Refining bias correction in genome-wide association analyses of case-control studies

Behrooz Darbani[1,2,3]*, Ole Birger Vesterager Pedersen[4,5], Sisse Rye Ostrowski[5,6], Qihua Tan[7,8], Vibeke Andersen[3,9]*

[1]Department of Clinical Research, Hospital Sønderjylland, DK-6200 Aabenraa, Denmark

[2]Department of Regional Health Research, University of Southern Denmark, DK-6200 Aabenraa, Denmark

[3]OPEN, Odense University Hospital, DK-5000 Odense, Denmark

[4]Department of Clinical Immunology, Zealand University Hospital, DK-4600 Køge, Denmark

[5]Department of Clinical Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, DK-2200 Copenhagen, Denmark

[6]Department of Clinical Immunology, Copenhagen University Hospital – Rigshospitalet, DK-2100 Copenhagen, Denmark

[7]Research Unit for Epidemiology, Biostatistics and Biodemography, Department of Public Health, University of Southern Denmark, DK-5000 Odense, Denmark

[8]Unit of Human Genetics, Department of Clinical Research, University of Southern Denmark, DK-5000 Odense, Denmark

[9]Department of Clinical Research, University of Southern Denmark, DK-5000 Odense, Denmark

*Corresponding authors. BD: Department of Clinical Research, Hospital Sønderjylland, Department of Regional Health Research, University of Southern Denmark (Email: bds@rsyd.dk / behroozdarbani@gmail.com), Tel.: +45 – 53578055, VA: OPEN, Odense University Hospital, Department of Clinical Research, University of Southern Denmark (Email: va@rsyd.dk)

Abstract

Genome-wide association studies are vulnerable to confounding factors. This study provides evidence-based guidance for minimizing bias associated with genetic relatedness, SNP-specific non-additive allelic interactions, predisposed genotypes among controls, and multi-allelic polymorphism in case-control studies. The analyses demonstrated that genetic similarity within case or control groups introduces experimental bias, whereas genetic relatedness across case-control samples reduces this bias. These findings establish a general framework that can filter genetically related sub-communities or paired samples, whilst preserving maximal statistical power. Moreover, the skewed odds ratios resulting from predisposed genotypes among controls underscored the importance of age-related filtering to minimize this confounding effect. To ensure accurate genetic estimates, such as polygenic risk scores, the identification of SNP-specific

allelic interaction models was also emphasized in case-control studies, contingent on normalization for within-population differences in genotype frequencies. Finally, a strategy is recommended to capture genetic effects at multi-allelic genomic positions accurately.

Introduction

Genome-wide association studies are designed to identify genetic variants that contribute to trait variation. The advent of advanced sequencing technologies marked a paradigm shift from traditional molecular markers, such as Restriction Fragment Length Polymorphism (RFLP), to whole genome sequences (1–4). This transition facilitated the identification of millions of polymorphic genomic loci, driving a surge in genome-wide association studies. Despite recent advancements in genomic and genotypic coverage, i.e., analyzing large cohorts using millions of single-nucleotide polymorphisms (SNPs) (5–11), methodological considerations in genetic data analysis are indispensable to reduce bias and thereby, the number of false-positive and negative SNP hits. This study outlines strategies to maximize statistical power for the discovery of causal SNPs.

Like population-level genetic differences (12), individual-level genetic similarities can bias genetic assessments. Genetic relatednesses, such as pedigree and cryptic similarities, have been identified as major confounders in genome-wide association studies, leading to inflated type I errors (13–20). Although mixed models (18, 20–27) have been introduced to control for sample relatedness, sample filtering by retaining one individual per related group of individuals has been the most common approach (28–35). In contrast, family-based association studies can offer greater statistical power than case-control association studies of genetically unrelated individuals (36–42). However, the advantage of family-based association studies can be easily undermined by the typically small sample sizes inherent to such designs. In this study, we explored optimal strategies for filtering genetically related samples to control the incidence of false SNP hits effectively.

The manifestation of specific traits, such as the onset of some diseases, occurs at different ages throughout the human lifespan. This can introduce bias in case-control genome-wide studies

through recruiting individuals mistakenly classified as controls, e.g., healthy individuals who do not have a given chronic disease but carry causal genetic elements that will express phenotypically at higher ages. This impurity among controls, i.e., the presence of predisposed genotypes, can lead to inaccurate genetic measures such as underestimated odds ratios. The presence of predisposed genotypes is independent of sampling and can occur under either sampling strategy, i.e., active participation (non-random) or passive participation used to establish cohorts. Accordingly, methods that address active participation-mediated volunteer bias (34), also known as participation bias (43), do not interfere with bias mediated by predisposed genotypes. Furthermore, the GWAS-by-proxy (GWAX) methodology (44), which uses parental diagnostic history, can potentially introduce biases due to uncertainty in phenotypic concordance between parents and offspring. Accordingly, the application of GWAX has been found to generate misleading outcomes (45). Here, we propose a filtering strategy to reduce impurity among control samples within cohorts.

Another important aspect that can easily influence the accuracy of genetic measures is the type of interaction between the reference and alternative alleles at a given polymorphic genomic position. These interactions should be distinguished from epistatic interactions among SNPs (46, 47). Recent studies have reported dominance effects in mammals (48) and humans (49). In contrast to pervasive non-linear interactions among genes or SNPs (47, 50), there has been a widespread assumption that SNP-specific non-additive allelic effects are negligible (51). Nevertheless, this assumption does not rest on any inherent genetic or biological principles. Sample-size enlargement is required for capturing dominance effects efficiently at the levels observed for additive effects (52). As illustrated, compared with the observed allelic effects, the estimated small non-additive allelic effects may be attributable to methodological limitations and artefacts (53). Here, we highlight and adjust the application of recently introduced observed-level allelic effects (53) to efficiently capture additive and non-additive allelic effects, including dominance, overdominance, and heterosis-like effects, in case-control studies. We also shed light on the importance of normalizing differences in genotype frequencies to accurately capture additive and non-additive effects in case-control studies, such as those with binary phenotypes. We finally provide a brief methodological discussion on extracting the effect of each alternative

allele at multi-allelic genomic positions. Each allele at a multi-allelic site is often encoded relative to all other alleles at that specific genomic position, which can hinder the accurate estimation of allele-specific genetic effects.

Results

Genetic relatedness and sample filtering

Within-cohort genetic relatedness is recognized as a major confounding factor, leading to inflated type I error rates in genome-wide association studies (13–20). It has been a common strategy to avoid erroneous association outcomes by genetic relationship inference involving control and case samples, i.e., filtering out individuals with second-degree or higher levels of genetic relatedness and retaining only one individual per group of related individuals, with priority given to the case samples over controls (28–35). As our results also prove (Figure 1-3), we must note that the confounding effects of relatedness are relevant only for the case-samples or control-samples specific relatedness, and not for relatedness across case and control groups. Sample-relatedness confined to either the case or control group can artificially inflate the frequency of unassociated alleles within the respective group, thereby generating spurious association signals. In contrast, the relatedness across case-control samples provides an even distribution of unassociated alleles between case and control groups; statistically significant association signals from unassociated SNPs are a common problem arising from low genotypic coverage. Therefore, genetic relatedness across case-control samples not only does not lead to false positives but also reduces them, mimicking the effect of family-based genetic studies. This mimicked effect is similar to the statistical power superiority reported for family-based genetic studies (36–42, 54), assuming equivalent sample sizes.

