong

S. S. Cherny · P. C. Sham
State Key Laboratory for Cognitive and Brain Sciences, The University of Hong Kong, Pokfulam, Hong Kong

P. C. Sham (✉)
Department of Psychiatry, LKS Faculty of Medicine, University of Hong Kong, Pokfulam, Hong Kong
e-mail: pcsham@hkucc.hku.hk

Introduction

Genome-wide association studies (GWAS) can now directly assay up to 2.5 million single nucleotide polymorphisms (SNPs) using high-throughput genotyping arrays (Ragoussis 2009). The number of SNPs is further boosted by statistical genotype-imputation algorithms that make use of large SNP reference datasets such as the HapMap Project and 1000 Genomes Project (Anderson et al. 2008; Howie et al. 2009). The number of SNPs is set to increase further with recent advances in resequencing technology (Metzker 2010). The testing of such huge numbers of SNPs results in a massive multiple-testing burden in statistical analysis.

The Bonferroni correction, which resets the significance threshold from α to α/M in the presence of M independent tests, is probably the most popular method for multiple-testing adjustment. However, the Bonferroni correction assumes independence among the tests considered, so that

it is inherently conservative when considering SNPs in linkage disequilibrium (LD). Adjustment for multiple testing by permutation appropriately takes account of marker dependency and results in a more powerful test (Pahl and Schafer 2010), but is computationally expensive. There have been a number of attempts to extend the conventional Bonferroni procedure to handle correlated tests, by replacing the actual number of markers being tested (M) by a smaller value called the effective number of independent markers (M_e). This results in a test-wise significance threshold of $\alpha' = \alpha/M_e$, which controls the family-wise error rate (FWER) at α. Conversely, the test-wise error rate α' is related to the family-wise error by $\alpha = 1 - (1 - \alpha')^{M_e} \approx M_e\alpha'$ Efforts were made to assess the genome-wide significance thresholds after Bonferroni correction for early GWAS (Dudbridge and Gusnanto 2008; Pe'er et al. 2008). However, it is not known whether these thresholds are still applicable to current or future GWAS in which much more SNPs are assayed.

Several methods have been proposed for estimating M_e from the correlations between the genetic markers. Duggal et al. (2008) suggested the simple method of counting 1 SNP per LD block in addition to all the SNPs outside of blocks. Other proposed methods involved the eigenvalues of the LD measure r^2 or Pearson correlation matrix of allele counts calculated from all possible pairs of SNPs (Cheverud 2001; Gao et al. 2008; Li and Ji 2005; Nyholt 2004; Galwey 2009). Two of these methods used the variance of the eigenvalues (λ) to estimate M_e (Cheverud 2001; Nyholt 2004). An important limitation of these variance-based approaches is that they do not result in additive M_e estimates across contiguous sets of SNPs. Li and Ji (2005) suggested summing the eigenvalues, after substituting 1 for the eigenvalues that are greater than 1. While generally more accurate than the variance-based approaches, this method can be both conservative and liberal in different situations (Li and Ji 2005). Gao et al. suggested defining M_e as the number of eigenvalues which can explain $C\%$ of the variation for SNP genotype data. However, it is unclear how C should be set, as overly large or small value of C would result in an FWER that is overly conservative or liberal, respectively (2008). Galwey (2009) proposed a measure of M_e based on an eigenvalue ratio function. Moskvina and Schmidt suggested a formula to approximate M_e based on the conditional probability of a Type 1 error in one marker given the test outcome of a second marker (Moskvina and Schmidt 2008). Several studies have concluded that the available measures of M_e were not sufficiently accurate as a valid substitute for a permutation procedure (Han et al. 2009; Salyakina et al. 2005; Galwey 2009).

Here we propose a new method to more accurately and rapidly estimate the effective number of independent tests, M_e, from a given set of SNPs. The ratio of M_e to the actual number of SNPs in a genotyping array is suggested as an index of the tagging efficiency of an array. Extensive simulation studies based on both artificial and real LD patterns were conducted to compare the performance of this method against five alternative approaches. We then systematically investigated the M_e for 13 popular commercial genotyping arrays from Illumina and Affymetrix, as well as for the HapMap Project and 1000 Genomes Project genotype datasets which are widely used as reference panels in genotype imputation. From this, we provide a series of suggested Bonferroni p-value thresholds to correct for the multiple-testing burden in different populations, when using these arrays and imputed datasets.

Methods and materials

Construction of a new measure of the effective number of independent tests

Our method is similar to that of Li and Ji (2005), except that the used eigenvalues are those of the correlation matrix of association test p values, rather than the correlation matrix of allele counts, between SNPs. In a previous paper (Li et al. 2011), we described a polynomial approximation that allows the correlation matrix of association test p values to be calculated from the correlation matrix of allele counts. If the eigenvalues of the correlation matrix of M association test p values are denoted by λ_i, then the effective number of tests, M_e is estimated to be $M - \sum_{i=1}^{M}[I(\lambda_i > 1)(\lambda_i - 1)]$, where $I(x)$ is an indicator function. The second part of this formula estimates the redundant number of tests as a result of marker dependency. The p-value threshold to control FWER to α, using M_e in a Bonferroni procedure, would then be α/M_e. The ratio $R_e = M_e/M$, called "effective ratio" for convenience, measures the extent that the M markers are non-redundant.

