Published in final edited form as:
Nat Genet. ; 44(3): 243–246. doi:10.1038/ng.1074.

Differential confounding of rare and common variants in spatially structured populations

Iain Mathieson[1,*] and **Gil McVean**[1,2]
[1]Wellcome Trust Centre for Human Genetics, University of Oxford

[2]Department of Statistics, University of Oxford

Abstract

Well-powered genome-wide association studies, now possible through advances in technology and large-scale collaborative projects, promise to reveal the contribution of rare variants to complex traits and disease. However, while population structure is a known confounder of association studies, it is unknown whether methods developed to control stratification are equally effective for rare variants. Here we demonstrate that rare variants can show a systematically different and typically stronger stratification than common variants, and that this is not necessarily corrected by existing methods. We show that the same process leads to inflation for load-based tests and can obscure signals at truly associated variants. We show that populations can display spatial structure in rare variants even when F_{ST} is low, but that allele-frequency dependent metrics of allele sharing can reveal localized stratification. These results underscore the importance of collecting and integrating spatial information in the genetic analysis of complex traits.

Introduction

Quantifying the contribution to of rare variants to the heritability of different traits is an important and open question in complex trait genetics[1]. While there is no universally accepted definition of what constitutes a ‘rare’ variant, a minor allele frequency (MAF) of 1% is the conventional definition of polymorphism[2]. At this frequency, the power of the current generation of genome wide association studies (GWAS) is negligible for modest effect sizes[3]. Therefore, although a small number of associations with rare variants have been reported, for example with type 1 diabetes[4] and cholesterol levels[5,6], it has not been possible to test the hypothesis that rare variants account for a significant proportion of the missing heritability for most complex traits. Recently, however, four factors have combined to make the direct investigation of rare variants possible. First, the increasing size of GWAS samples and meta-samples, now approaching cohort sizes of 100,000 through large-scale international collaborations, boosts power. Second, the ascertainment of many rare variants through the 1000 Genomes project[7], has enabled imputation of millions of rare and low frequency variants and led to the development of a new generation of low-cost genotyping platforms that interrogate rare variants directly. Third, the decline in the cost of sequencing technology has enabled large scale sequencing studies to be performed which in principle allow the detection of all variants in a sample. Finally, the recent development of new statistical tests for association aimed at rare variants[8-13] (reviewed in ref. 14) potentially

[*]Corresponding author: mathii@well.ox.ac.uk.

Author Contributions

GM conceived and designed the study. IM ran simulations and collected results. GM and IM jointly wrote the simulation code and manuscript.

provides power to detect genes or pathways harbouring multiple rare variants for which there would be individually low power to detect association.

The large sample sizes required for such studies typically require combining information across multiple geographic locations, within and across countries. Population structure, which can lead to spurious correlations between allele frequencies and non-genetic risk factors, has long been known to be a major potential confounding factor for association studies[15-17]. The effects of stratification have been studied extensively[18-20] and testing and correcting for structure is now standard practice in GWAS through methods such as genomic control (GC)[21,22], principal component analysis (PCA)[23] and mixed models[24]. However, analyses of these methods have typically concentrated on common variants and there has been little investigation of the effect that structure might have specifically on rare variants. Informally, rare variants, through being typically recent, may tend to have different geographic distributions than more common and typically older variants.

We therefore set out to investigate (a) under what conditions population structure will lead to differential test-statistic inflation for variants of different frequencies, (b) whether methods effective for controlling stratification of common variants are also appropriate for rare variants, (c) whether different ways of analyzing rare variants (single-marker versus aggregating) are equally affected by structure and (d) how best to measure population structure in empirical data in a manner that is informative about differential stratification. We used a simple lattice model to approximate population structure across a geographical region and investigated the interaction between the spatial distribution of non-genetic risk and inflation of standard association tests under the null model of no genetic risk (Online methods). We contrasted the situation where non-genetic risk is smoothly distributed (for example, a latitudinal effect) with the situation where the same overall risk is concentrated into one or more small, sharply defined regions (for example, localized environmental pollution).

Results

As is well documented, population structure leads to inflation of association test statistics under the null and hence systematic underestimation of P-values. When the risk has a wide and smooth distribution, rare variants show less inflation than common variants (Fig. 1a, c). In contrast, when the risk has a small, sharp spatial distribution, rare variants show more inflation than common variants, particularly for small P-values (Fig. 1b, d). The magnitude of inflation increases as the P-value decreases in both scenarios and the greatest inflation is for variants with frequency approximately equal to the fraction of the area with elevated risk (Fig. 1c, d). As the size or smoothness of the area of risk increases, the inflation is spread over a wider range of P-values (Supplementary Fig. 1 and 2).