Figures 1-3 illustrate the advantage of genetic relatedness across case-control samples, in contrast to the disadvantages of case-samples specific or control-samples specific genetic relatedness in genome-wide association studies. Our results demonstrate the usefulness of relatedness between case and control samples for reducing false positive rates (Figure 1). Interestingly, a subset of 100 genetically related samples across case and control groups, accounting for just 1% of the cohort, exhibited a lower false-positive rate (Figure 1A, 2A). A sharp

reduction in false positive rate was also observed in the presence of genetic relatedness across case and control groups, including all samples (Figure 1F, 2A). On the other hand, including ≥ 500 case or control group-specific genetically related paired samples, i.e., comprising ≥ 10% of that group, did raise the false-positive rate (Figure 2). Small genetically related sub-communities containing up to 25 related samples also did not elevate the false-positive rate (Figure 3). Taken together, these results provide a general guidance to handle the sample relatedness within cohorts via sample filtering: i) sample relatedness across case-control groups reduces false positive rates and thus obviates the need for filtering; ii) genetically related small sub-communities, e.g., sub-communities constituting 0.25% of controls or cases, do not lead to false positive SNP hits; and, iii) genetically related sample pairs confined to either the case or control group can lead to false-positive associations and warrant filtering when they constitute ≥ 10% of their respective group, i.e., case or control. For these simulations, we used a relatively high level of relatedness, i.e., 25% genetic similarity, among all the related samples, and thereby, the mentioned thresholds of 0.25% and 10% are quite stringent; higher threshold levels can be considered acceptable in many studies having lower levels of genetic relatedness among individuals.

Sample filtering has been the most common approach for controlling relatedness (28–35), and one of the primary goals of this study was to establish an efficient sample filtering strategy. For the purpose of comparison, we also performed analyses using the mixed-model fastGWA-GLMM (23) to control for relatedness without the need for sample filtering. As expected, the mixed-model results were comparable to those from logistic regression when analyzing the dataset with no relatedness (Figure 2A, S1A) or with relatedness across case-control groups (Figure 1F, S1B). The analyses by fastGWA-GLMM also demonstrated effective control of relatedness without sample filtering when we examined the previously described cohort, which contained a genetically related sub-community of 50 related samples within either the control or case group (Figure S1A,C). It is important to note that when mixed models are used to account for sample-relatedness, scenarios with case-specific or control-specific relatedness are also expected to yield similar numbers of random association hits as the corresponding scenarios lacking relatedness. We nevertheless observed an elevated false-negative risk when applying the mixed model to

examine the previously described cohort with 1,000 genetically related paired samples within either the control or case group (Figure S1D). Specifically, the number of random associations with *p*-values less than 10^{-8} decreased from 90 in the scenario without related samples (Figure 2A and Figure S1A) to 58 in the scenario with 1,000 genetically related paired samples (Figure S1D), despite no underlying reason, i.e., without any related samples being shared between the case and control groups. We additionally detected a markedly increased number of random associations upon applying the mixed model following sample filtering (Figure S2). These analyses indicate that there is no single prescribed approach for choosing between sample filtering and mixed models for controlling relatedness. When applying the commonly used approach of sample filtering (28–35), our study recommends avoiding the removal of related samples across case and control groups.

The effect of predisposed genotypes, i.e., individuals among controls with yet not expressed but inherited causal genetic variants, on the accuracy of genetic measures

Age has been commonly used as a covariate in numerous case-control genome-wide association studies. Nevertheless, it is important to note that covariates should be preventive or promotive factors imposing independent effects on outcomes (55–57). In the absence of a standalone impact of age or other variables on the outcome of interest, including these variables as covariates is therefore unwarranted, as it may unnecessarily reduce statistical power (55, 56). In conditions like age-independent disease incidence and in cohorts with wide age ranges, impurity among control samples, i.e., individuals with yet not expressed causal genetic variants, is expected. Therefore, implementing an age cutoff for filtering of control samples, e.g., excluding individuals under 65 years of age, can effectively reduce the impurity levels. It is worth noting that including age as a covariate cannot be considered a corrective measure for predisposed genotypes, as it does not eliminate or reduce the bias introduced by them. The level of impurity among controls depends on the disease prevalence, the disease incidence pattern across different age groups, and the age range and distribution of control individuals within cohorts.

Table 1 provides a simulation of age-independent disease incidence across 10-year age intervals. Here, we used the world population statistics and assumed the same age distribution within the

cohort, as well as an even, i.e., age-independent, incidence rate of 12.5% of disease prevalence rate for each 10-year age interval. These conditions result in a control group containing the predisposed genotypes at rates up to ≈ 50% of the population disease prevalence (number of predisposed genotypes, i.e., samples, within the control group = 0.5 × DP × N; DP: disease prevalence, N: total number of samples within the control group). As already mentioned, predisposed genotypes harbour causal genetic variants and will later in life develop the case group phenotype.

Therefore, reducing the impurity by filtering out younger individuals from the control group is a reasonable strategy. This typically demands a much larger control group than the case group, a condition common in most genetic cohorts. To demonstrate the potential bias introduced by impurity in control groups and the filtering-by-age efficiency of control samples, we simulated case-control data, based on the information from Table 1 and as discussed in Methods and Text-Box S1. Our analyses revealed a downward bias in odds ratio estimates due to impurity in the control group within cohorts (Figure 4). Simulations at four different population disease prevalence rates of 15%, 10%, 5%, and 1%, each at 9 different SNP-phenotype association levels, revealed the inaccuracy of the estimated odds ratios for the associated SNP (Figure 4). The analyses also indicated the effectiveness of age-dependent filtering within cohorts (Figure 4).

It is important to note that the proposed age-filtering method is incompatible with experiments that do not apply the commonly used 1% MAF filter. Cohorts may contain individuals with another specific disease, independent of the study's disease of interest, which causes early death. Therefore, it is important to highlight that the proposed age-dependent filtering inadvertently depletes the control group of individuals affected by a given disease associated with early mortality. Nonetheless, this should not pose a potential risk of false associations, as a residual number of individuals with an independent second disease linked to early death would still be present among the cases. For example, if the independent second disease, which is entirely associated with early death but excluded from the control group due to age-based filtering, has a prevalence of less than 2%, then the likelihood that cases experienced early death due to this independent disease would also be less than 2%. It should be noted that a 2% prevalence for a disease that always leads to early death is considered exceptionally high and overestimates real-

world scenarios. Consequently, genomic variants linked to early death exhibit very low allele frequencies within cohorts; a likelihood of less than 2% among cases is equivalent to a minor allele frequency of less than 1% in a cohort with an equal number of cases and controls. Accordingly, early-death-associated genomic variants are expected to be excluded by applying a quality filter of 1% minor allele frequency.

Furthermore, the death rate among individuals under 65 years old accounts for 40% of all human mortality (Data from Our World in Data at https://ourworldindata.org/grapher/annual-deaths-by-age). Given the same 40% premature mortality rate also among cases, a causal variant would need to contribute to at least 5% of these deaths to have a minor allele frequency of larger than 1% in a cohort with an equal number of cases and controls, following age-related filtering of the control group ($0.05 \times 0.40 \div 2 = 0.01$, the division by 2 accounts for the rate being halved after combining the control and case groups). This is also an exceptionally rare condition and overestimates real-world scenarios. Among chronic diseases, only ischaemic heart disease and stroke have global mortality rates exceeding 5%, i.e., account for more than 5% of deaths, including both premature and old-age deaths (6, 58, 59). However, these diseases are also highly polygenic, with hundreds of genetic variants being responsible for less than 50% of disease events (6, 58). Taken together, false positive hits originating from the causal genetic variants of early death are not expected after age-related filtering of control groups. This is because a minimum 5% contribution to total deaths in a monogenic scenario and a significantly higher contribution rate in an oligogenic scenario are required for the death-associated genetic variants to pass the minimum 1% minor allele frequency filter. For a polygenic disease that causes early death through hundreds of genetic variants, it is almost impossible to find a causal genetic variant that passes the minimum 1% minor allele frequency filter after age-related filtering of the control group.

