European Journal of Human Genetics (2011) 19, 807–812

www.nature.com/ejhg

ARTICLE

Genomic inflation factors under polygenic inheritance

Jian Yang*[,1], Michael N Weedon[2], Shaun Purcell[3,4], Guillaume Lettre[5], Karol Estrada[6], Cristen J Willer[7], Albert V Smith[8], Erik Ingelsson[9], Jeffrey R O'Connell[10], Massimo Mangino[11], Reedik Mägi[12], Pamela A Madden[13], Andrew C Heath[13], Dale R Nyholt[1], Nicholas G Martin[1], Grant W Montgomery[1], Timothy M Frayling[2], Joel N Hirschhorn[3,14,15], Mark I McCarthy[12,16], Michael E Goddard[17], Peter M Visscher[1] and the GIANT Consortium

Population structure, including population stratification and cryptic relatedness, can cause spurious associations in genome-wide association studies (GWAS). Usually, the scaled median or mean test statistic for association calculated from multiple single-nucleotide-polymorphisms across the genome is used to assess such effects, and 'genomic control' can be applied subsequently to adjust test statistics at individual loci by a genomic inflation factor. Published GWAS have clearly shown that there are many loci underlying genetic variation for a wide range of complex diseases and traits, implying that a substantial proportion of the genome should show inflation of the test statistic. Here, we show by theory, simulation and analysis of data that in the absence of population structure and other technical artefacts, but in the presence of polygenic inheritance, substantial genomic inflation is expected. Its magnitude depends on sample size, heritability, linkage disequilibrium structure and the number of causal variants. Our predictions are consistent with empirical observations on height in independent samples of ~4000 and ~133 000 individuals.
European Journal of Human Genetics (2011) **19,** 807–812; doi:10.1038/ejhg.2011.39; published online 16 March 2011

Keywords: genome-wide association study; genomic inflation factor; polygenic inheritance

INTRODUCTION

Genome-wide association studies (GWAS) have led to the discovery of hundreds of genetic variants that are associated with complex diseases and traits.[1] In total, however, the identified variants explain only a fraction of total risk or phenotypic variance, resulting in the so-called 'missing heritability'.[2,3] One explanation is that most complex diseases and traits are caused by a large number of variants, the effects of which are too small to pass a stringent genome-wide significance level.[3] Therefore, large sample sizes are required and many collaborations have been established to achieve this, resulting in published meta-analyses for a range of diseases and traits.[4–8]

One standard quality-control measure for GWAS and meta-analysis is genomic control (GC).[9–11] The concept behind this method is that apart from a small number of SNPs that show a true association with the trait or disease, the test statistics for other SNPs should follow the distribution under the null hypothesis of no association between a SNP and the trait. However, artificial differences in allele frequencies due to population stratification, cryptic relatedness and genotyping errors will affect all SNPs and so the test statistics will be inflated across the whole genome.[12–14] For instance, the mean and median χ^2 value over all SNPs will be inflated by these artificial differences above their expectations under the null hypothesis of 1.0 and 0.455. This inflation can be detected and corrected for when testing for alleles that are associated with disease. The genomic control method was first proposed before GWAS, when it was hypothesised that the genetic architecture of complex traits was likely to consist of a small number of causal variants (in, eg, candidate genes) comprising a small proportion of the genome, and that a small number of non-associated null SNPs could be chosen to reflect most of the genome that was not associated with the trait. Before large-scale GWAS being conducted, this method was examined in the studies with hundreds of stratified individuals[13,15] and soon became a standard approach to quantify and adjust for population structure. In the first wave of GWAS, the genomic inflation factors observed in GWAS with thousands of individuals were usually <1.1, which were usually interpreted to be due to subtle population structure.[16] Much larger inflation factors have been observed in GWAS with large sample size especially when pooling a number of GWAS into a meta-analysis.[4,5] For example, the GIANT meta-analysis of height observed a genomic inflation factor of 1.42 even after GC-correction in each of the participating studies.[5]

The logic of GC relies on the fact that only a small fraction of the SNPs show a true association with the disease. However, published results from GWAS clearly indicate that there are many causal variants for a particular disease or trait. We therefore addressed the

[1]Queensland Institute of Medical Research, Brisbane, Queensland, Australia; [2]Genetics of Complex Traits, Peninsula College of Medicine and Dentistry, University of Exeter, Exeter, UK; [3]Broad Institute, Cambridge, MA, USA; [4]Center for Human Genetic Research, Massachusetts General Hospital, Boston, MA, USA; [5]Montreal Heart Institute (Research Center), Université de Montréal, Montréal, Québec, Canada; [6]Departments of Internal Medicine, Rotterdam, The Netherlands; [7]Department of Biostatistics, University of Michigan, Ann Arbor, MI, USA; [8]Icelandic Heart Association, Heart Preventive Clinic and Research Institute, Kopavogur, Iceland; [9]Department of Medical Epidemiology and Biostatistics, Karolinska Institute, Stockholm, Sweden; [10]Department of Medicine, University of Maryland School of Medicine, Baltimore, MD, USA; [11]Twin Research and Genetic Epidemiology Department, King's College London St Thomas' Hospital, London, UK; [12]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK; [13]Department of Psychiatry, Washington University, St Louis, MO, USA; [14]Program in Genomics and Divisions of Genetics and Endocrinology, Children's Hospital, Boston, MA, USA; [15]Department of Genetics, Harvard Medical School, Boston, MA, USA; [16]Oxford Centre for Diabetes, Endocrinology and Metabolism, Oxford, UK; [17]Department of Food and Agricultural Systems, University of Melbourne, Parkville, Victoria, Australia