Such differential behaviour can be understood as a result of the interaction between the spatial distribution of risk and the spatial distribution of variants. Small P-values occur when a variant shows strong correlation with the non-genetic risk. Rare variants, through being recent, tend to show greater geographic clustering than common ones (Fig. 2a-c). When non-genetic risk varies on a large scale, rare variants cannot be highly correlated with it (Fig. 2d). In contrast, when non-genetic risk varies on a small scale, although most variants are uncorrelated with the risk, rare variants have a tail of highly correlated variants (Fig. 2e), which drive the inflation

Several methods for correcting for population stratification in GWAS have been developed. The most popular are genomic control (GC)[21,22], principal component analysis (PCA)[23] and linear mixed models[24]. These corrections are known to be effective in the standard GWAS setting and we find they are all effective when non-genetic risk has a large and smooth distribution (Fig. 3a). However, none of them are effective for the small, sharp distribution of risk (Fig. 3b). GC fails in this case because most variants, even rare ones, have correlation with the non-genetic risk of close to 0 (Fig. 2e). PCA and mixed models fail because they effectively try to correct based on linear functions of relatedness. In the simulations, the first few principal components always include the axes of the grid, so can correct for any non-genetic risk which can be expressed as a linear function of these axes. However, the small, sharp region of risk would require a highly non-linear function to be expressed in these terms, which cannot be achieved simply by including the top components. Ultimately, including a large enough number of principal components will remove virtually all stratification (here, between 20 and 100 PC's is sufficient; Supplementary Fig. 3), but it is not possible to know how many are required and inclusion of many components will lead to substantial reduction in power to detect true associations.

Where variants are sufficiently rare that they are unlikely to be observed in more than a few samples, adequate power to detect true association can only be obtained by combining information across multiple variants within a gene, though this can be approached in many ways[8-14]. To assess the effects of stratification of such aggregating tests, we considered one of the simplest 'load-based' tests[11], which tests association with the number of rare variants carried in a region, typically a gene. For smoothly-varying Gaussian risk, test-statistic inflation is largely independent of the number of variants considered (Fig. 3c). For sharply-defined risk, test-statistic inflation is reduced as more variants are considered, but still increases sharply for low P-values (Fig. 3d). Given that some versions of these tests cannot easily accommodate relatedness information and that the problems of spatial structure will increase as allele frequency decreases, these results suggest that similar care need be taken when interpreting enrichment of either single or multiple variants within cases or controls.

The results discussed so far relate to inflation under the null. However, another implication of differential structure is that causal rare variants may be geographically localized. Thus even when there is no spatial structure to non-genetic risk, test-statistic inflation will be observed. When there are many loci with rare variants contributing to the background genetic effect, inflation is typically stronger for common variants and will be corrected for by standard approaches. However, when there are only a few loci driving risk, inflation is greater for rare variants (Supplementary Fig. 4). Consequently, if genetic risk is driven by small numbers of rare variants, then true signals are more likely to be obscured by rare variants that show association even though they are not physically linked to the causal variants.

Discussion

We have demonstrated that under certain conditions rare and common variants exhibit differential patterns of stratification. However these results are qualitative and we must also ask whether these conditions are likely to be met in practice. While the data that would be required to investigate this effect directly are not yet available, we can nonetheless consider metrics that could be used to relate our simulations to real populations. Historically, approaches to summarizing population structure in genetic data have focused on simple statistics, such as Wright's fixation index F_{ST}, which measures the proportion of overall genetic variation that results from between-population variation. Among human populations, F_{ST} is typically estimated to be <0.1 (for example, 0.071 between the 1000 Genomes CEU and YRI populations[7] and typically <0.02 within Europe[25]). Dividing the simulated grid into two equal sub-populations, for the migration parameter used for Figure 1 (M=0.01) F_{ST} is approximately 0.1, which is comparable to a worldwide sample. Within a European sample, a more appropriate migration parameter might be M=10, which gives F_{ST}<0.01, a value which would be considered negligible. However, F_{ST} estimates are driven by common variants, and also depend on the relative sizes and number of the sub-populations (Supplementary Fig. 5). Analysis of allele sharing by distance as a function of distance shows that while common variants show effectively no excess allele sharing at short ranges, even with M=10 rare variants still show excess clustering (Fig. 4) and although stratification is much reduced compared to a low migration rate, it is still greatest for rare variants (Supplementary Fig. 6). These results are consistent with empirical observations that show very low rare-allele sharing even between very closely related human populations[26]. The fact that excess allele sharing increases as frequency decreases implies that even for relatively unstructured populations, this effect will be observed below some, sufficiently low, variant frequency. These results highlight the need for methods for explicitly showing spatial structure, such as the allele-sharing plot (Fig. 4) or other spatial correlation measures such as Moran's I statistic[27], which provides a much richer and more informative representation than any single statistic.