It is worth noting that the death rate among individuals under 65 years old in developed countries is significantly lower than the global premature death rate of 40%. This rate for countries like Denmark, Finland, Germany, Italy, Japan, Spain, Sweden, Switzerland, and United Kingdom is below 15% (Data from Our World in Data at https://ourworldindata.org/grapher/annual-deaths-by-age and from Statistics Denmark at https://www.dst.dk/en/Statistik/emner/borgere/befolkning/doedsfald). This

underscores the advantage of our proposed age-related filtering: the lower the death rate among individuals under 65, the greater the likelihood that the minor allele frequency filter will exclude variants associated with early death. With 15% of the total deaths occurring at age 0–64, even the extreme assumption that the polygenic ischaemic heart disease which is responsible for 13% of the world's total deaths (59), is monogenic and that all its related death occur before age 65, cannot lead to false positive hits due to the exclusion of the associated variant after applying 1% minor allele frequency filter ($0.15 \times 0.13 \div 2 = 0.00975 < 1\%$ MAF).

Non-additive interactions between reference and alternative alleles are a potential source of bias when genetic measures are estimated using the additive-only modeling assumption

Most genome-wide association studies have focused on additive allelic effects, largely because of the prevailing assumption that non-additive effects are negligible (51). It is important to note that capturing additive-only effects in the presence of both additive and non-additive allelic effects can introduce varying degrees of bias, potentially resulting in false SNP hits and inaccurate genetic measures. We simulated genome-wide association studies (see Text-Box S1 and section Methods) at five different association levels, not only for the additive allelic effect scenario but also for dominance, overdominance, and heterosis-like non-additive allelic effects as described previously (53). In the presence of non-additive allelic effects, restricting analysis to additive effects led to inaccurate risk scores for genotypes carrying the causal allele, as observed across five association levels and two disease prevalence rates (Table 2). However, this is not the case when applying the allelic adjustment method, which estimates the true allelic effect using the lowest association *p*-value for each SNP across different effect-model scenarios (53).

To jointly capture additive and non-additive allelic effects, we applied the SAFE-h^2 allelic adjustment approach, which uses observed allelic effects and subsequent allelic adjustments to fit the observed allelic effect-models (53). This means that every SNP is analyzed according to its specific allelic interaction model, i.e., whether additive or one of the non-additive models. The advantage of the mentioned allelic adjustment approach is its reliance on a linear association model while effectively capturing both additive and non-additive allelic effects. Here, it is important to highlight the methodological difference between quantitative and qualitative traits.

For quantitative traits, it is the quantities, including every observation, that influence the regression output. In contrast, the frequency of each phenotypic category plays an important role in the logistic regression output. Logistic regression, which is used for qualitative traits, is therefore sensitive to the frequency of each phenotypic category relative to the genotypes, i.e., the three SNP-specific genotypes (Ref/Ref, Ref/Alt, and Alt/Alt) observed for both the control and case groups. Due to differences in genotype (Ref/Ref, Ref/Alt, and Alt/Alt) frequencies within the population, the sensitivity to the frequency of phenotypic categories hinders the identification of the best-fitted model when comparing allelic interaction models. Therefore, we introduce a normalization step which is essential when analyzing qualitative phenotypes to account for differences in genotype frequencies within the population. This process ensures a fair comparison of allelic interaction models, enabling the identification of the model that most accurately represents the observed data and provides the best fit. For normalization, genotype frequencies within the population are required, as explained in section Methods and Text-Box S2. Assuming known population allele frequencies, we can calculate population genotype frequencies (see Text-Box S1) and use these values to normalize the observed genotype frequencies (see Methods and Text-Box S2). If within-population allele frequencies are unknown, the population disease prevalence can be used to build iterative control-case combinations, each representing the original population. Thus, the average genotype frequencies can be used as estimates of population genotype frequencies. Please note that normalization is relevant only when comparing allelic interaction models to identify the best model for each SNP. The original unnormalized data should be used, as previously described, for the allelic adjustment strategy to obtain genetic measures for SNPs with additive allelic effects and for SNPs exhibiting non-additive allelic effects.

In contrast to the common approach, i.e., additive-effects based association analysis, which performs poorly in the presence of non-additive allelic interactions and yields inaccurate risk scores, the SAFE-h^2 allelic adjustment approach enabled clear differentiation of the allelic interaction models (Table 2). Here, we simulated and applied predesignated true associations using various genetic effect models. By evaluating different effect-model scenarios for every SNP, the lowest association *p*-values were obtained in the scenario that matched the true allelic effect

(Table 2). In Table 2, the expected (true) risk scores also corresponded to those obtained using the allelic adjustment method. The results indicate that the allelic adjustment strategy, when combined with normalization can be applied in case-control studies to achieve greater accuracy. SAFE-h^2 performs model selection based on association *p*-values, effectively distinguishing among additive and various non-additive genetic effects at varying association levels. In contrast to association *p*-values (Table S1), the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) failed to distinguish between predesignated genetic models of additive and overdominance (Table S2).

The normalization of genetic data to correct for differences in population genotype frequencies and the subsequent logistic regression for every SNP are considerable tasks. A more practical and efficient approximation is to leverage the case-to-control ratios for each genotype (Ref/Ref, Ref/Alt, & Alt/Alt). Afterwards, we can define quartiles (Q1 to Q4) by dividing the range of case-to-control ratios into four equal parts. Finally, the distribution of genotypes across these quartiles can be used to define allelic effects as shown in Table 2. We found complete agreement between the association *p*-value obtained after normalization of genotype frequencies and the case-to-control ratios (Table 2 and Table S3).

Accurate analysis of multi-allelic genomic loci necessitates decomposing variants into individual allelic comparisons relative to a defined reference allele

In comparison to biallelic SNPs, there is a higher complexity in the association analyses of multi-allelic SNPs. Higher levels of genotypic diversity cause this complexity, which can be further exacerbated by non-additive allelic effects or by a combination of additive and non-additive allelic effects. Given the same cohort size, the greater genotypic diversity at multi-allelic sites also leads to lower statistical power in association analysis relative to biallelic SNPs. Simultaneous association analysis for all alleles of a multi-allelic SNP demands complex models. The most common coding format for multi-allelic SNPs only allows comparison of each allele versus the remaining alleles collectively as the reference. This particular allele coding format may introduce an additional source of bias. Comparing every allele versus the remaining alleles of a multi-allelic genomic position can elevate false negative rates and compromise the accuracy of estimates for

the allele's genetic effects. The expected inaccuracy is due to possible contrasting genetic effects among the alleles, collectively defined as the reference allele. Therefore, an ideal solution is to set one of the alleles, e.g., the most frequent allele, as the reference allele and perform association analyses for the remaining alleles, each versus the reference allele. In addition to the association analyses using a collective reference allele for multi-allelic genomic positions, we accordingly recommend performing an association analysis for each allele in a multi-allelic genomic position versus a defined single-allele reference. This is feasible through reformatting the dataset to decompose multi-allelic genomic positions. For example, four association analyses are performed for a genomic position with four alleles. In all these four association analyses, the reference allele is a combination of three alleles. In this example, we recommend conducting three additional association analyses, each using the most frequent allele as the reference, and comparing it individually against each of the other three alleles.

Discussion

A key challenge in genome-wide association studies is the low statistical power due to insufficient genotypic coverage; approximately 3^N individuals with distinct genotypes are required for precise genetic estimation in a diploid genome with N biallelic SNPs, given that mutations are independent and conditional lethal SNPs are exceedingly rare. Meeting such cohort sizes is unfeasible due to the relatively small size of the human population. For instance, insufficient genotypic coverage is the reason for regional association signals from multiple neighbouring SNPs, and the developed SNP clumping and fine-mapping strategies (60–65) only partially overcome this challenge by identifying the most likely causal SNP or SNP_s within each regional association signal. Replicating experiments, e.g., using effect-sizes of associated SNPs obtained from one cohort to calculate polygenic risk scores for individuals in a second cohort, is also a common approach to mitigate overfitting caused by inadequate genotypic coverage. Other factors may also introduce bias into the experiments, e.g., the misclassification of predisposed genotypes as control samples, genetically related samples within cohorts (13–20), and non-additive allelic interaction (53). Unlike the challenge posed by insufficient genotypic coverage,

which is an insurmountable obstacle, the remaining factors can be efficiently addressed to reduce bias in genetic assessments.