*Correspondence: Dr J Yang, Queensland Statistical Genetics Laboratory, Queensland Institute of Medical Research, 300 Herston Road, Herston, Brisbane, Queensland 4006, Australia.
Tel: +61 73 845 3573; Fax: +61 73 362 0101; E-mail: jian.yang@qimr.edu.au
Received 28 July 2010; revised 24 January 2011; accepted 28 January 2011; published online 16 March 2011

question of what genomic inflation would be expected under polygenic inheritance. We used analytical derivations and simulation studies to quantify the expected mean (λ_{mean}) and median (λ_{median}) of a χ^2-statistic for association from a GWAS under polygenic inheritance, in the absence of population structure and genotyping errors. We show that the predicted genomic inflation factors are consistent with those observed in practice.

METHODS

Prediction of genomic inflation factors in quantitative trait and case–control association studies

In association analysis of a quantitative trait (QT), the non-centrality parameter (NCP) of χ^2-statistic for a causal variant is

$$\mathrm{NCP}_{\mathrm{C}}^{\mathrm{QT}}=\frac{Nq^2}{1-q^2} \tag{1}$$

where N is the sample size and q^2 is the proportion of phenotypic variance explained by a causal variant. Therefore, the NCP for a SNP that is in linkage disequilibrium (LD) with the causal variant is[17,18]

$$\mathrm{NCP}_{\mathrm{S}}^{\mathrm{QT}}=\frac{Nq^2r^2}{1-q^2r^2} \tag{2}$$

where r is the correlation coefficient between the SNP and the causal variant due to LD.

Under the assumption that the causal variants have not been genotyped on the current commercial SNP arrays, the mean of χ^2-statistics (λ_{mean}) in GWAS is

$$\lambda_{\mathrm{mean}}^{\mathrm{QT}} = 1+\frac{1}{n}\sum_{j}^{m}\sum_{k}^{s_j}\frac{Nq_j^2r_{jk}^2}{1-q_j^2r_{jk}^2} \tag{3}$$

where m is the number of causal variants, s_j is the number of SNPs in LD with the j-th causal variant, n is the total number of SNPs, q_j^2 is the variance explained by the j-th causal variant and r_{jk}^2 is the LD r^2 between the j-th causal variant and the k-th of the SNPs that are in LD with it.

Let s_0 be the number of SNPs that are completely in linkage equilibrium with the causal variants so that their test statistics are distributed as χ_1^2. In the absence of population structure or technical artefacts, the χ^2-statistics of all the SNPs will be a mixture of s_0 null SNPs and $n-s_0$ non-null SNPs (distributed as non-central χ^2) with a cumulative probability function of

$$Q(x) = \frac{s_0}{n}\Phi(x,1,0)+\frac{1}{n}\sum_{j}^{m}\sum_{k}^{s_j}\Phi\left(x,1,\ \frac{Nq^2r_{jk}^2}{1-q^2r_{jk}^2}\right) \tag{4}$$

where $\Phi(x, 1, \theta)$ is the cumulative probability of non-central χ^2-distribution with NCP of θ.

The median of χ^2-statistics (λ_{median}) is defined as $x=c$ so that $Q(c)=0.5$. The genomic inflation factor with respect to the median of χ^2-statistics is $\lambda_{\mathrm{median}}^{\mathrm{QT}}=c/\mathrm{median}(\chi_1^2)$

For a case–control (CC) association study, we assume an underlying threshold-liability model of disease and a multiplicative model of genotype relative risk (GRR). If GRR is small, the variance explained on an underlying liability scale for a genetic variant is[19]

$$q^2 \approx 2p(1-p)\,(GRR-1)^2/i^2 \tag{5}$$

where p is the allele frequency of the variant, and $i=z/K$ with K being the disease prevalence and z being the height of the standard normal curve at the truncating point pertaining to a probability of K.