There are three ways in which non-genetic risk might show sharply-defined boundaries of the type for which we have shown differential inflation. First, localized environmental exposure may be highly patchy, for example associated with urban areas. Second, there may be systematic measurement bias at a single recruitment centre. Third, and more subtly, there may be local variation in recruitment policy or rates of misclassification (the effect of which can be thought of as changing the background disease risk). Although we have simulated quantitative trait data, case-control studies are subject to the same issues of population structure and a case-control study that randomly misclassifies cases and controls will bias effect size estimates[28]. When this misclassification is restricted to a particular spatial area, for example a single recruitment center in a large study, it will produce the effects described here. In fact, if we add additional disconnected small areas of risk of the same size as the first, the inflation in P-value has the same distribution with respect to frequency (Supplementary Fig. 7) so this observation would extend to the case where multiple collection centers were making biased measurements or random misclassifications. Because the extent and clustering of non-genetic risk will differ between phenotypes and study designs, it is not possible to predict any general influence of differential stratification. The principal problem with trying to account for known non-genetic risk (i.e. to include these as covariates within the analysis) is that while information about broad-scale risk factors may be available, typically, the more localized a risk factor is, the less we are likely to know about it and the greater effect this will have on rare variants.

Given that existing methods can fail to correct for rare variant stratification, what approaches can be taken to guard against its effects? One approach is to use methods that are robust to stratification (though at a cost to power and ease of experimental design), such as family-based association, perhaps only for replication. Another is to adapt existing methods to work better with rare variants. For example, although PCA with rare variants does not effectively control inflation if we linearly correct using the top components (Fig. 3b), in principal, more sophisticated methods for selecting non-linear functions of components could correct appropriately.

Alternatively, we might look to the development of new measures of relatedness more sensitive to recent ancestry and fine-scale structure. Whichever approach is taken, it is likely to require fine-grained information about the geographic origins and recruitment path of each sample. The collection of such information must be an important consideration in the design of future studies.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by the Wellcome Trust through grants 089250/Z/09/Z to IM, 086084/Z/08/Z to GM and 090532/Z/09/Z to the Wellcome Trust Centre for Human Genetics. The authors thank Matti Pirinen, Chris Spencer, Zam Iqbal and Cecilia Lindgren for discussion and Matti Pirinen for providing software to fit the linear mixed models used in this analysis.

Appendix

Online Methods

Simulations of association studies

We simulate genotypes and quantitative traits by starting with a number of individuals and their locations on the grid and work backwards in time to generate random genealogical events. Each event is either a coalescence of two lineages or a migration of a single lineage from one square to another. The relative rates of coalescence and migration depend on the population-scaled migration rate M and the number and distribution of lineages on the grid.

More precisely, suppose we have a $K \times K$ grid and we wish to simulate a sample of $C = cK^2$ individuals where c is the number of individuals in each grid square. We simulate L loci on each of G genealogies for a total of LG loci. Each genealogy represents an independent genomic region, with no recombination inside each region. Index the grid squares by i, j and denote the number of lineages in grid square i, j at time t by $n^t_{i,j}$. Let $s_{i,j}$ represent the number of grid squares adjacent to i, j in a Manhattan sense, so $s_{i,j} \in \{2, 3, 4\}$.

Now we start at $t = 0$ and repeat the following steps until only one lineage remains:

1. At time t, the rate of coalescence within grid square i, j is $\lambda^t_{i,j} = \frac{n^t_{i,j}\left(n^t_{i,j} - 1\right)}{2}$ and the total rate of coalescence is $\lambda^t_{\bullet,\bullet} = \sum_{i,j} \frac{n^t_{i,j}\left(n^t_{i,j} - 1\right)}{2}$ where we use dots to represent summation over indices. The rate of migration for each grid square is $\mu^t_{i,j} = \frac{M n^t_{i,j} s_{i,j}}{2}$ and the total rate of migration is $\mu^t_{\bullet,\bullet}$. The next event occurs at time $t + T$ where $T \sim Exp\left(\lambda^t_{\bullet,\bullet} + \mu^t_{\bullet,\bullet}\right)$. It is coalescence with probability $\frac{\lambda^t_{\bullet,\bullet}}{\lambda^t_{\bullet,\bullet} + \mu^t_{\bullet,\bullet}}$ and migration with probability $\frac{\mu^t_{\bullet,\bullet}}{\lambda^t_{\bullet,\bullet} + \mu^t_{\bullet,\bullet}}$.