Our results indicate the disadvantage of the standard bulk filtering procedure for genetically related individuals within cohorts. We present evidence that genetic relatedness across case-control groups effectively reduces false positive rates and, therefore, should not be removed through bulk filtering. Ideally, the only source of variation among experimental groups or treatments should be the phenotype of interest. We often include covariates to account for known sources of variation, allowing for a less biased comparison of samples with respect to the phenotype of interest. Genetic relatedness across case-control groups serves a similar purpose by controlling for both known and hidden sources of variation, thereby establishing an experimental framework for more robust and unbiased evaluation of the phenotype of interest among samples. In other words, genetic relatedness between case and control groups selectively reduces allele-frequency differences for non-causal variants, where all observed disparities are attributable to experimental bias. Considering 25% genetic relatedness, we found no bias leading to false-positive hits attributable to genetically related sub-communities constituting up to 0.25% of controls or cases, or by genetically related paired-samples constituting less than 10% of the case or control groups. The mentioned thresholds of 0.25% and 10% impose a high degree of stringency, given that the simulated 25% genetic similarity exceeds typical expectations and represents the upper-bound of relatedness levels in real-world cohorts.

The proposed age-dependent filtering also effectively mitigated bias caused by predisposed genotypes that were incorrectly classified as control samples. Age-based filtering is applicable when age is not a covariate and, as a characteristic commonly observed in contemporary genetic cohorts, the control group is substantially larger than the case group. This method can thus be effectively employed to correct for bias arising from predisposed genotypes. In general, age-based filtering is incompatible with experiments that omit the commonly used 1% MAF filter, as this may lead to false positives, particularly for phenotypes with high within-population prevalence. Nevertheless, age-based filtering can still be applied in rare variant analyses by conducting a complementary GWAS using age-based filtered samples, i.e., deceased samples versus the remaining filtered samples. This complementary GWAS can identify potential variants

linked to premature death that might otherwise yield false positives after age-based filtering in rare variant analyses.

Cox regression is a robust method for analyzing time-constant variables that influence time-to-event outcomes (66, 67). When disease age-of-onset is defined as a time-to-event phenotype to perform survival analysis in GWAS, individuals/genotypes are treated as components of a time-constant variable. Here, special attention should be given to the age of onset, as it requires a strong correlation between disease incidence and specific ages. In other words, defining age-of-onset as a time-to-event phenotype requires that age serve as a covariate that mediates the genetic influence on disease incidence. Otherwise, disease incidence across all ages, i.e., with no clear correlation or enrichment at specific ages, reflects the dependence of disease age-of-onset on non-genetic factors. For example, we have found an even distribution of incidence for inflammatory bowel disease across all ages (54). Notably, Cox regression cannot account for predisposed genotypes, i.e., impurity in control groups, where disease incidence is driven by non-genetic factors. Predisposed genotypes consistently lead to underestimation of effect sizes; consequently, accurate inference methods should adjust for this bias and report appropriately larger effect sizes. However, Cox regression often leads to smaller effect sizes. For instance, when conducting association analyses for coronary heart disease, Cox regression yielded smaller effect size estimates compared to logistic regression (66). In agreement, the hazard ratios from Cox regression on the simulated data (see figure 4 for the dataset) were also smaller than the odds ratios (Figure S3).

Genome-wide association studies have predominantly modeled additive allelic effects, reflecting the commonly held assumption that non-additive effects contribute minimally to trait variations (51). However, there is no inherent genetic or biological rationale to rule out non-additive allelic effects. Given the roles of evolutionary processes through random genetic mutations and the complexity of molecular interactions at all levels that modulate biosynthetic, structural, and compositional properties, the occurrence of non-additive allelic effects is highly plausible. Here, as an example, we refer to two recent studies on dominance effects in mammals (48) and humans (49). Methodological limitations have been introduced as the main driver for the trivial estimates of non-additive allelic effects (52, 53). We further introduce the impact of population-level

frequency disparities among SNP genotypes (Ref/Ref, Ref/Alt, Alt/Alt), which complicate the identification of allelic interaction modes for individual SNPs via logistic regression in case-control analyses. The results indicate that a pre-normalization step for SNP-specific genotype frequencies enables accurate detection of allelic interaction modes and thereby reduces bias in capturing phenotypic variance. It is important to note that capturing additive-only effects in the presence of additive and non-additive allelic effects can introduce varying degrees of bias, potentially resulting in false SNP hits and inaccurate genetic measures.

Materials and methods

Simulation of data

We used LDAK to simulate unrelated genotypes (68). Afterwards, unrelated individuals were applied to simulate genetically related individuals. We replicated ≈ 25% of the SNPs from one individual in an unrelated individual to obtain two genetically related individuals. As expected, genetic similarity among related individuals was 24.8%–25.1%, calculated using the KING-robust kinship estimator of PLINK (14). To analyze the relatedness-based filtering, case and control allocations were randomized across individuals. Therefore, any increase in the number of significant SNP hits, e.g., due to the inclusion of genetically related case or controls, indicates an elevated false-positive rate. On the other hand, any decrease in the number of significant SNP hits, e.g., due to the inclusion of individuals that introduce genetic relatedness across case and control groups, indicates an reduced false-positive rate. However, any decrease in the number of significant SNP hits in the absence of an underlying cause, i.e., without any related samples being shared between the case and control groups, is an indication for increased false negative rates. We performed the analyses in 60 independent replicates. Association analyses were performed using the GLM function of PLINK (69). Minor allele frequencies were ≥ 0.01 in all the analyses. For mixed model based analysis, GCTA-fastGWA-GLMM was used (23).

The age distribution of the simulated cohorts was based on the world age distribution, as shown in Table 1. Varying levels of true association were simulated by applying a factor, *Z*, representing the percentage of disease prevalence attributable to the causal variant. Disease prevalence rates were simulated at 15%, 10%, 5%, and 1%, with corresponding causal allele frequencies of 22.5%,

15%, 7.55%, and 1.5%, respectively. The simulations required causal allele frequencies to exceed disease prevalence to account for predisposed genotypes, i.e., individuals carrying the causal alleles but not yet affected, within the control group. The calculation of population and cohort level genotype frequencies for genotypes within population (XX_P), genotypes of healthy controls after excluding cases from the population (XX_{PH}), case genotypes within the cohort (XX_C), control genotypes within the cohort (XX_H), and control genotypes within the cohort after removing individuals under 65 years of age (XX_H) were performed as follows, in steps 1 to 5:

1. Population (*Disease prevalence = 1%, Frequency of Reference allele A = 0.985 & frequency of causal allele C = 0.015*)

Genotype frequencies →

$AA_P = (0.985 \times 0.985)$

$AC_P = 2(0.985 \times 0.015)$

$CC_P = (0.015 \times 0.015)$

Patient and non-patient frequencies →

$Healthy_P = 0.99$, Healthy with predisposed genotypes = 0.005

$Patients_P = 0.01$

2. Calculating genotype (i.e., AA, AC, & CC) frequencies for cases witihin cohort (*n = 10,000: 0.0% controls + 100% cases, proportion of cases being associated with allele C = Z*)