A divide-and-conquer algorithm was developed to speed up the calculation of eigenvalues of large correlation matrices. SNPs on a chromosome can be partitioned into multiple loose LD blocks. Within a block, a SNP has strong or moderate LD with at least one other SNP while SNPs in different LD blocks are in weak LD (say, $r^2 < 0.1$). Theoretically, assume a large correlation matrix, $P = \begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix}$, has an eigenvalue, λ, and an associated eigenvector $\begin{bmatrix} X \\ Y \end{bmatrix}$. According to the definition of eigenvalue, $\begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix}\begin{bmatrix} X \\ Y \end{bmatrix} = \begin{bmatrix} AX \\ BY \end{bmatrix} = \begin{bmatrix} \lambda X \\ \lambda Y \end{bmatrix}$. Therefore, $AX = \lambda X$ and

$BY = \lambda Y$. This indicates that matrixes P, A and B share the same eigenvalues and the LD block partition will not change the eigenvalues and thus the resulting M_e, provided the blocks are independent. The M_e of whole genome is equal to the summation of the M_e calculated within each LD block. This divide-and-conquer strategy substantially speeds up the analysis by avoiding calculating eigenvalues in a huge matrix with thousands of rows and columns, although, in principle, if blocks cannot be formed the proposed measure of M_e could still be implemented.

Datasets

Local genotype dataset

In the simulation study, we used a genotype dataset of 2,514 Chinese subjects typed by the Illumina Human610-Quad BeadChip. This sample was originally prepared for several independent disease-gene mapping projects [(Kung et al. 2010) and unpublished data]. After standard quality control procedures of GWAS scan for common variants, 473,931 SNPs were left in the simulation analysis.

HapMap LD dataset

We downloaded the latest version of pair-wise LD data (r^2) of the 11 HapMap panels (http://hapmap.ncbi.nlm.nih.gov/downloads/ld_data/latest/, Release 27). For the JPT, CHB, CEU and YRI panels, this release merged SNPs of phases I + II + III and had more SNPs than other 7 panels which entered the HapMap Project at phase III. Therefore, we used the LD data of the 4 panels to derive the *Me* on the 13 commercial genotyping arrays. The numbers of unique SNPs contained in the 4 LD dataset for JPT, CHB, CEU and YRI panels were 2,509,881, 2,554,939, 2,776,528 and 3,114,362, respectively. But to provide a reference for GWAS imputation in more populations, we estimated the M_e and corresponding p-value thresholds in all of the 11 panels as well.

1000 Genomes Project genotype dataset

We downloaded genotypes of 1000 Genomes Project (released by August 2010) from the website of MACH (http://www.sph.umich.edu/csg/abecasis/MACH/download/). In this dataset, there were total 651 individuals separated in three different panels according to ancestry, ASN (Asian, 194), EUR (European, 283), and AFR (African, 174). The numbers of overall SNPs in the three panels are 10,832,281 (ASN), 11,914,767 (EUR), and 17,042,857 (AFR), respectively. However, only around half of the SNPs have the minor allele frequencies over 0.05. We estimated the M_e among SNPs with minor allele frequencies $\geq$0.05 because SNPs with too small minor allele frequency are generally underpowered in GWAS.

Examining the relationship between LD r^2 and correlation of p values from association tests

Genotype data of two bi-allelic SNPs were simulated for a number of subjects, for a set of LD coefficients, r, and allele frequencies, under Hardy–Weinberg equilibrium. For a case–control study, we randomly assigned disease status to generate 3,000 cases and 3,000 controls; for a quantitative trait study the 6,000 subjects were randomly given phenotypic scores sampled from the standard normal distribution $N(0, 1)$. That is, we simulated no correlation between trait/disease and genotype. An allelic association test was then performed for each of the two SNPs in the case–control study and the Wald test for parameters in a linear regression model was used to examine association in the quantitative trait study. The procedure was repeated 100,000 times to obtain 100,000 sets of p values, from which the correlation coefficient of the p values of the two SNPs,ρ, was calculated. The allele frequencies and the LD coefficients, r (Hill and Robertson 1968), were incremented in steps of 0.05 to generate a series of data points. Repeated simulations using samples of different sizes were also conducted.

The relationship between LD r^2 and p-value correlation coefficients was extrapolated by least-squares fitting using a 6th order polynomial function of the squared pair-wise allelic correlation coefficient,r'^2, in Microsoft Excel 2007. We found that under the null hypothesis p-value correlation coefficient,ρ, can be accurately approximated by a 6th order polynomial function of the squared pair-wise allelic correlation coefficient r^2 (coefficient of determination $R^2 = 0.9987$) (Supplementary Fig. 1), regardless of allele frequencies, sample size and study design.

Comparison of type 1 error of various measures by simulation and permutation

Given the LD patterns and allele frequencies (see supplementary Table 1), a program based on the HapSim algorithm (Montana 2005) was written to generate genotype data under Hardy–Weinberg equilibrium. We simulated regions with 1 LD block (6 SNPs), 2 LD blocks (10 SNPs), 6 LD blocks (30 SNPs) or 24 LD blocks (120 SNPs). We considered the null model where no SNP had an effect on disease risk. For each scenario, a population of 4,000,000 individuals was generated. A random sample of 3,000 cases and 3,000 controls was drawn from the population, without replacement, and subjected to the different methods of multiple testing. Type 1 error rates under the different

scenarios were obtained from the proportion of simulated datasets that resulted in at least one significant p value (set at 0.05), from 1,000 simulated populations.