2. If the next event is coalescence, it occurs in grid square i, j with probability $\frac{\mu^t_{i,j}}{\mu^t_{\bullet,\bullet}}$. In this grid square, choose two lineages uniformly, and join them together. Return to 1) with t replaced by $t + T$.

3. If the next event is migration, it occurs in grid square i, j with probability $\frac{\mu_{i,j}^{t}}{\mu_{\bullet,\bullet}^{t}}$. In this grid square, choose one lineage uniformly, and move it uniformly to one of the adjacent grid squares. Return to 1) with t replaced by $t + T$.

Once we have simulated a single instance of the genealogy, we generate genotypes at L random loci by sampling L nodes from the genealogy with replacement, selecting each node with probability proportional to the length of the branch above that node, and setting each individual's genotype to 0 or 1 at each locus according to whether they are descended from that node or not so that a genotype of 0 represents an ancestral allele and a genotype of 1 represents a derived allele.

We generate quantitative traits for each locus, for each individual by sampling from a standard normal distribution. We shift the mean of the distribution for each individual according to the non-genetic risk in the square that that individual came from. Let $\phi: [1, C] \rightarrow [1, K] \times [1, K]$ be a function which maps each individual to the grid square which they originated in. Then, for individual k, the trait value $Y_k \sim N\left(R_{\phi(k)}, 1\right)$ where $R_{i,j}$ is the non-genetic risk in grid square i, j. In a real experiment, each individual would have one value of Y_k which would be tested against every locus, but to reduce the uncertainty due to sampling error, in our simulations, we resample the trait independently for each locus, except where that would be inappropriate, for example when testing corrections, in which case we average the results over many experiments instead.

We perform association tests for this locus by fitting a simple linear model $Y_k^{l,g} = \mu^{l,g} + \beta^{l,g} X_k^{l,g} + \varepsilon_k^{l,g}$ where $\varepsilon_k^{l,g} \sim N\left(0, \sigma_{l,g}^2\right)$ IID for some $\sigma_{l,g}$ and computing the P-values of the beta estimates. We then repeat this for $l = 1\ldots L$ and $g = 1\ldots G$.

The results in Figure 1 use the following parameters; $K = 20$, $c = 2$ so $C = 800$, $M = 0.01$, $G = 100{,}000$ and $L = 1000$. This gives us 10^8 points in each QQ plot. The maximum non-genetic risk for the Gaussian risk was 0.4 standard deviations and 1 standard deviation for the small, sharp risk. We computed the statistic $Q = \frac{\mathrm{var}\left(R_{\phi(k)}\right)}{\mathrm{var}\left(Y_k\right)}$, which is the proportion of the phenotypic variance explained by the non-genetic risk, and was equal to 1.4% for the Gaussian risk and 2.2% for the small, sharp risk.

Correcting for stratification

In order to investigate the effect of corrections for population structure, we sample genotypes for multiple loci and genealogies as described above. However now we sample only one realization of the quantitative trait $Y_k \sim N\left(R_{\phi(k)}, 1\right)$ to use for every l, g. We compute single marker test statistics as described above, and then perform the following corrections

1. Genomic control. We take the P-values for each locus $p^{l,g}$ and compute chi-squared statistics $\mathrm{X}^{l,g} = F_{\chi^2}^{-1}\left(1 - p^{l,g}\right)$ where F_{χ^2} is the cumulative distribution function of the chi-squared distribution with one degree of freedom. We then compute adjusted test statistics $\tilde{X}^{l,g} = \frac{X^{l,g}}{\lambda}$ where the genomic inflation constant $\lambda = \frac{\left\langle X^{l,g} \right\rangle}{F_{\chi^2}^{-1}(0.5)}$ and $\langle \bullet \rangle$ represents the median and compute adjusted P-values $\tilde{p}^{l,g} = 1 - F_{\chi^2}\left(\tilde{X}^{l,g}\right)$.