Causal-allele C independent cases →

$AA_C = (AA_P \div (AA_P + 0.5AC_P)) \times n(1-Z)$

$AC_{C1} = (0.5AC_P \div (AA_P + 0.5AC_P)) \times n(1-Z)$

Causal-allele C dependent cases →

$AC_{C2} = (0.5AC_P \div (CC_P + 0.5AC_P)) \times nZ$

$CC_C = (CC_P \div (CC_P + 0.5AC_P)) \times nZ$

3. Calculating genotype, i.e., AA, AC, & CC, frequencies for pure (with no predisposed genotype) controls (*n = 100,000: 100% controls + 0.0% cases*)

$AA_{PH} = ((100{,}000 \times AA_P - 0.1AA_C)) \times 10 \div 9.9$

$AC_{PH} = (((100{,}000 \times 0.5AC_P) - (0.1AC_{C1})) \times 10 \div 9.9) + (((100{,}000 \times 0.5AC_P) - (0.1AC_{C2})) \times 10 \div 9.9)$

$CC_{PH} = ((100{,}000 \times CC_P) - (0.1CC_C)) \times 10 \div 9.9$

4. Calculating genotype, i.e., AA, AC, & CC, frequencies for controls within cohort (*n = 100,000: 99.5% controls + 0.5% cases*)

$AA_H = 0.995AA_{PH} + 0.05AA_C$

$AC_H = 0.995AC_{PH} + 0.05AC_{C1} + 0.05AC_{C2}$

$CC_H = 0.995CC_{PH} + 0.05CC_C$

5. Calculating genotype, i.e., AA, AC, & CC, frequencies for filtered controls (*age ≥ 65 years*) *within cohort* (*n = 10,000: 99.99% controls and 0.01% cases*)

$AA_H = 0.0999AA_{PH} + 0.001AA_C$

$AC_H = 0.0999AC_{PH} + 0.001AC_{C1} + 0.001AC_{C2}$

$CC_H = 0.0999CC_{PH} + 0.001CC_C$

The same procedure was applied for simulations with disease prevalences of 5%, 10% and 15%.

Additive and non-additive allelic interactions

Here, we adopted the approach described above to simulate and apply predesignated true associations, this time using various genetic-effect models. We used SAFE-h^2 for joint-capturing of additive and non-additive allelic effects (53). This is done through allelic adjustments for linear approximation of allelic effects, including dominance for either the reference or alternative allele, overdominance for either the reference or alternative allele, and heterosis-like effects. When linear models are augmented with some allelic adjustments tailored to the observed non-additive effects, e.g., by converting the heterozygote nucleotide positions into one of the homozygotes at

the genomic positions with dominance allelic effects, we can capture genetic variance at levels comparable to those of the exponential model. SAFE-*h*2 employs the same allelic adjustment algorithm to extract non-additive effects within the framework of linear models as described previously (53). Briefly, SAFE-h^2 examines all scenarios, i.e., additive (original genotypes), adjusted genotypes for dominance effects, adjusted genotypes for overdominance effects, and adjusted genotypes for heterosis-like effects, to find the best-fitted linear model for every SNP position. Before association analysis, combinatory genotype datasets can be built based on a selected adjustment or the original genotype, (i.e., the one with the smallest association *p*-value, for each SNP position). In case-control studies employing logistic regression for genome-wide association analyses, the SAFE-*h*2 allelic adjustment algorithm requires a pre-normalisation step. In contrast to the linear regression on quantitative phenotypes, where the quantity of every genotype is highly influential, logistic regression is sensitive to the frequency of each phenotypic category in relation to the genotypes, i.e., SNP specific three genotypes of Ref/Ref, Ref/Alt, Alt/Alt among control samples and among case samples. Therefore, population level frequency differences among SNP specific genotypes of Ref/Ref, Ref/Alt, and Alt/Alt can hinder the identification of the best fitted model when comparing allelic interaction models through logistic regression. Therefore, a normalization step is necessary when analyzing qualitative phenotypes to account for genotype frequency differences within population before being able to compare the allelic interaction models and identify either additive or one of the non-additive models that provides the best fit. The correction for genotype frequency differences can be carried out as outlined below:

Considering XX_P for genotype frequencies within population, XX_C for case-samples specific genotype frequencies within the cohort, XX_{CN} for normalized case-samples specific genotype frequencies within the cohort, XX_H for control-samples specific genotype frequencies within the cohort, and XX_{HN} for normalized control-samples specific genotype frequencies within the cohort, we correct the frequencies as follows, in steps 1 and 2:

1. Population (*Reference Allele frequency = R, Alternative Allele frequency = A*)

$RR_P = R \times R$

$RA_P = 2 \times R \times A$

$AA_P = A \times A$

When within-population allele frequencies are unknown, the population disease prevalence can be used for iterative control-case combinations, each representing the original population. In this way, the average genotype frequencies will provide estimates for population genotype frequencies.

2. Cohort

2-1. Observed genotype frequencies among controls within the cohort:

RR_H, RA_H, AA_H

Normalized genotype frequencies among controls within the cohort (Assuming R > A):

$RR_{HN} = RR_H$

$RA_{HN} = RA_H \times (RR_P \div RA_P)$

$AA_{HN} = AA_H \times (RR_P \div AA_P)$

2-2. Observed genotype frequencies among cases within the cohort:

RR_c, RA_c, AA_c

Normalized genotype frequencies among cases within the cohort (Assuming R > A):

$RR_{CN} = RR_C$

$RA_{CN} = RA_C \times (RR_P \div RA_P)$

$AA_{CN} = AA_C \times (RR_P \div AA_P)$

This normalization creates a new cohort dataset that can be used to determine the best-fitting interaction models for alleles of each SNP using the SAFE-h^2 allelic adjustment algorithm. Finally, the identified interaction models for each SNP can be applied to the original dataset (not normalized) using the SAFE-h^2 allelic adjustment algorithm to obtain accurate genetic measures.

As an efficient approximation of the regression line, we also used the case-to-control ratios (CCR) for SNP-specific genotypes ($CCR_{Ref/Ref}$, $CCR_{Ref/Alt}$, $CCR_{Alt/Alt}$). In rare instances of complete association or when predisposed genotypes are absent among control samples, certain genotypes may be missing from either the control or case group. To address this and avoid zero or undefined ratios, a small numeric constant, e.g., 10^{-6}, should be introduced before obtaining ratios. Four quartiles were defined for the whole range of CCR, i.e., Max_CCR − Min_CCR. Finally, the distribution of SNP-specific $CCR_{Ref/Ref}$, $CCR_{Ref/Alt}$, and $CCR_{Alt/Alt}$ ratios across these quartiles was used to define allelic effects as shown in Table 2. The additive model can be confirmed when $CCR_{Ref/Ref}$, $CCR_{Ref/Alt}$, and $CCR_{Alt/Alt}$ are distributed respectively in Q1, Q2-Q3, and Q4, or alternatively in Q4, Q2-Q3, and Q1. The dominance model can also be confirmed when $CCR_{Ref/Ref}$ falls within Q1, while both $CCR_{Ref/Alt}$ and $CCR_{Alt/Alt}$ fall within Q4, or when $CCR_{Ref/Ref}$ falls within Q4, while both $CCR_{Ref/Alt}$ and $CCR_{Alt/Alt}$ fall within Q1. Alternatively, the dominance model can be supported when $CCR_{Ref/Ref}$ and $CCR_{Ref/Alt}$ fall within Q1 while $CCR_{Alt/Alt}$ falls within Q4, or when $CCR_{Ref/Ref}$ and $CCR_{Ref/Alt}$ fall within Q4 while $CCR_{Alt/Alt}$ falls within Q1. The overdominance model is confirmed when $CCR_{Ref/Ref}$, $CCR_{Alt/Alt}$, and $CCR_{Ref/Alt}$ are distributed respectively in Q1, Q2-Q3, and Q4; or in Q2-Q3, and Q4, and Q1; or alternatively in Q4, Q2-Q3, and Q1; or in Q2-Q3, Q1, and Q4. The heterosis-like model can also be supported when $CCR_{Ref/Ref}$ and $CCR_{Alt/Alt}$ fall in Q1 while $CCR_{Ref/Alt}$ in Q4, or alternatively when $CCR_{Ref/Ref}$ and $CCR_{Alt/Alt}$ fall in Q4 and $CCR_{Ref/Alt}$ falls in Q1.