We compared the performance of the proposed measure to 4 different estimates of M_e as well as the conventional Bonferroni correction approach. A permutation procedure was also carried out for the comparison. The four previous proposed M_e measures have been described in the "Introduction". In the permutation procedure, the phenotypes of subjects were permuted 1,000 times and the smallest SNP p value in a region at each permutation was chosen to generate the empirical distribution. The resulting permuted p value is equal to the proportion of the generated p values less than the observed one.

Examining type 1 error using a real dataset

The allelic association test was used to examine association at each SNP with simulated disease status in the real genotype dataset of 2,514 Chinese subjects. The pair-wise LD coefficients, r^2, were approximated by the square of Pearson correlation of genotypes coded as the number of minor alleles (0, 1 and 2). The M_e and type 1 error were assessed at five regions containing 100–300 SNPs in different chromosomes sampled randomly. SNPs with minor allele frequency less than 0.05, Hardy–Weinberg equilibrium p value less than 0.001, or genotype call rate less than 90% were excluded for this analysis. Type 1 error rates for these regions were obtained from the proportion of simulated phenotype datasets that resulted in significant p values (at FWER 0.05), from 50,000 simulated datasets.

Comparison of type 1 error using multivariate normal distribution (MVN)

On each chromosome, we randomly draw 500 regions with a random number of SNPs ranging from 2 to 100 in the same sample of 2,514 Chinese subjects mentioned above. At each region, M_e was estimated by five different methods and the corresponding p-value threshold, α', for individual SNPs to control the FWER (α) at 0.05, was calculated by Bonferroni correction method, $\alpha' = \alpha/M_e$. Given α', the FWER was calculated by the standard cumulative distribution function of MVN$(0, \sum)$:

$$1 - \int_{A}^{-A} f(x)\mathrm{d}u = 1 - \int_{A}^{-A} \frac{1}{(2\pi)^{M/2}|\sum|^{1/2}} \exp[-\frac{1}{2}(x-\mu)^T \times \sum^{-1}(x-\mu)]\mathrm{d}u,$$

where $A = \left[\Phi^{-1}(\alpha'/2), \ldots, \Phi^{-1}(\alpha'/2)\right]^T$ is a M dimension vector and $\sum$ is the genotypic Pearson correlation coefficient matrix of the M SNPs. We used the R package mvtnorm (http://cran.r-project.org/web/packages/mvtnorm/index.html) for the numerical integration of the MVN.

Estimating M_e and genome-wide significance thresholds in 13 genotyping arrays

It was noted that some SNPs in the genotyping arrays were not in the HapMap Project. For each array, a pair-wise r^2 was extracted into a subset from the HapMap LD dataset if both of its SNPs appeared on the genotyping array. The M_e and effective ratio were first estimated for SNPs in the subset. The total M_e of the genotyping array was then approximated by the number of SNPs on the array multiplied by the effective ratio. The p-value thresholds for genome-wide significant and highly significant association were equal to 0.05 and 0.001 divided by the total M_e of the genotyping array.

Results

Comparison of FWER in simulated data

The proposed method was compared to several existing methods as well as permutation testing (the gold standard) by simulation studies. Genotypes were simulated according to artificial LD patterns (Supplementary Table 1), and phenotypes were randomly assigned. As shown in Table 1, the use of the proposed M_e for Bonferroni correction produced FWER values that are generally close to the correct value of 0.05. As expected, standard Bonferroni correction for M SNPs is conservative. The correction based on Nyholt's M_e was liberal when there is only one LD block, but conservative in the multiple-LD-block scenarios. The Li and Ji (2005) method was liberal in all the simulated situations, while the Moskvina and Schmidt (2008) method was slightly conservative in the one-block scenario but became less conservative in the multiple-LD-block scenarios. Generally speaking, the type 1 error rates of Moskvina and Schmidt (2008), Galwey (2009), and the proposed method, along with those obtained via permutation, were comparable in the simulated dataset.

Comparison of FWER in real data

We further examined the family-wise type I error rates of the modified Bonferroni procedure by M_e in a real GWAS genotype dataset, where the phenotypes were re-assigned at random. The real GWAS data used were on a sample of 2,514 Chinese subjects typed by the Illumina Human610-Quad BeadChip. Five regions on different chromosomes were randomly chosen for an empirical validation.

Table 1 Empirical family-wise type 1 error rates (percentages) of alternative multiple testing corrections in simulated datasets

#SNP	Bonferroni for # SNP	Nyholt (2004)	Li and Ji (2005)	Moskvina and Schmidt (2008)	Galwey (2009)	Permutation	Proposed M_e
6	2.14	6.10	5.94	4.02	4.68	4.95	4.81
10	2.70	3.82	6.24	4.45	4.96	4.98	5.01
30	2.84	3.11	6.59	4.67	5.22	4.91	5.28
120	2.89	3.06	6.80	4.94	5.56	4.74	5.60

The nominal FWER is 0.05. We simulated 4 different LD patterns, in which a region may have 1 LD block (including 6 SNPs), 2 LD blocks (including 10 SNPs), 6 LD blocks (including 30 SNPs) and 24 LD blocks (including 120 SNPs), respectively, 40,000 replicates for each scenario. See Supplementary Table 1 for the LD patterns

Table 2 Family-wise error rates and effective number of independent tests in real genotype datasets