2. Principal component analysis. We compute the principal components of the $LG \times C$ genotype matrix $X=\{X_k^{l,g}\}$, say $P^1,\ldots,P^N$ and then fit the linear model $Y_k=\mu^{l,g}+\beta^{l,g}X_k^{l,g}+\sum_{i=1}^{10}\gamma_i^{l,g}P_i+\varepsilon_k^{l,g}$ where $\varepsilon_k^{l,g}\sim N\left(0,\sigma_{l,g}^2\right)$ for some $\sigma_{l,g}^2$. We test the significance of the beta estimates as before. We also tried PCA, but calculating the principal components only from rare markers with MAF<4%.
3. Mixed model analysis. The linear mixed model has the form $Y=\mu+\beta X+\varepsilon_G+\varepsilon_R$ where $\varepsilon_G \sim MVN(0,\sigma_G^2 R)$ and $\varepsilon_R \sim MVN(0,\sigma_R^2 I)$ for some σ_G^2 and σ_R^2 and where R is the fixed kinship matrix. Here we used the correlation matrix of the genotype vectors as an approximation to the kinship matrix.

We fitted the model using software kindly provided by Matti Pirinen. This fits the same model as the EMMA package[24], but with a more efficient numerical algorithm (Matti Pirinen, Peter Donnelly and Chris Spencer; Efficient Computation with a Linear Mixed Model on Large-scale Data Sets with Applications to Genetic Studies; Manuscript in revision).

Because we are sampling a single trait for all loci, there is some variance in the amount of inflation and the effectiveness of the corrections, mainly because of the sampling variance of the phenotype. So that we give an accurate idea of the overall effect, we perform these simulations 100 times and show the pointwise average of the QQ plots.

Load based test

We implement a test for association of a quantitative trait with rare variant load (sometimes described as a "collapsing" or "burden" test) as described in reference [11].

We simulate variants $X=\{X_k^{l,g}\}$ as described above, but include only rare variants (MAF<4%). We imagine that each genealogy represents an independent genomic region (say, a gene) and that each of the L variants on that genealogy represents a rare variant that segregates in the population. We then compute the rare variant load (or "burden") B_k^g for each individual for each region by counting the number of derived alleles at each locus so that $B_k^g=X_k^{\bullet,g}$. We simulate traits as described above and test association with the rare variant load by fitting the model $B_k^g=\mu^g+\beta^g B_k^g+\varepsilon_k^g$ where $\varepsilon_k^g\sim N\left(0,\sigma_g^2\right)$.

Excess allele sharing by distance

We calculate the probability that two individuals at a given distance share an allele, compared to what would be expected from a randomly mating population.

Specifically, given the $LG \times C$ genotype matrix $X=\{X_k^{l,g}\}$, we first divide the variants into rare, low frequency and common variants based on their allele frequencies. Suppose there are R rare variants in an $R \times C$ genotype matrix $\tilde{X}=\{\tilde{X}_k^r\}$ with allele frequencies $f_r=\frac{\tilde{X}_\bullet^r}{C}$ and the spatial distance between individuals i and j is given by $D_{i,j}$. Then we compute the excess allele sharing at distance d, Q_d as:

$$Q_d=\frac{1}{R}\sum_{r=1}^{R}\frac{\sum_{i=1}^{c}\left[\chi\left\{\tilde{X}_r^i=1\right\}\sum_{j>1}\chi\left\{\left(\tilde{X}_r^j=1\right)\cap\left(D_{i,j}=d\right)\right\}\right]}{f_r\sum_{i=1}^{C}\left[\chi\left\{\tilde{X}_r^i=1\right\}\sum_{j>i}\chi\left\{\left(D_{i,j}=d\right)\right\}\right]}$$

Where $\chi\{A\}$ is the indicator function of the event A and recalling that $\tilde{X}_k^r\in\{0,1\}$ with 0 and 1 representing the ancestral and derived alleles respectively. So for a given distance, for each derived allele, we count the number that are shared at a given distance, and divide by the total number of individuals at that distance to get the allele sharing probability. We then divide by the allele frequency, which is the allele sharing probability in an infinite homogenous population, to get the excess allele sharing probability. Figure 3c-d shows $\log_{10} Q_d$ for rare, low frequency and common variants, with simulation parameters described in the figure legend.