Therefore, in a CC association study, the NCP for a causal variant is[20]

$$\mathrm{NCP}_{\mathrm{C}}^{\mathrm{CC}} \approx \frac{i^2v(1-v)Nq^2}{(1-K)^2} \tag{6}$$

where v is the proportion of cases in the sample. Therefore, the NCP of a SNP in LD with the causal variant in a case–control study is

$$\mathrm{NCP}_{\mathrm{S}}^{\mathrm{CC}} \approx \frac{i^2v(1-v)Nq^2r^2}{(1-K)^2} \tag{7}$$

The mean of χ^2-statistics from a genome-wide CC association study is

$$\lambda_{\mathrm{mean}}^{\mathrm{CC}} = 1+\frac{1}{n}\sum_{j}^{m}\sum_{k}^{s_j}\frac{i^2v(1-v)Nq_j^2r_{jk}^2}{(1-K)^2} \tag{8}$$

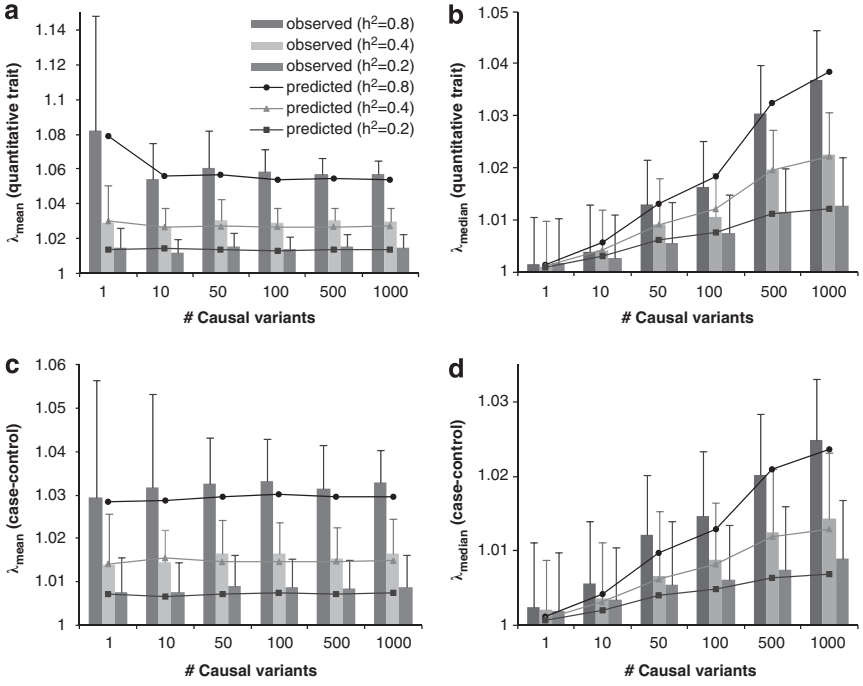

Figure 1 Genomic inflation factor observed in simulation *versus* that predicted by theory. Data are simulated based on real genotypes of 3925 individuals and 294 831 SNPs with different numbers of causal variants (m=1, 10, 50, 100, 500 and 1000) and heritabilities (h^2=0.2, 0.4 and 0.8). Each column represents the average of λ_{mean} (**a** and **c**) or λ_{median} (**b** and **d**) observed from 100 simulations. Error bars are SD. Each marked line represents the predicted λ_{mean} or λ_{median} averaged over 100 prediction replicates given m and h^2. For case–control studies (**c** and **d**), h^2 refers to heritability of liability on the underlying scale.