Chromosome	Position[a]	The observed		Nyholt (2004)		Li and Ji (2005)		Moskvina and Schmidt (2008)		Galwey (2009)		Proposed M_e	
		#SNP	Error	#SNP	Error	#SNP	Error	#SNP	Error	#SNP	Error	#SNP	Error
1	5733711 6877920	137	2.74%	128.8	2.97%	52.0	6.56%	63.1	5.48%	48.7	7.04%	68.5	5.16%
2	105304539 106766191	271	3.31%	264.7	3.40%	110.0	7.45%	147.1	5.81%	100.1	8.01%	159.3	5.43%
3	178265666 179728246	186	2.97%	180.8	3.06%	85.0	6.40%	99.3	5.54%	78.3	6.84%	106.7	5.11%
6	100078150 102098421	282	2.68%	271.7	2.78%	88.0	7.74%	126.2	5.66%	78.6	8.51%	137.3	5.21%
21	30821453 31663481	118	3.01%	113.2	3.14%	51.0	6.86%	62.9	5.73%	48.8	7.08%	68.1	5.26%

[a] The coordinates of NCBI Human Reference Genome Build 36.3 was used to denote the regions. The 5 regions were randomly selected. The Nominal FWER is 0.05; 50,000 simulated replicates were produced for each region

As shown in Table 2, the proposed measure of M_e led to FWERs much closer to the nominal $\alpha = 0.05$ for all regions in 50,000 simulated datasets. The simple Bonferroni correction for number of SNPs was conservative, as expected, as was the Bonferroni correction using Nyholt's M_e. The methods of Li and Ji (2005) and Galwey (2009) resulted in quite liberal FWERs. The FWERs based on Moskvina and Schmidt (2008) were only slightly more liberal than those based on our new method.

Comparison of FWER via MVN

For some common tests of association, the vector of test statistics for a single trait over multiple markers asymptotically follows a MVN, or can be transformed to follow a MVN (Lin 2005; Seaman and Muller-Myhsok 2005); the covariance matrix of this MVN can be approximated from the matrix of correlation coefficients between the markers (Moskvina and Schmidt 2008; Han et al. 2009; Seaman and Muller-Myhsok 2005; Conneely and Boehnke 2007). Given a fully characterized MVN, the FWER for any specified SNP-wise error rate can be calculated by multivariate integration. However, this is only feasible for a limited number of SNPs because of the computational burden in calculating the probabilities from a large-dimensional MVN. We randomly drew 500 genomic regions on each of the 22 autosomes and the X chromosome from the real GWAS dataset mentioned above. The number of markers within each region was random, ranging from 2 to 100. At each region, the five different methods were used to estimate M_e and to calculate the test-wise p-value threshold required to obtain a nominal FWER of 0.05. An estimate of the FWER corresponding to each test-wise p-value threshold is then obtained from the MVN. Figure 1 shows a Box plot of the MVN-derived FWER for the different methods over the 11,500 randomly selected regions. The proposed method of estimating M_e appears to give MVN-derived FWERs that agree most closely with the nominal level of 0.05, with least bias and small variance (Fig. 1). Consistent with the results obtained via simulation and permutation, the Bonferroni correction using Nyholt's M_e was generally conservative; the methods of Li and Ji (2005) and Galwey (2009) resulted in liberal FWERs, and all three have larger variance across genomic regions. The FWERs based on Moskvina and Schmidt (2008) were slightly more liberal but had comparable variance as the proposed method.

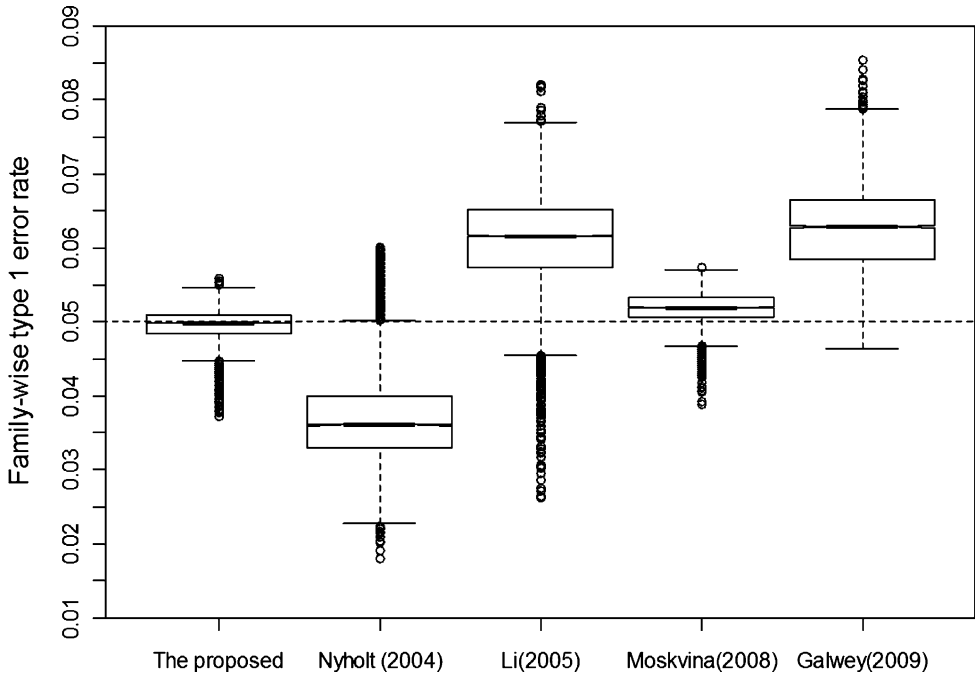

Fig. 1 Box plot of MVN-derived FWERs for five different methods. For each method, the nominal FWER was set to be 0.05. The *bottom* and *top* of each *box* mark the 25th and 75th percentile, respectively, and the *band* in the *box* denotes the 50th percentile. The *lines* above and below each *box* denote the upper and lower 1.5 interquartile range (IQR)