References

1. Manolio TA, et al. Finding the missing heritability of complex diseases. Nature. 2009; 461:747–753. [PubMed: 19812666]
2. Bodmer W, Bonilla C. Common and rare variants in multifactorial susceptibility to common diseases. Nat. Genet. 2008; 40:695–701. [PubMed: 18509313]
3. 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. [PubMed: 19492015]
4. Nejentsev S, Walker N, Riches D, Egholm M, Todd JA. Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. Science. 2009; 324:387–389. [PubMed: 19264985]
5. Cohen JC, et al. Multiple rare alleles contribute to low plasma levels of HDL cholesterol. Science. 2004; 305:869–872. [PubMed: 15297675]
6. Wang J, et al. Common and rare ABCA1 variants affecting plasma HDL cholesterol. Arterioscler. Thromb. Vasc. Biol. 2000; 20:1983–1989. [PubMed: 10938021]
7. Durbin RM, et al. A map of human genome variation from population-scale sequencing. Nature. 2010; 467:1061–1073. [PubMed: 20981092]
8. Ionita-Laza I, Buxbaum JD, Laird NM, Lange C. A new testing strategy to identify rare variants with either risk or protective effect on disease. PLoS Genet. 2011; 7:e1001289. [PubMed: 21304886]
9. Li B, Leal SM. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am. J. Hum. Genet. 2008; 83:311–321. [PubMed: 18691683]
10. Madsen BE, Browning SR. A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet. 2009; 5:e1000384. [PubMed: 19214210]
11. Morris AP, Zeggini E. An evaluation of statistical approaches to rare variant analysis in genetic association studies. Genet. Epidemiol. 2010; 34:188–193. [PubMed: 19810025]
12. Mukhopadhyay I, Feingold E, Weeks DE, Thalamuthu A. Association tests using kernel-based measures of multi-locus genotype similarity between individuals. Genet. Epidemiol. 2010; 34:213–221. [PubMed: 19697357]
13. Neale BM, et al. Testing for an unusual distribution of rare variants. PLoS Genet. 2011; 7:e1001322. [PubMed: 21408211]
14. Bansal V, Libiger O, Torkamani A, Schork NJ. Statistical analysis strategies for association studies involving rare variants. Nat. Rev. Genet. 2010; 11:773–785. [PubMed: 20940738]

15. Knowler WC, Williams RC, Pettitt DJ, Steinberg AG. Gm3;5,13,14 and type 2 diabetes mellitus: an association in American Indians with genetic admixture. Am. J. Hum. Genet. 1988; 43:520–526. [PubMed: 3177389]
16. Lander ES, Schork NJ. Genetic dissection of complex traits. Science. 1994; 265:2037–2048. [PubMed: 8091226]
17. Pritchard JK, Donnelly P. Case-control studies of association in structured or admixed populations. Theor. Popul. Biol. 2001; 60:227–237. [PubMed: 11855957]
18. Cardon L, Palmer L. Population stratification and spurious allelic association. Lancet. 2003; 361:598–604. [PubMed: 12598158]
19. Clayton DG, et al. Population structure, differential bias and genomic control in a large-scale, case-control association study. Nat. Genet. 2005; 37:1243–1246. [PubMed: 16228001]
20. 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. [PubMed: 15052271]
21. Bacanu SA, Devlin B, Roeder K. The power of genomic control. Am. J. Hum. Genet. 2000; 66:1933–1944. [PubMed: 10801388]
22. Devlin B, Roeder K. Genomic control for association studies. Biometrics. 1999; 55:997–1004. [PubMed: 11315092]
23. Price A, Patterson N, Plenge R. Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. 2006
24. Kang HM, et al. Efficient control of population structure in model organism association mapping. Genetics. 2008; 178:1709–1723. [PubMed: 18385116]
25. Nelis M, et al. Genetic structure of Europeans: a view from the North-East. PLoS One. 2009; 4:e5472. [PubMed: 19424496]
26. Bustamante CD, Burchard EG, De la Vega FM. Genomics for the world. Nature. 2011; 475:163–165. [PubMed: 21753830]
27. Moran PAP. Notes on continuous stochastic phenomena. Biometrika. 1950; 37:17–23. [PubMed: 15420245]
28. Copeland KT, Checkoway H, McMichael AJ, Holbrook RH. Bias due to misclassification in estimation of relative risk. Am. J. Epidemiol. 1977; 105:488–495. [PubMed: 871121]