Contributions

BD and VA conceived the study. BD, OBVP, SRO, QT, and VA discussed the experimental designs and results. BD performed the analyses. BD and VA wrote the manuscript in consultation with OBVP, SRO, and QT.

Competing interests

The authors declare no competing interests.

Acknowledgements

Behrooz Darbani acknowledges Louis-Hansens Fond (grant nr. 23-2B-14148). Vibeke Andersen acknowledges Sundhedsdonationer (grant nr. 2024-0379) and Region of Southern Denmark

(grant nr. 23/7359). The authors thank Caroline Moos (Research advisor, Department of Clinical research, Hospital Sønderjylland) for invaluable language-editing contributions.

References:

1. E. S. Lander, *et al.*, Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).

2. N. Patil, *et al.*, Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* **294**, 1719–1723 (2001).

3. †The International HapMap Consortium, The International HapMap Project. *Nature* (2003). Available at: https://www.nature.com/articles/nature02168 [Accessed 12 September 2018].

4. D. A. Hinds, *et al.*, Whole-genome patterns of common DNA variation in three human populations. *Science* **307**, 1072–1079 (2005).

5. R. Tadros, *et al.*, Large-scale genome-wide association analyses identify novel genetic loci and mechanisms in hypertrophic cardiomyopathy. *Nat Genet* **57**, 530–538 (2025).

6. C. Tcheandjieu, *et al.*, Large-scale genome-wide association study of coronary artery disease in genetically diverse populations. *Nat Med* **28**, 1679–1692 (2022).

7. S. L. Zheng, *et al.*, Genome-wide association analysis provides insights into the molecular etiology of dilated cardiomyopathy. *Nat Genet* **56**, 2646–2658 (2024).

8. H. Hautakangas, *et al.*, Genome-wide analysis of 102,084 migraine cases identifies 123 risk loci and subtype-specific risk alleles. *Nat Genet* **54**, 152–160 (2022).

9. V. Lammi, *et al.*, Genome-wide association study of long COVID. *Nat Genet* **57**, 1402–1417 (2025).

10. D. P. Wightman, *et al.*, A genome-wide association study with 1,126,563 individuals identifies new risk loci for Alzheimer's disease. *Nat Genet* **53**, 1276–1282 (2021).

11. J. Ghouse, *et al.*, Genome-wide meta-analysis identifies 93 risk loci and enables risk prediction equivalent to monogenic forms of venous thromboembolism. *Nat Genet* **55**, 399–409 (2023).

12. N. A. Rosenberg, *et al.*, Genetic structure of human populations. *Science* **298**, 2381–2385 (2002).

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

14. A. Manichaikul, *et al.*, Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).

15. J. McClellan, M.-C. King, Genetic heterogeneity in human disease. *Cell* **141**, 210–217 (2010).

16. R. Thomson, R. McWhirter, Adjusting for Familial Relatedness in the Analysis of GWAS Data. *Methods Mol Biol* **1526**, 175–190 (2017).

17. L. Sun, A. Dimitromanolakis, W.-M. Chen, Identifying Cryptic Relationships. *Methods Mol Biol* **1666**, 45–60 (2017).

18. J. H. Sul, L. S. Martin, E. Eskin, Population structure in genetic studies: Confounding factors and mixed models. *PLoS Genet* **14**, e1007309 (2018).

19. V. Tam, *et al.*, Benefits and limitations of genome-wide association studies. *Nat Rev Genet* **20**, 467–484 (2019).

20. H. Xu, *et al.*, SPAGRM: effectively controlling for sample relatedness in large-scale genome-wide association studies of longitudinal traits. *Nat Commun* **16**, 1413 (2025).

21. A. Korte, *et al.*, A mixed-model approach for genome-wide association studies of correlated traits in structured populations. *Nat Genet* **44**, 1066–1071 (2012).

22. H. Chen, *et al.*, Control for Population Structure and Relatedness for Binary Traits in Genetic Association Studies via Logistic Mixed Models. *The American Journal of Human Genetics* **98**, 653–666 (2016).

23. L. Jiang, Z. Zheng, H. Fang, J. Yang, A generalized linear mixed model association tool for biobank-scale data. *Nat Genet* **53**, 1616–1621 (2021).

24. W. Zhou, *et al.*, Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat Genet* **50**, 1335–1341 (2018).

25. P.-R. Loh, *et al.*, Efficient Bayesian mixed model analysis increases association power in large cohorts. *Nat Genet* **47**, 284–290 (2015).

26. X. Zhou, M. Stephens, Genome-wide efficient mixed-model analysis for association studies. *Nat Genet* **44**, 821–824 (2012).

27. H. M. Kang, *et al.*, Variance component model to account for sample structure in genome-wide association studies. *Nat Genet* **42**, 348–354 (2010).

28. P. M. A. Sleiman, *et al.*, GWAS identifies four novel eosinophilic esophagitis loci. *Nat Commun* **5**, 5593 (2014).

29. J. Vijayakrishnan, *et al.*, A genome-wide association study identifies risk loci for childhood acute lymphoblastic leukemia at 10q26.13 and 12q23.1. *Leukemia* **31**, 573–579 (2017).

30. J. Vijayakrishnan, *et al.*, Genome-wide association study identifies susceptibility loci for B-cell childhood acute lymphoblastic leukemia. *Nat Commun* **9**, 1340 (2018).

31. S. E. Graham, *et al.*, The power of genetic diversity in genome-wide association studies of lipids. *Nature* **600**, 675–679 (2021).

32. A. J. Forstner, *et al.*, Genome-wide association study of panic disorder reveals genetic overlap with neuroticism and depression. *Mol Psychiatry* **26**, 4179–4190 (2021).

33. B. R. Gorman, *et al.*, Multi-ancestry GWAS meta-analyses of lung cancer reveal susceptibility loci and elucidate smoking-independent genetic risk. *Nat Commun* **15**, 8629 (2024).

34. S. van Alten, B. W. Domingue, J. Faul, T. Galama, A. T. Marees, Correcting for volunteer bias in GWAS increases SNP effect sizes and heritability estimates. *Nat Commun* **16**, 3578 (2025).

35. K. S. O'Connell, *et al.*, Genomics yields biological and phenotypic insights into bipolar disorder. *Nature* **639**, 968–975 (2025).

36. H. Zhao, Family-based association studies. *Stat Methods Med Res* **9**, 563–587 (2000).

37. J. Lasky-Su, *et al.*, On Genome-wide Association Studies for Family-Based Designs: An Integrative Analysis Approach Combining Ascertained Family Samples with Unselected Controls. *Am J Hum Genet* **86**, 573–580 (2010).

38. J. Ott, Y. Kamatani, M. Lathrop, Family-based designs for genome-wide association studies. *Nat Rev Genet* **12**, 465–474 (2011).

39. E. M. Wijsman, Family-based approaches: design, imputation, analysis, and beyond. *BMC Genetics* **17**, S9 (2016).

40. E. M. Pedersen, *et al.*, Accounting for age of onset and family history improves power in genome-wide association studies. *Am J Hum Genet* **109**, 417–432 (2022).

41. C. Veller, M. Przeworski, G. Coop, Causal interpretations of family GWAS in the presence of heterogeneous effects. *Proceedings of the National Academy of Sciences* **121**, e2401379121 (2024).