Estimating M_e and genome-wide significance thresholds in 13 genotyping arrays

Applying the proposed method, we systematically estimated M_e for 7 Illumina and 6 Affymetrix genotyping arrays, which have been widely used in GWAS in various populations. The r^2 values in the HapMap LD dataset (released on April 19, 2009) were used to calculate *p*-value correlation coefficients. Similar to the criteria proposed for genome-wide linkage studies (Lander and Kruglyak 1995), we calculated *p*-value thresholds for two genome-wide significance levels, significant association, and highly significant association, in which the FWER per scan are 0.05 and 0.001, respectively. Table 3 shows results based on HapMap CEU LD dataset. The thresholds for genome-wide significant association for all genotyping arrays (except for the Illumina HumanOmni2.5) range from 8.21×10^{-8} to 1.11×10^{-6}, which are all slightly less stringent than the widely-adopted one, 5.0×10^{-8}. An association scan based on the densest Affymetrix array requires a *p*-value threshold of 1.08×10^{-7} to declare a significant hit and the corresponding threshold for Illumina HumanHap 1 M is 8.21×10^{-8}. When combining all of the six Affymetrix arrays (1,011,854 unique SNPs in total), the *p*-value threshold for significant association is 1.04×10^{-7}. The Illumina HumanOmni2.5 seems to have an efficient design for effective SNPs. Its effective ratio is comparable with the Illumina HuamHap 1 M although it has doubled the SNP amount. When all of the seven Illumina arrays were combined, the *p*-value threshold turned out to be $\sim 3.5 \times 10^{-8}$. This amount of markers often happens in GWAS with genotype imputation, particularly in meta-analysis of GWAS. Consistent with observation in Barrett and Cardon (2006), Illumina arrays have larger effective ratio and require more stringent *p*-value thresholds to declare a significant finding than the Affymetrix arrays with similar number of SNPs. Results based on the HapMap CHB, JPT and YRI LD datasets are shown in Supplementary Table 2. Similarly, except for the Illumina HumanOmni2.5, the thresholds for genome-wide significant association using the other available genotyping arrays are all slightly less stringent than the widely adopted threshold, 5.0×10^{-8}.

Estimating M_e and significance thresholds in datasets of HapMap and 1000 Genomes Project

We then measured the M_e in datasets of HapMap and 1000 Genomes project. As shown in Table 4, although the number of unique SNPs in the HapMap LD dataset is over 2.5 million in the JPT, CHB and CEU panels, the M_e is less than 1 million and the ratio of M_e to the observed number (i.e., the effective ratio) is low, ranging from 0.26 to 0.30. The *p*-value thresholds for significant association are looser than 5.0×10^{-8}. The YRI panel has both the largest number of SNPs and effective ratio in the HapMap data, which makes the stringent *p*-value threshold 3.44×10^{-8}. Supplementary Table 3 shows the estimation results in the other 7 HapMap panels. There are only around 1.5 million SNPs in each panel and the effective ratios range from 0.41

Table 3 Estimated effective number of SNPs and *p*-value thresholds using the HapMap CEU sample

	Array Name	#SNP			Effective ratio	*p*-value thresholds*	
		In total	In HapMap	M_e		Significant association	Highly significant association
Illumina HumanHap	Omni2.5	2,450,000	969,415	544,311	0.561	3.63E−08	7.27E−10
	1 M	1,199,187	964,612	513,911	0.533	7.83E−08	1.57E−09
	650Y	660,557	609,860	393,752	0.646	1.17E−07	2.34E−09
	p610-Quad	598,821	561,716	374,316	0.666	1.25E−07	2.51E−09
	p550-Duo	561,122	540,047	370,501	0.686	1.30E−07	2.60E−09
	CNV370	353,188	338,660	258,305	0.763	1.86E−07	3.71E−09
	300-Duo	318,117	317,804	251,244	0.791	1.99E−07	3.98E−09
Affymetrix array	Array 6.0	934,968	783,702	388,751	0.496	1.08E−07	2.16E−09
	Array 5.0	443,816	384,423	211,592	0.550	2.05E−07	4.09E−09
	250 K Nsp	262,264	227,290	141,440	0.622	3.06E−07	6.13E−09
	250 K Sty	238,304	204,969	136,228	0.665	3.16E−07	6.31E−09
	50 K Hind 240	56,936	48,917	38,773	0.793	1.11E−06	2.22E−08
	50 K Xba 240	58,625	53,624	41,219	0.769	1.11E−06	2.22E−08
Merged illumina arrays		3,048,319	1,316,091	617,409	0.469	3.50E−08	6.99E−10
Merged affymetrix arrays		1,011,854	853,412	404,187	0.474	1.04E−07	2.09E−09

* *p*-value threshold is equal to the FWER/(Total number of SNPs × effective ratio)

Table 4 Estimated effective number of SNPs and genome-wide significance thresholds