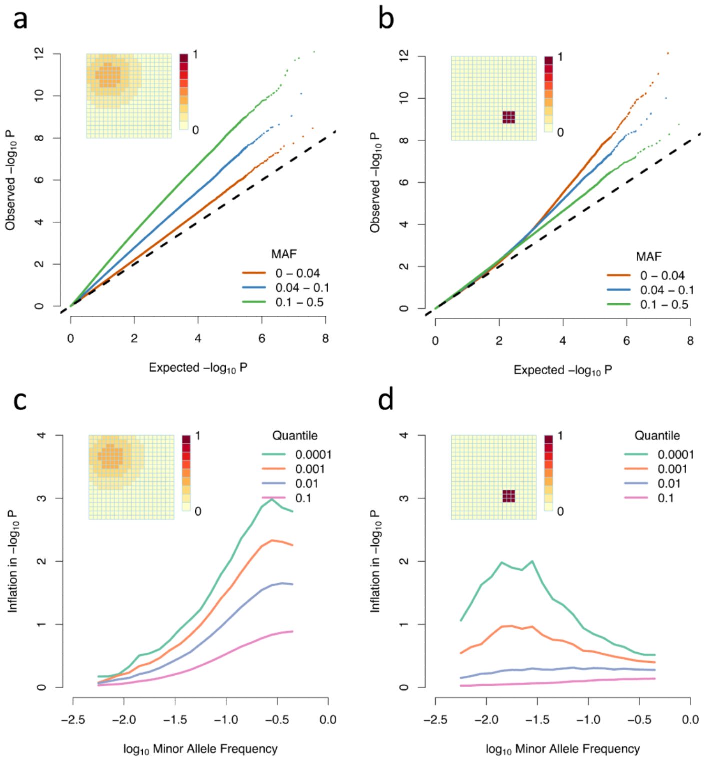

Figure 1. Differential inflation of rare and common variants

(**a**-**b**) QQ plots of association test P-values, broken down by allele frequency for (**a**) a broad, smoothly (Gaussian) varying non-genetic risk factor and (**b**) a small, sharply defined region of constant non-genetic risk; (**c**-**d**) Inflation plots showing the amount by which the observed $-\log_{10}$ P-value exceeds the expected value across allele frequencies. Different lines represent different levels of significance, with $-\log_{10}$ P-value equal to 1,2,3 or 4; The grids in the top left of the pictures represent the spatial distribution of risk and the scale indicates by how many standard deviations the phenotypic mean is shifted in each grid square. The populations simulated here are uniformly distributed over the grid, with two individuals in each square, and a migration rate of 0.01.

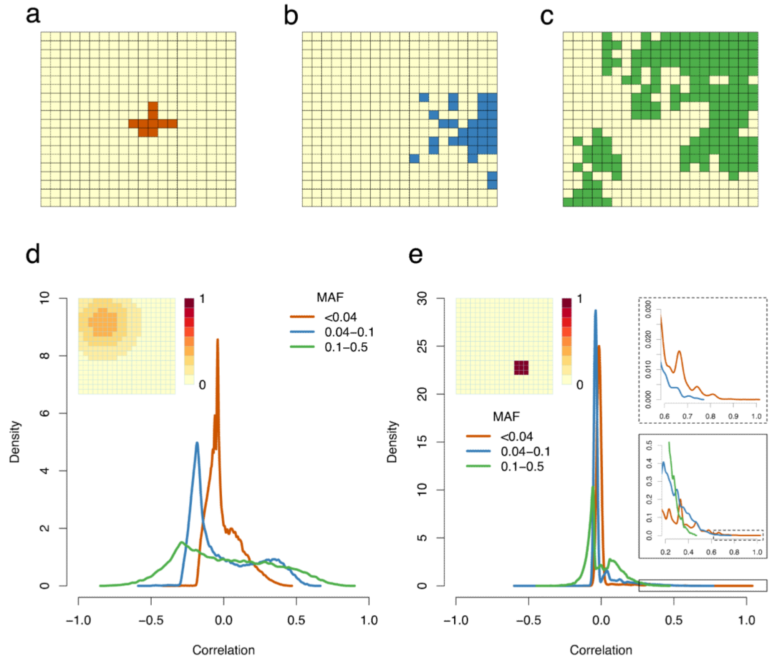

Figure 2. Spatial distribution of rare and common variants

(**a-c**) Examples from simulations of the spatial distribution of (**a**) rare, (**b**) low frequency and (**c**) common variants. In each case, grid squares where the allele is present are in colour; (**d-e**) The distribution of the correlation coefficient between genotypes and non-genetic risk for rare, low frequency and common variants. These are kernel density estimates of the distribution of the correlation between genotypic value (0/1) and associated environmental risk for individuals from the simulations described in Figure 1; (**d**) Gaussian risk; (**e**) Small, sharply defined risk. The inset panels in **e** show successive enlargements of the boxed areas in the tail of the distribution. All parameters are the same as in Figure 1. **Abbreviations**: MAF: minor allele frequency.