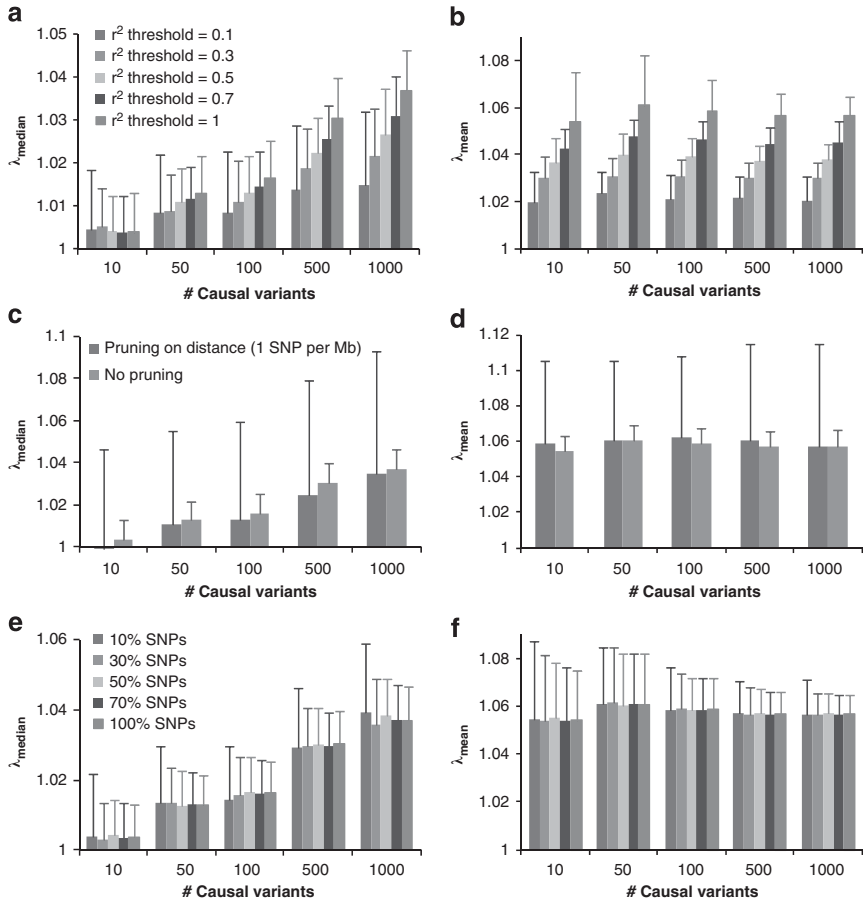

Figure 2 Genomic inflation factor for pruned (or selected) SNPs in simulation study. GWAS for quantitative trait is simulated based on real genotypes of 3925 individuals and 294 831 SNPs with heritability of 0.8 and with different numbers of causal variants (10, 50, 100, 500 and 1000). Each column represents an average of λ_{mean} (**b**, **d** and **f**) or λ_{median} (**a**, **c** and **e**) observed from 100 simulations. Error bars are SD. In (**a** and **b**), SNPs are pruned for LD using PLINK[22] with threshold r^2 value of 0.1, 0.3, 0.5 and 0.7. In (**c** and **d**), SNPs are pruned based on physical distance so that any pair of SNPs are at least 1 Mb away from each other. In (**e** and **f**), 10, 30, 50 and 70% SNPs are randomly sampled from all of the SNPs.

Analogous to that in the QT association study, the cumulative probability function of χ^2-statistics in a case–control study is

$$P(x) = \frac{s_0}{n}\Phi(x,1,0)+\frac{1}{n}\sum_{j}^{m}\sum_{k}^{s_j}\Phi\left(x,1,\frac{i^2v(1-v)Nq_j^2r_{jk}^2}{(1-K)^2}\right) \quad (9)$$

and $\lambda^{CC}_{median} = c/\text{median}(\chi^2_1)$ when $P(c)=0.5$.

Assume that the causal variants have a property that is similar to random SNPs in terms of allele frequency spectrum and LD structure. We randomly sampled m SNPs across the genome to mimic m causal variants. For each 'causal variant', we searched SNPs for LD within a d Mb region in either direction. Let y denote the genotype code for the causal variant and x for a SNP nearby. We tested for LD between the SNP and causal variant by simple regression, $y=b_0+b_1x+e$. We accepted a SNP in LD with the causal variant if the regression P-value <0.05. Obviously, there is a multiple-test problem, but it is unnecessary to correct for it because any SNP in significant LD with the causal variant will be inflated in single-SNP-based association tests.

For a QT, given the heritability (h^2) and sample size, we sampled q^2 for m causal variants from an exponential distribution with mean of m/h^2 and weighted each q^2 by $h^2/\Sigma q^2$ to constrain the sum of weighted q^2 to be h^2. Further, we predicted λ^{QT}_{mean} and λ^{QT}_{median} by equations (3) and (4). For a CC study, given disease prevalence, h^2 (heritability of liability on the underlying scale), sample size and number of cases, we predicted λ^{CC}_{mean} and λ^{CC}_{mean} by equations (8) and (9). When m becomes large, it is very likely that some SNPs will be in LD with multiple causal variants. In that case, we calculated s_0 as the number of SNPs that were not in LD with the causal variants rather than by using the equation $n-\sum_j^m s_j$ because otherwise we will underestimate s_0 and violate the definitions of equations (4) and (9), that is, $Q(x\to\infty)$ and $P(x\to\infty)$ would be >1. We then calculated the variance explained by a SNP in LD with s causal variants by $\left(\sum_j^s r_j\right)^2\sum_j^s q_j^2/s$ where we summed r rather than r^2 because the effects of two causal variants could be either in the same direction or opposite direction. We repeated the procedure 100 times and calculated the mean and SD of the predicted λ_{mean} and λ_{median}.

In the sections above, we showed how the genomic inflation factors can be predicted on the basis of the LD structure estimated from random SNPs and the heritability. When m is large (ie, q^2 is small), equations (3) and (8) are approximately equal to

$$\lambda^{QT}_{mean} \approx 1+\frac{Nh^2\overline{r^2}\bar{s}}{n} \quad (10)$$

$$\lambda^{CC}_{mean} \approx 1+\frac{Nh^2\overline{r^2}\bar{s}i^2v(1-v)}{n(1-K)^2} \quad (11)$$

where $\bar{s}$ is the average number of SNPs that are in LD with the causal variants (mimicked by a set of random SNPs) with average r^2 of $\overline{r^2}$. Since $\bar{s}$ and $\overline{r^2}$ are correlated, in practice, we use $\overline{r^2}\bar{s}$ instead of $\overline{r^2}\bar{s}$.