		#SNP*	M_e	Effective ratio	*p*-value thresholds	
					Significant association	Highly significant association
HapMap	JPT panel	2,509,881	664,279.75	0.26	7.53E−08	1.51E−09
	CHB panel	2,554,939	693,418.45	0.27	7.21E−08	1.44E−09
	CEU panel	2,776,528	820,888.14	0.30	6.09E−08	1.22E−09
	YRI panel	3,114,362	1,452,799.72	0.47	3.44E−08	6.88E−10
1000 Genomes	ASN (Asian)	5,367,975	1,442,762.66	0.27	3.47E−08	6.93E−10
	EUR (European)	5,730,196	1,634,900.82	0.29	3.06E−08	6.12E−10
	AFR (African)	7,961,101	3,091,723.20	0.39	1.62E−08	3.23E−10

*In the 1000 Genomes dataset, 50.4, 51.9 and 53.2% SNPs with minor allele frequency below 0.05 were filtered out in the ASN, EUR and AFR panels, respectively

to 0.65. However, the *p*-value thresholds of significant association are still close to 5.0×10^{-8} in four panels (LWK, MKK, ASW and MEX). The CHD panel has the smallest number of SNPs and its *p*-value threshold for significant association is close to 10^{-7}.

The 1000 Genomes Project samples are divided into three panels according to their population ancestry. The common SNPs with minor allele frequency over 0.05 in the 1000 Genomes Project is over twice as large as the number of SNPs in the HapMap data. The effective ratios in the 1000 Genomes Project datasets of ASN and EUR panel are similar to that in the HapMap dataset of the corresponding populations although the amount of SNPs of the former is much more than that of the latter. The effective ratio in the 1000 Genomes Project AFR panel is smaller than that of HapMap YRI panel. The large M_e in the 1000 Genomes Project datasets entails stringent *p*-value thresholds below 5.0×10^{-8} for significant association. These *p*-value thresholds are useful reference for GWAS based on the genotype imputation using genotypes from HapMap and 1000 Genomes as reference sample.

We also estimated potential effective number of SNPs within known genes. Gene regions were defined according to the reference genome coordinates (GRCh37) of its transcripts with 2000 bp extension at both sides. The RefGene dataset was used in this analysis, including 37,322 transcripts of 22,610 genes. Table 5 lists the estimated effective number of SNPs and significance thresholds in the

Table 5 Estimated effective number of SNPs and significance thresholds in gene regions

	#SNP*	M_e	Effective ratio	*p*-value thresholds	
				Significant association	Highly significant association
ASN (Asian)	2,427,784	675845.93	0.28	7. 40E−8	1.48E−9
EUR (European)	2,591,410	765,693.14	0.30	6.53E−8	1.31E−9
AFR (African)	3,603,810	1,448,010.91	0.40	3.45E−8	6.91E−10

*51.0, 52.5 and 53.8% SNPs with minor allele frequency below 0.05 were excluded in the ASN, EUR and AFR panels, respectively

datasets of 1000 Genomes Project. The effective ratios in the gene regions are slightly higher than those in the whole genome. The *p*-value thresholds for SNPs in gene regions are roughly twice than that for the whole genome SNPs, back to a level close to 5×10^{-8}.

As expected from known LD patterns of populations worldwide (Frazer et al. 2007), the effective ratio in the Asia population is smaller than that in the European population and the African population has the largest effective ratio in both the HapMap and 1000 Genomes Project datasets. In principle, the effective ratio measures the average LD degree between SNPs in a marker set. A lower effective ratio is resulted from higher degree and/or longer-range of LD between markers. Given the same set of markers, a larger R_e may imply, on average, more miosis and recombination evens per genome happened in a population as a result of longer population history. Therefore, the largest effective ratio in the African population also indirectly supports the longest history of this population and is consistent with the 'Out of Africa' event hypothesis (Tishkoff et al. 1996; Reich et al. 2001). Correspondingly, the required *p*-value threshold for significant association in the African population is the more stringent those in the Asian and European population.

A software tool to estimate M_e and type I error

We have implemented the proposed measure of effective number of independent tests and the improved Bonferroni correction procedure in a software tool named genetic type 1 error calculator (GEC, http://statgenpro.psychiatry.hku.hk/gec/). Users can input actual genotype data [in either the conventional linkage pedigree format or PLINK binary format (Purcell et al. 2007)] or the HapMap LD data into GEC to quickly calculate M_e of the whole genome or at some specified genomic regions. Table 6 lists the running time GEC took to estimate M_e by the proposed method on 6 Illumina genotyping arrays using HapMap CEU LD data. If a set of SNP *p* values for genetic association tests is input, GEC will straightforwardly report the significant SNPs according to the improved Bonferroni correction procedure. GEC has both user-friendly command line and web-based graphic online interface.

Table 6 The running time GEC need to scan various genotyping arrays

Array name	#SNP	Running time (min)[a]
1 M-Duo	1,199,187	~7.8
650Y	660,557	~3.1
p610-Quad	598,821	~2.7
p550-Duo	561,122	~2.5
CNV370	353,188	~1
300-Duo	318,117	~0.9

The configuration of the computer doing the test is Intel(R) Xeon(R) X5670 @ 2.97 GHz, and Ubuntu 11.04 64bit; One GB maximal memory was set for GEC

[a] The time needed to read HapMap LD data was also included

Discussion

In the present study, we proposed a more robust measure of the effective number of independent tests, M_e, to control FWER for genetic association studies. Compared with previous methods (Gao et al. 2008; Li and Ji 2005; Moskvina and Schmidt 2008; Nyholt 2004; Galwey 2009), our measure is more robust to variable LD patterns in real datasets. Moreover, the new measure is additive across multiple distinct LD blocks. Capitalizing on this property, we developed a divide-and-conquer algorithm to handle large datasets, which can substantially relieve the computational burden when scanning millions of SNPs by avoiding calculating eigenvalues of the massive correlation matrix. We have demonstrated that this new method yields correct type I error rates and behaves similarly to the gold standard of permutation.