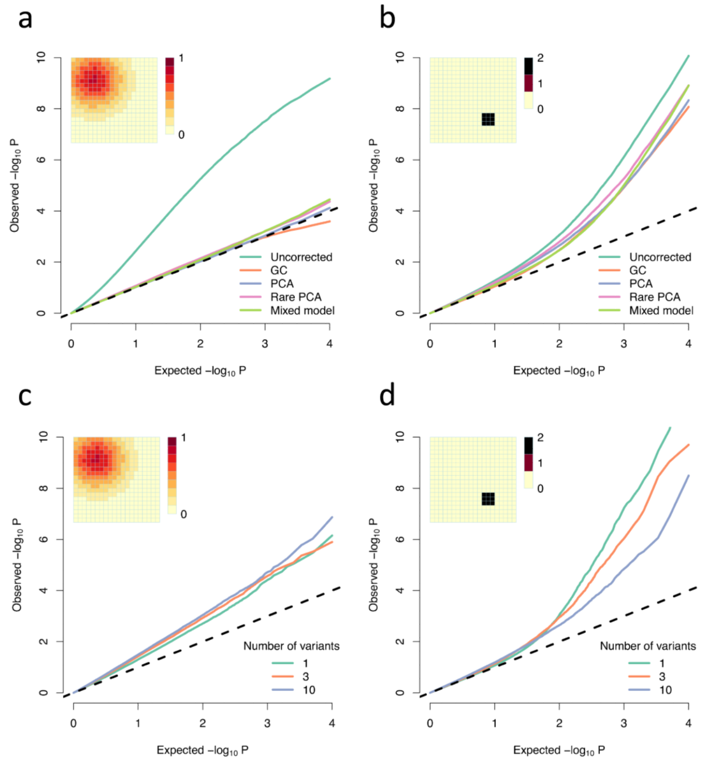

Figure 3. Comparison of methods for correcting for population structure

(**a-b**) QQ plots of $-\log_{10}$ P-values showing the uncorrected values and the values under different corrections; (**c-d**) Simulated rare variant load tests (Online methods); All parameters are the same as in Figure 1, except the non-genetic risk is doubled so for the Gaussian risk **a** and **c** the phenotypic mean is shifted by at most 0.8 standard deviations, while for the small, sharp risk in **b** and **d** it is shifted by at most 2 standard deviations; These are both averaged over multiple simulations in order to show the average effect. Individual experiments may vary due to the sampling variance of the trait. (**a-b**) averaged over 100 simulations, each testing one trait at 10,000 loci in total (10 loci on each of 1000 genealogies, representing independent genomic regions). (**c-d**) averaged over 10 simulations, each one testing 10,000 genealogies with either 1,3, or 10 variants in each; **Abbreviations**: GC, genomic control; PCA principal component analysis, using the first 10 principal components; Rare PCA, as PCA but using only variants with MAF < 4%.

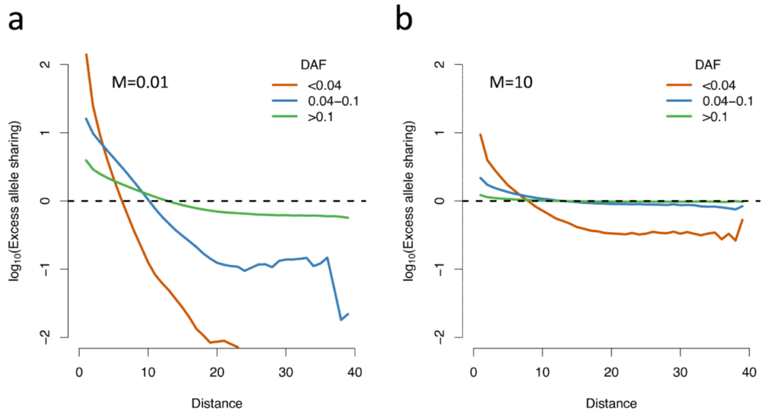

Figure 4. Excess allele sharing

A ratio measuring how much more likely two individuals at a given spatial distance are to share a derived allele, compared to what would be expected in a homogenous population (Methods). The parameters are the same as those used in Figure 1, apart from migration rate, which is (**a**) M=0.01, (**b**) M=10; In **a**, F_{ST}=0.1 and in **b**, F_{ST}<0.01; **Abbreviations**: DAF: derived allele frequency.