Samples and genotyping

We selected 3925 unrelated individuals (3248 adults and 677 16-year olds) from several GWAS conducted at the Queensland Institute of Medical Research

(QIMR).[21] All the samples had measured or self-reported height and were genotyped on the Illumina 370K or 610K SNP arrays (Illumina Inc., San Diego, CA, USA). All the samples were collected with informed consent and appropriate ethical approval. A total of 294 831 autosomal SNPs were retained for analysis after stringent quality control. Principal component analysis of ancestry showed that all of these 3925 individuals are of European descent (see ref. 21 for details of the data and quality control procedures). The phenotypes were corrected for age and sex, and standardised to *z*-scores in the adult and adolescent cohorts separately.

Simulation schemes

We performed simulation studies based on the observed genotype data of 3925 individuals and ~295K SNPs. We randomly sampled m SNPs as causal variants and generated the effect of each causal variant (b) from a standard normal distribution. We calculated the genetic value of each individual by $g=\sum_j^m x_j b_j$, where x is coded as 0, 1 or 2 for genotype qq, Qq or QQ (allele is arbitrarily called Q or q), respectively. We generated residual effects (e) from $N(0, \mathrm{var}(g)(1-h^2)/h^2)$ and calculated the simulated phenotype by $y=g+e$.

For CC studies, we generated the disease liability in the same way as above. We ranked the individuals by liability and assigned the top 1000 individuals as cases and the remaining individuals as controls.

We used different settings of heritability (h^2=0.2, 0.4 and 0.8) and number of causal variants (m=1, 10, 50, 100, 500 and 1000). For each setting, we repeated the simulation 100 times, randomising the positions of causal variants in each simulation replicate. We performed association analyses of the simulated data in PLINK[22] and calculated mean and median of χ^2-statistics with exclusion of the causal variants.

RESULTS

Under the assumption of polygenic inheritance of a quantitative trait and disease liability, we derived analytical equations to predict the genomic inflation factor in GWAS for QT and CC study. We show that in the absence of population structure, the genomic inflation factor, either λ_{mean} or $\lambda_{\mathrm{median}}$, is not expected to be unity, but is a function of sample size, LD structure, number of causal variants (m) and heritability (h^2) for both QT and CC association studies. For the CC study, it depends further on disease prevalence and the proportion of cases in the sample. When $m>10$, λ_{mean} is independent of the number of causal variants and depends only on the heritability, LD structure in the genome and the experimental sample size.

We demonstrate our method using a data set of 3925 unrelated individuals and 294 831 SNPs selected from several GWAS at the QIMR.[21] We validated the analytical equations by simulations based on the actual genotype data. Results show that both λ_{mean} and $\lambda_{\mathrm{median}}$ increase with h^2 (Figure 1), decrease when pruning SNPs for LD, but do not change when selecting SNPs at random or based on physical distance (Figure 2). Conditional on h^2, λ_{mean} is approximately constant, but $\lambda_{\mathrm{median}}$ increases with m, as predicted by theory. The reason is that when m increases, more SNPs (in LD with the causal variants) will depart from the null distribution (χ_1^2) so that the median of χ^2-statistics will deviate more from the expected median of (χ_1^2), whereas the effect of each locus decreases as constrained by the heritability, so that the mean test statistic remains the same. Given h^2 and m, we predicted λ_{mean} and $\lambda_{\mathrm{median}}$ by theory, but conditional on the observed LD structure. The LD structure is important because there are many SNPs in LD with each causal variant and so many SNPs have an inflated χ^2 and this increases the mean and to a less extent the median. We used the LD between SNPs as a proxy for the LD between SNPs and causal variants. In general, the predicted λ_{mean} and $\lambda_{\mathrm{median}}$ agree well with those observed from simulations (Figure 1). For a particular data set, when m is large (eg, $m>10$), λ_{mean} depends only on trait heritability.

We performed standard GWAS of height using the QIMR data set and observed λ_{mean}=1.035 and $\lambda_{\mathrm{median}}$=1.029 (Figure 3). We have

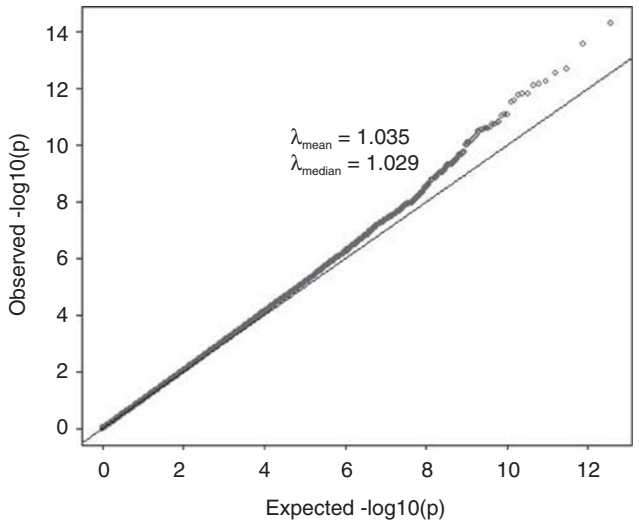

Figure 3 Quantile–quantile plot of height association result for QIMR data set (3925 unrelated individuals and 294 831 SNPs). All the SNPs passed stringent quality control and all the individuals are of European ancestry as verified by SNP data. The mean and median of χ^2-statistics are 1.035 and 1.029, respectively.