Pe'er et al. (2008) estimated the multiple testing burden in GWAS through simulation studies using data on the HapMap ENCODE regions to emulate an infinitely dense map, analogous to the Lander and Kruglyak approach for linkage analysis (Lander and Kruglyak 1995), and arrived at the commonly accepted genome-wide significance threshold of 5×10^{-8}. Similarly, by subsampling genotypes at increasing density and extrapolating to infinite density, Dudbridge and Gusnanto, 2008) estimated the genome-wide significance threshold to be about 7.2×10^{-8}. We noted that for 12 arrays widely used by previous GWAS, the recommended threshold for a

genome-wide error rate of 0.05, 5.0×10^{-8}, is conservative. For some studies, using arrays of $\sim$500,000 or 600,000 SNPs, a p-value threshold of $\sim 10^{-7}$ can be safely adopted without inflation of type I error. However, for GWAS using one of the latest Illumina arrays, HumanOmni2.5, a threshold as stringent as 5×10^{-8} or even slightly smaller is needed. This is also true for GWAS employing imputed common SNPs based on HapMap data. For GWAS with several million imputed SNPs from the 1000 Genomes Project dataset, a slightly more stringent p-value threshold (10^{-8}) is necessary. However, if one only examines the imputed SNPs within known genes, the threshold 5.0×10^{-8} can be used.

As previously noted (Pe'er et al. 2006; Hao et al. 2008), GWAS employing Affymetrix arrays allow use of a less stringent p-value threshold than those employing Illumina with similar amount of markers because Affymetrix randomly selected their SNPs while Illumina used a tagging approach in designing their arrays. Consistent with previous reports (Barrett and Cardon 2006; Pe'er et al. 2008), the multiple-testing burden for a sample from Japanese and Chinese populations is less heavy than that for a sample from Caucasian and African populations. Hence, the exact thresholds of individual GWAS slightly vary across genotyping platforms and sampling populations. We provided a user friendly tool, GEC, to quickly calculate exact genome-wide thresholds.

The effective ratio, R_e, we proposed can aid in marker selection for genetic association study design as well. Undoubtedly, a design with R_e close to 0 is not cost-efficient because this implies that most typed markers will be redundant and little independent information will be obtained. The larger the R_e is, the more independent genotype information the SNP marker set will have. Meanwhile, it should be noted that solely using R_e to evaluate a design may not be sufficient. It is possible that only one of two imperfectly correlated markers is in strong LD with an untyped disease susceptibility locus (DSL). Exclusion of one marker can definitely increase the R_e but can result in a loss of statistical power if the only marker in strong LD with a DSL is removed. Therefore, there is not a perfect relationship between R_e and statistical power.

In the present paper, we did not investigate the performance of the proposed method in an imputed GWAS dataset. The imputation quality, which is often related to imputation quality thresholds employed to clean the dataset, imputation algorithms and even matching degree between the study sample and reference panels, may affect the estimation of M_e in an imputed genotype dataset. If the quality of imputed genotypes are poor and the pair-wise LD between SNPs calculated by the imputed genotypes is largely different from that by actual genotypes, the estimation of M_e using the imputed dataset will be not reliable. On the other hand, if the imputation quality is good and the pair-wise LD between SNPs calculated according to the imputed genotypes is very similar to that by actual genotypes, the proposed method can be safely applied to estimate the M_e in the imputed GWAS datasets. In the present study, we estimated the M_e in the public datasets (including the HapMap Project panels and 1000 Genomes Project panels) which are widely used as reference panels for GWAS imputation. The M_e and p-value thresholds in these reference panels can be regarded as reference boundaries for the imputed GWAS datasets. In practice, one can employ our tool, GEC, to quickly estimate the M_e in an specific imputed GWAS datasets given the imputation quality is good.

Acknowledgments We thank HapMap Project for the LD data and 1000 Genomes Project for the genotype data used in this project. This work was funded HKU 7672/06 M, the Small Project Funding HKU 201007176166, the European Community's Seventh Framework Programme under grant agreement No. HEALTH-F2-2010-241909 and The University of Hong Kong Strategic Research Theme on Genomics.