42. L. Zhang, D. Conley, Within-Family GWAS does not Ameliorate the Decline in Prediction Accuracy across Populations. *bioRxiv* 2024.12.12.628188 (2024). https://doi.org/10.1101/2024.12.12.628188.

43. S. Song, S. Benonisdottir, J. S. Liu, A. Kong, Participation bias in the estimation of heritability and genetic correlation. *Proceedings of the National Academy of Sciences* **122**, e2425530122 (2025).

44. J. Z. Liu, Y. Erlich, J. K. Pickrell, Case–control association mapping by proxy using family history of disease. *Nat Genet* **49**, 325–331 (2017).

45. Y. Wu, *et al.*, Pervasive biases in proxy genome-wide association studies based on parental history of Alzheimer's disease. *Nat Genet* **56**, 2696–2703 (2024).

46. L. Wienbrandt, C. Prieß, D. Ellinghaus, GWAIS-Web: A Fast and Secure Web Service for Epistasis Detection in Genome-wide Association Interaction Studies. *bioRxiv* 2023.05.16.540964 (2023). https://doi.org/10.1101/2023.05.16.540964.

47. C. Carré, *et al.*, Next-Gen GWAS: full 2D epistatic interaction maps retrieve part of missing heritability and improve phenotypic prediction. *Genome Biology* **25**, 76 (2024).

48. L. Cui, *et al.*, Dominance is common in mammals and is associated with trans-acting gene expression and alternative splicing. *Genome Biol* **24**, 215 (2023).

49. R. Ohta, Y. Tanigawa, Y. Suzuki, M. Kellis, S. Morishita, A polygenic score method boosted by non-additive models. *Nat Commun* **15**, 4433 (2024).

50. O. Zuk, E. Hechter, S. R. Sunyaev, E. S. Lander, The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences* **109**, 1193–1198 (2012).

51. L. Varona, A. Legarra, M. A. Toro, Z. G. Vitezica, Non-additive Effects in Genomic Selection. *Frontiers in Genetics* **9** (2018).

52. D. S. Palmer, *et al.*, Analysis of genetic dominance in the UK Biobank. *Science* **379**, 1341–1348 (2023).

53. B. Darbani, M. Nicolaisen, On the Genetic Origins of Phenotypes in Genome-Wide Association Studies: The SNP Allocation For Estimating Heritability (SAFE-h2) Tool to Explore Additive-only Allelic Effects or Additive and Non-Additive Allelic Effects. *bioRxiv* 2023.08.28.555092 (2025). https://doi.org/10.1101/2023.08.28.555092.

54. B. Darbani, *et al.*, Cohort profile: Copenhagen Hospital Biobank-chronic inflammatory disease—inflammatory bowel disease (CHB-CID:IBD) genetic cohort. *Eur J Epidemiol* **40**, 721–734 (2025).

55. G. Xing, C. Xing, Adjusting for covariates in logistic regression models. *Genet Epidemiol* **34**, 769–771; author reply 772 (2010).

56. J. Mefford, J. S. Witte, The Covariate's Dilemma. *PLoS Genet* **8**, e1003096 (2012).

57. C.-L. Kuo, E. Feingold, What's the best statistic for a simple test of genetic association in a case-control study? *Genet Epidemiol* **34**, 246–253 (2010).

58. A. Mishra, *et al.*, Stroke genetics informs drug discovery and risk prediction across ancestries. *Nature* **611**, 115–123 (2022).

59. The top 10 causes of death. Available at: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death [Accessed 2 January 2026].

60. D. J. Schaid, W. Chen, N. B. Larson, From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat Rev Genet* **19**, 491–504 (2018).

61. M. van de Bunt, A. Cortes, M. A. Brown, A. P. Morris, M. I. McCarthy, Evaluating the Performance of Fine-Mapping Strategies at Common Variant GWAS Loci. *PLoS Genet* **11**, e1005535 (2015).

62. J. B. Maller, *et al.*, Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nat Genet* **44**, 1294–1301 (2012).

63. A. Hutchinson, J. Asimit, C. Wallace, Fine-mapping genetic associations. *Hum Mol Genet* **29**, R81–R88 (2020).

64. G. Wang, A. Sarkar, P. Carbonetto, M. Stephens, A simple new approach to variable selection in regression, with application to genetic fine mapping. *J R Stat Soc Series B Stat Methodol* **82**, 1273–1300 (2020).

65. FINEMAP: efficient variable selection using summary data from genome-wide association studies | Bioinformatics | Oxford Academic. Available at: https://academic.oup.com/bioinformatics/article/32/10/1493/1743040 [Accessed 4 November 2025].

66. J. R. Staley, *et al.*, A comparison of Cox and logistic regression for use in genome-wide association studies of cohort and case-cohort design. *Eur J Hum Genet* **25**, 854–862 (2017).

67. R. Dey, *et al.*, Efficient and accurate frailty model approach for genome-wide survival association analysis in large-scale biobanks. *Nat Commun* **13**, 5437 (2022).

68. D. Speed, G. Hemani, M. R. Johnson, D. J. Balding, Improved Heritability Estimation from Genome-wide SNPs. *Am J Hum Genet* **91**, 1011–1021 (2012).

69. C. C. Chang, *et al.*, Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, s13742-015-0047–8 (2015).

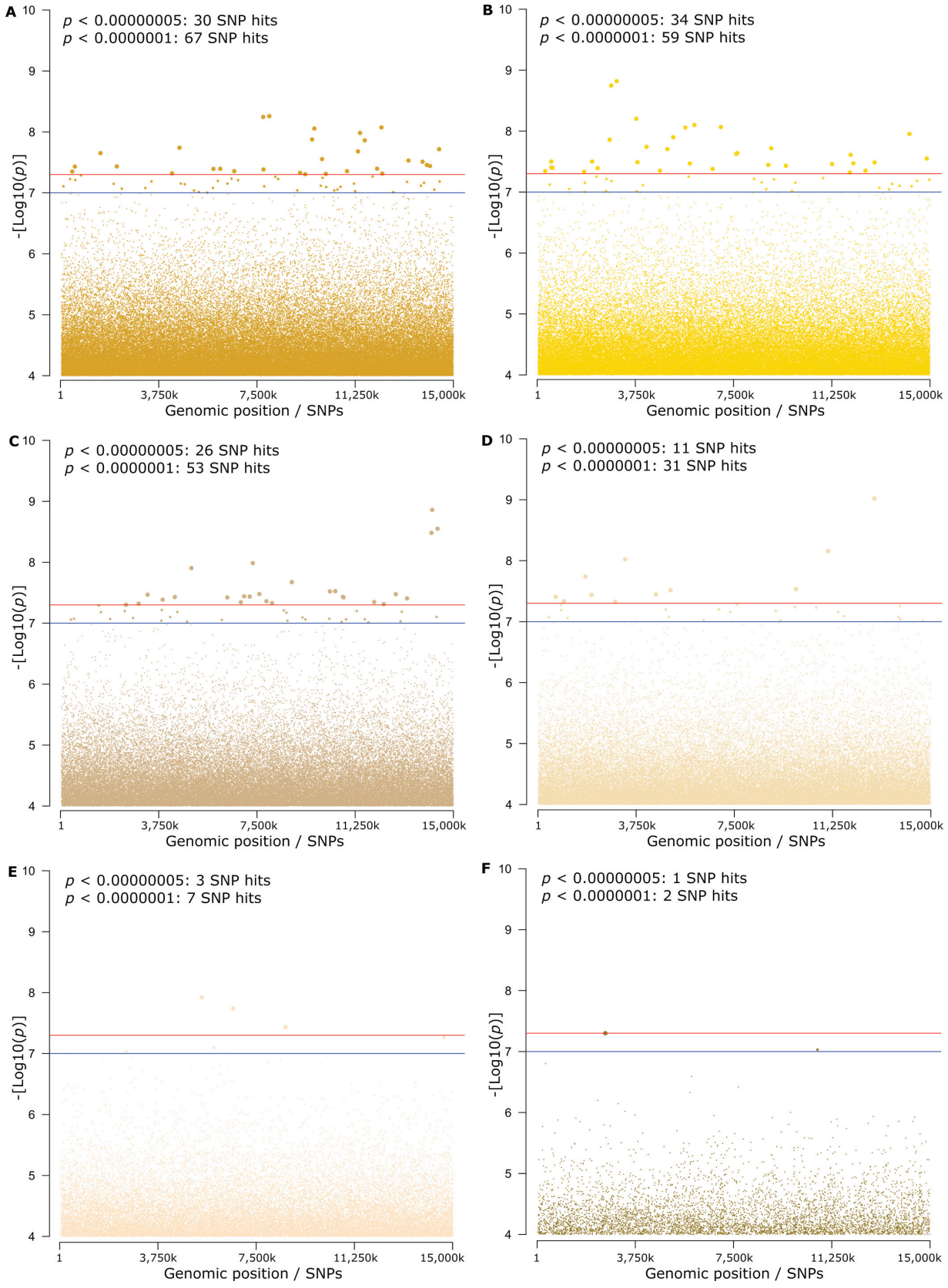
A
p < 0.00000005: 30 SNP hits
p < 0.0000001: 67 SNP hits
B
p < 0.00000005: 34 SNP hits
p < 0.0000001: 59 SNP hits
C
p < 0.00000005: 26 SNP hits
p < 0.0000001: 53 SNP hits
D
p < 0.00000005: 11 SNP hits
p < 0.0000001: 31 SNP hits
E
p < 0.00000005: 3 SNP hits
p < 0.0000001: 7 SNP hits
F
p < 0.00000005: 1 SNP hits
p < 0.0000001: 2 SNP hits
-[Log10(p)]
10
9
8
7
6
5
4
1
3,750k
7,500k
11,250k
15,000k
Genomic position / SNPs