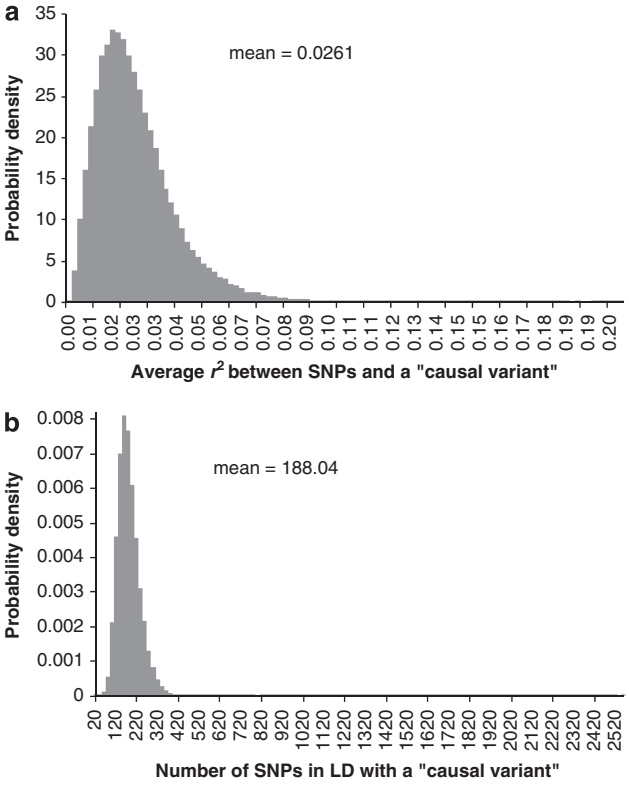

Figure 4 Histograms of (**a**) number of SNPs in significant LD with a 'causal variant' and (**b**) average r^2 between these SNPs and the 'causal variant'. The 'causal variants' are mimicked by randomly sampling (without replacement) 100 000 out of 294 831 SNPs across the genome. Simple regression is used to test for SNPs in LD with each 'causal variant' within 5-Mb distance in either direction.

shown previously that there is no evidence of population structure in this data set (Supplementary Figures 2 and 3 and Supplementary Tables 1 and 2 of ref. 21). We searched for SNPs in LD with 100 000 randomly selected loci and estimated an average of 188 SNPs that are in LD with each locus, with an average r^2 of 0.026 (Figure 4). We have previously estimated in this data that 45% of phenotypic variation for height can be explained by ~295K common SNPs.[21] Assuming that the unobserved causal variants are similar to random SNPs with respect to allele frequency and LD, we estimated h^2=0.54 (SE=0.1) after adjustment for imperfect LD between the causal variants and SNPs.[21] Given h^2=0.54, we predicted λ_{median} to be from 1.028 to 1.035 assuming that the number of causal variants for height ranges from 1000 to 4000 (Figure 5a), consistent with an observed λ_{median} of 1.029 and with height being highly polygenic.

We accessed the test statistics of the discovery set of GIANT meta-analysis (MA) of height with ~133 000 individuals and ~2.8-M genotyped and imputed SNPs.[5] We excluded ~636K SNPs with effective sample sizes <126 000 and extracted ~270K SNPs in common with the QIMR data set. We observed λ_{mean}=1.95 and λ_{median}=1.55. Assuming that the LD structure that underlies the GIANT MA results is similar to that in the QIMR data and h^2=0.54, we predicted λ_{median} for the GIANT MA to be from 1.32 to 1.59 with the assumption of 1000–4000 causal variants (Figure 5b).

DISCUSSION

We have shown by theory, simulation studies and analysis of multiple data sets that a significant inflation of test statistics is to be expected under polygenic inheritance even when there is no population structure. We have provided options in our software tool GCTA[23] to estimate LD structure and perform GWAS simulations, and provided an *R*-script to implement the theoretical predictions as described above

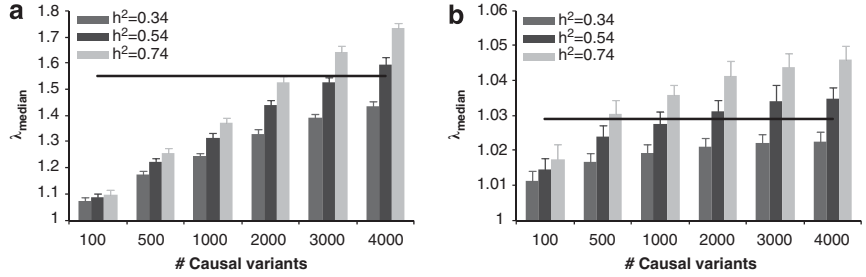

Figure 5 Predicted median of χ^2-statistics (λ_{median}) of height association study in (**a**) the QIMR data and (**b**) the GIANT meta-analysis. Each column is mean ± 2SD of 25 prediction replicates. The straight lines are the observed λ_{median} in real data analyses.