References

Anderson CA, Pettersson FH, Barrett JC, Zhuang JJ, Ragoussis J, Cardon LR, Morris AP (2008) Evaluating the effects of imputation on the power, coverage, and cost efficiency of genome-wide SNP platforms. Am J Hum Genet 83(1):112–119. doi:10.1016/j.ajhg.2008.06.008

Barrett JC, Cardon LR (2006) Evaluating coverage of genome-wide association studies. Nat Genet 38(6):659–662. doi:10.1038/ng1801

Cheverud JM (2001) A simple correction for multiple comparisons in interval mapping genome scans. Heredity 87(Pt 1):52–58 (901[pii])

Conneely KN, Boehnke M (2007) So many correlated tests, so little time! Rapid adjustment of P values for multiple correlated tests. Am J Hum Genet 81 (6). doi:10.1086/522036

Dudbridge F, Gusnanto A (2008) Estimation of significance thresholds for genomewide association scans. Genet Epidemiol 32(3):227–234. doi:10.1002/gepi.20297

Duggal P, Gillanders EM, Holmes TN, Bailey-Wilson JE (2008) Establishing an adjusted p-value threshold to control the family-wide type 1 error in genome wide association studies. BMC Genomics 9:516. doi:10.1186/1471-2164-9-516

Frazer KA, Ballinger DG, Cox DR, Hinds DA et al (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449(7164):851–861. doi:10.1038/nature06258

Galwey NW (2009) A new measure of the effective number of tests, a practical tool for comparing families of non-independent significance tests. Genet Epidemiol 33(7):559–568. doi:10.1002/gepi.20408

Gao X, Starmer J, Martin ER (2008) A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. Genet Epidemiol 32(4):361–369. doi:10.1002/gepi.20310

Han B, Kang HM, Eskin E (2009) Rapid and accurate multiple testing correction and power estimation for millions of correlated markers. PLoS Genet 5(4):e1000456. doi:10.1371/journal.pgen.1000456

Hao K, Schadt EE, Storey JD (2008) Calibrating the performance of SNP arrays for whole-genome association studies. PLoS Genet 4(6):e1000109. doi:10.1371/journal.pgen.1000109

Hill WG, Robertson A (1968) Linkage disequilibrium in finite populations. Theor Appl Genet 38(4):226–231

Howie BN, Donnelly P, Marchini J (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet 5(6):e1000529. doi:10.1371/journal.pgen.1000529

Kung AW, Xiao SM, Cherny S, Li GH, Gao Y, Tso G, Lau KS, Luk KD, Liu JM, Cui B, Zhang MJ, Zhang ZL, He JW, Yue H, Xia WB, Luo LM, He SL, Kiel DP, Karasik D, Hsu YH, Cupples LA, Demissie S, Styrkarsdottir U, Halldorsson BV, Sigurdsson G, Thorsteinsdottir U, Stefansson K, Richards JB, Zhai G, Soranzo N, Valdes A, Spector TD, Sham PC (2010) Association of JAG1 with bone mineral density and osteoporotic fractures: a genome-wide association study and follow-up replication studies. Am J Hum Genet 86(2):229–239. doi:10.1016/j.ajhg.2009.12.014

Lander E, Kruglyak L (1995) Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. Nat Genet 11(3):241–247. doi:10.1038/ng1195-241

Li J, Ji L (2005) Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. Heredity 95(3):221–227. doi:10.1038/sj.hdy.6800717

Li MX, Gui HS, Kwan JS, Sham PC (2011) GATES: a rapid and powerful gene-based association test using extended simes procedure. Am J Hum Genet 88(3):283–293. doi:10.1016/j.ajhg.2011.01.019

Lin DY (2005) An efficient Monte Carlo approach to assessing statistical significance in genomic studies. Bioinformatics 21(6):781–787. doi:10.1093/bioinformatics/bti053

Metzker ML (2010) Sequencing technologies—the next generation. Nat Rev Genet 11(1):31–46. doi:10.1038/nrg2626

Montana G (2005) HapSim: a simulation tool for generating haplotype data with pre-specified allele frequencies and LD coefficients. Bioinformatics 21(23):4309–4311. doi:10.1093/bioinformatics/bti689

Moskvina V, Schmidt KM (2008) On multiple-testing correction in genome-wide association studies. Genet Epidemiol 32(6):567–573. doi:10.1002/gepi.20331

Nyholt DR (2004) A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. Am J Hum Genet 74(4):765–769. doi:10.1086/383251

Pahl R, Schafer H (2010) PERMORY: an LD-exploiting permutation test algorithm for powerful genome-wide association testing. Bioinformatics 26(17):2093–2100. doi:10.1093/bioinformatics/btq399

Pe'er I, de Bakker PI, Maller J, Yelensky R, Altshuler D, Daly MJ (2006) Evaluating and improving power in whole-genome association studies using fixed marker sets. Nat Genet 38(6):663–667. doi:10.1038/ng1816

Pe'er I, Yelensky R, Altshuler D, Daly MJ (2008) Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. Genet Epidemiol 32(4):381–385. doi:10.1002/gepi.20303

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81(3):559–575. doi:10.1086/519795

Ragoussis J (2009) Genotyping technologies for genetic research. Annu Rev Genomics Hum Genet 10:117–133. doi:10.1146/annurev-genom-082908-150116

Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, Richter DJ, Lavery T, Kouyoumjian R, Farhadian SF, Ward R, Lander ES (2001) Linkage disequilibrium in the human genome. Nature 411(6834):199–204. doi:10.1038/35075590

Salyakina D, Seaman SR, Browning BL, Dudbridge F, Muller-Myhsok B (2005) Evaluation of Nyholt's procedure for multiple testing correction. Hum Hered 60 (1):19–25; discussion 61–12. doi:10.1159/000087540

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 76(3):399–408. doi:10.1086/428140

Tishkoff SA, Dietzsch E, Speed W, Pakstis AJ, Kidd JR, Cheung K, Bonne-Tamir B, Santachiara-Benerecetti AS, Moral P, Krings M (1996) Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. Science 271(5254):1380–1387