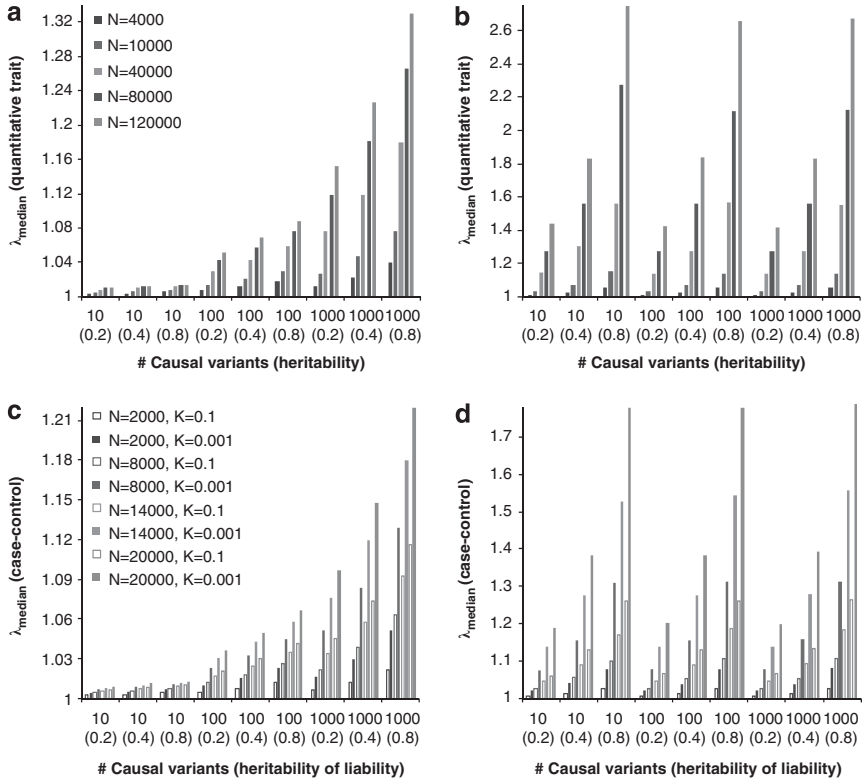

Figure 6 Predicted genomic inflation factor for quantitative trait (**a** and **b**) and case–control (**c** and **d**) association studies. Prediction is based on 294 831 SNPs with different numbers of causal variants and heritabilities (h^2), sample size (N) and disease prevalences (K, for case–control study). Each value is an average over 100 prediction replicates. For the case–control study, the number of cases and controls is equal.

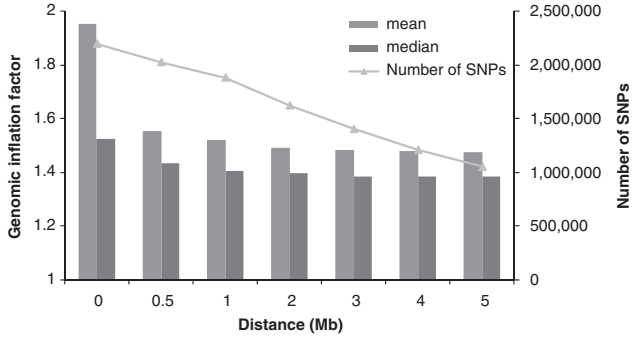

Figure 7 Genomic inflation factor for ~2.2-M SNPs (with exclusion of ~636K with effective sample sizes <126 000 from the total ~2.8 M SNPs) in GIANT meta-analysis for height with ~133 000 samples. A total of 318 top hits were identified by GIANT meta-analysis (genome-wide false discovery rate of 0.05).[5] Any SNP within *d* Mb distance (*d*=0.5, 1, …, or 5, *x*-axis) of the top hits is removed and genomic inflation factor is calculated using all of the remaining SNPs.

(http://gump.qimr.edu.au/gcta/gc). Of course, we are not denying that there may be spurious associations because of population structure for single SNPs,[14,16] but are questioning whether λ_{mean} or λ_{median} is an appropriate statistic to indict and adjust for population structure. In the absence of population structure, λ_{mean} reflects the trait heritability and λ_{median} further reflects the number of causal variants.

Standard GC theory predicts that the expected value of λ_{mean} and λ_{median} are the same,[9,10] because the distribution of the test statistic is a scaled (χ_1^2). Under polygenic inheritance, however, λ_{mean} and λ_{median} show explicitly different patterns with different sample size, heritability and disease prevalence (Figure 6). Results from the GIANT MA also show a much larger λ_{mean} than λ_{median}, as predicted from the polygenic model. When removing SNPs within *d* Mb (*d*=0.5, 1, …, or 5) of the 318 top hits (180 hits at genome-wide false-positive rate of 0.05 and additional 138 hits at genome-wide false discovery rate of 0.05) from ~2.2-M SNPs in the GIANT MA, λ_{mean} decreases from 1.95 to 1.48 and λ_{median} decreases from 1.53 to 1.39, but they do not converge, consistent with polygenic inheritance (Figure 7). Adjustment for GC in large meta-analyses may therefore be too conservative and reduce the power to detect significant SNP-trait associations.

In the presence of both population structure and polygenic inheritance (which may be regarded as a general case in practice), we cannot distinguish whether population structure or polygenic inheritance is the major cause of the genomic inflation unless we are able to estimate the proportion of phenotypic variance explained by all the SNPs and that attributed to population structure. It may be possible to discriminate polygenic inheritance from population structure by testing for associations between markers on different chromosomes. Population structure, including the presence of cryptic relationships among individuals in the sample, implies a correlation between alleles on different chromosomes. A genome-wide inflation of the test statistic with little or without such correlation is a strong support for polygenic variation.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

We thank all three reviewers for helpful comments. We acknowledge funding from the Australian National Health and Medical Research Council (NHMRC Grants 389891, 389892, 613672 and 613601), the Australian Research Council (ARC Grants DP0770096 and DP1093900) and the US National Institute of Health (NIH Grants AA13320, AA13321 and DA12854).

1 Hindorff LA, Sethupathy P, Junkins HA *et al*: Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci USA* 2009; **106**: 9362–9367.
2 Maher B: Personal genomes: the case of the missing heritability. *Nature* 2008; **456**: 18–21.
3 Manolio TA, Collins FS, Cox NJ *et al*: Finding the missing heritability of complex diseases. *Nature* 2009; **461**: 747–753.
4 Speliotes EK, Willer CJ, Berndt SI *et al*: Association analyses of 249 796 individuals reveal 18 new loci associated with body mass index. *Nat Genet* 2010; **42**: 937–948.
5 Lango Allen H, Estrada K, Lettre G *et al*: Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* 2010; **467**: 832–838.
6 Heid IM, Jackson AU, Randall JC *et al*: Meta-analysis identifies 13 new loci associated with waist-hip ratio and reveals sexual dimorphism in the genetic basis of fat distribution. *Nat Genet* 2010; **42**: 949–960.
7 Franke A, McGovern DPB, Barrett JC *et al*: Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat Genet* 2010; **42**: 1118–1125.
8 Teslovich TM, Musunuru K, Smith AV *et al*: Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* 2010; **466**: 707–713.
9 Devlin B, Roeder K: Genomic control for association studies. *Biometrics* 1999; **55**: 997–1004.
10 Reich DE, Goldstein DB: Detecting association in a case-control study while correcting for population stratification. *Genet Epidemiol* 2001; **20**: 4–16.
11 Zheng G, Freidlin B, Gastwirth JL: Robust genomic control for association studies. *Am J Hum Genet* 2006; **78**: 350–356.
12 Cardon LR, Palmer LJ: Population stratification and spurious allelic association. *The Lancet* 2003; **361**: 598–604.
13 Marchini J, Cardon LR, Phillips MS, Donnelly P: The effects of human population structure on large genetic association studies. *Nat Genet* 2004; **36**: 512–517.
14 Campbell CD, Ogburn EL, Lunetta KL *et al*: Demonstrating stratification in a European American population. *Nat Genet* 2005; **37**: 868–872.
15 Hao K, Li C, Rosenow C, Wong WH: Detect and adjust for population stratification in population-based association study using genomic control markers: an application of Affymetrix Genechip Human Mapping 10 K array. *Eur J Hum Genet* 2004; **12**: 1001–1006.
16 WTCCC: Genome-wide association study of 14 000 cases of seven common diseases and 3000 shared controls. *Nature* 2007; **447**: 661–678.
17 Chapman JM, Cooper JD, Todd JA, Clayton DG: Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. *Hum Hered* 2003; **56**: 18–31.
18 Spencer CC, Su Z, Donnelly P, Marchini J: Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. *PLoS Genet* 2009; **5**: e1000477.
19 Purcell SM, Wray NR, Stone JL *et al*: Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* 2009; **460**: 748–752.
20 Yang J, Wray NR, Visscher PM: Comparing apples and oranges: equating the power of case-control and quantitative trait association studies. *Genet Epidemiol* 2010; **34**: 254–257.
21 Yang J, Benyamin B, McEvoy BP *et al*: Common SNPs explain a large proportion of the heritability for human height. *Nat Genet* 2010; **42**: 565–569.
22 Purcell S, Neale B, Todd-Brown K *et al*: PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007; **81**: 559–575.
23 Yang J, Lee SH, Goddard ME, Visscher PM: GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* 2010; **88**: 76–82.