Figure 1

Figure 1. Genetic relatedness across case-control samples provides higher statistical power, leading to reduced random associations, i.e., false positive rates. Each Manhattan plot is illustrating the outputs from 60 independent replicates of case-control association analyses. The analyses were based on simulated data including 15 million SNPs (MAF ≥ 0.01) for 10,000 genetically unrelated case samples (genetic similarity ≤ 0.3%) and 10,000 genetically unrelated controls samples (genetic similarity ≤ 0.3%). The decreased number of random association signals in these datasets reflect reduced false positive rates resulting from random allocation of case and control samples. A) Genetic similarities between case and control samples were ≤ 0.3%, except for 100 control samples each sharing between 24.8%–25.1% genetic similarity with an individual case samples. B) Genetic similarities between case and control samples were ≤ 0.3%, except for 1,000 control samples each sharing between 24.8%–25.1% genetic similarity with an individual case samples. C) Genetic similarities between case and control samples were ≤ 0.3%, except for 2,000 control samples each sharing between 24.8%–25.1% genetic similarity with an individual case samples. D) Genetic similarities between case and control samples were ≤ 0.3%, except for 2,500 control samples each sharing between 24.8%–25.1% genetic similarity with an individual case samples. E) Genetic similarities between case and control samples were ≤ 0.3%, except for 5,000 control samples each sharing 24.8%–25.1% genetic similarity with an individual case samples. F) Genetic similarities between case and control samples were 24.8%–25.1%. MAF: Minor allele frequency.

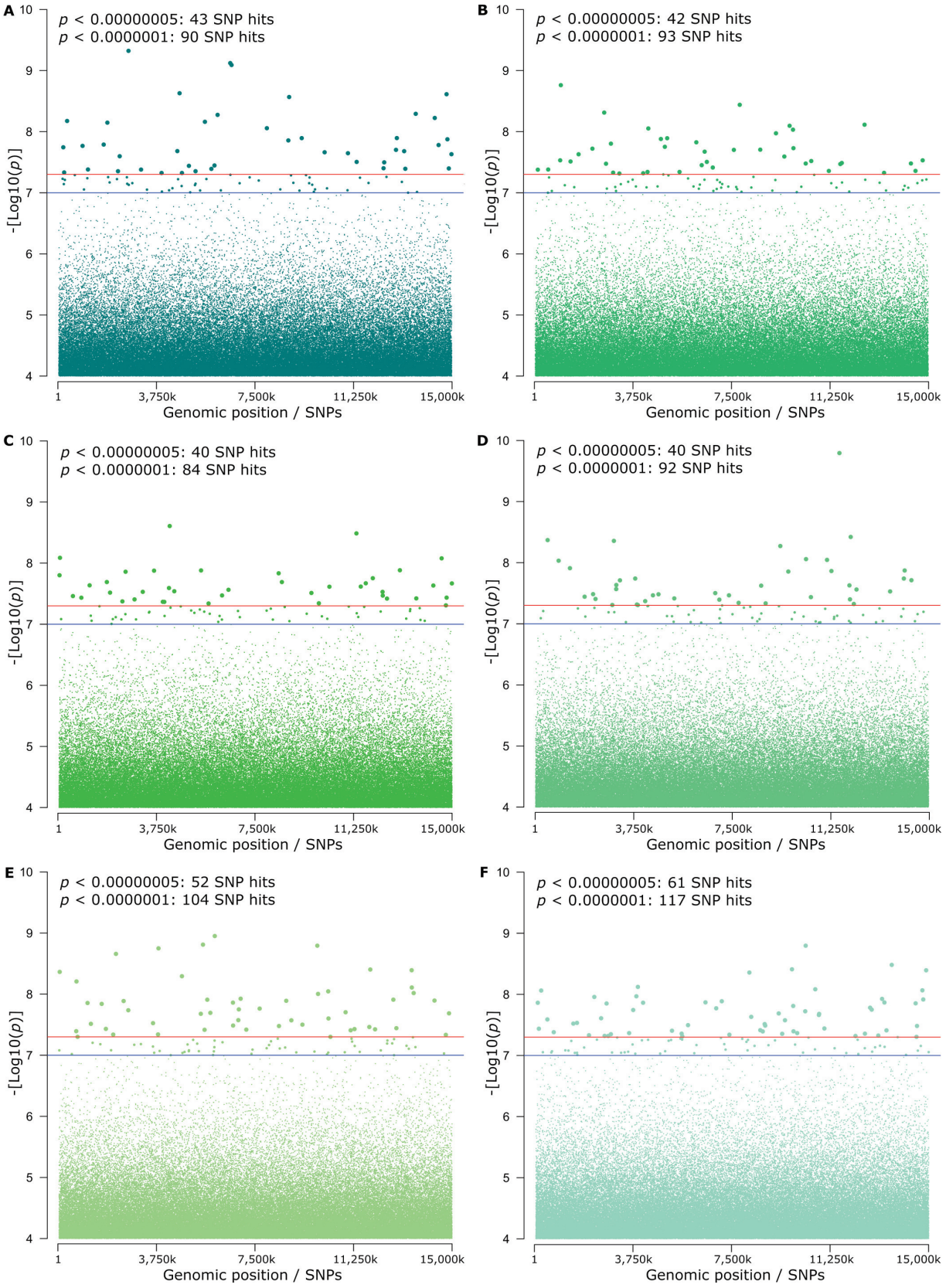
A
p < 0.00000005: 43 SNP hits
p < 0.0000001: 90 SNP hits
B
p < 0.00000005: 42 SNP hits
p < 0.0000001: 93 SNP hits
C
p < 0.00000005: 40 SNP hits
p < 0.0000001: 84 SNP hits
D
p < 0.00000005: 40 SNP hits
p < 0.0000001: 92 SNP hits
E
p < 0.00000005: 52 SNP hits
p < 0.0000001: 104 SNP hits
F
p < 0.00000005: 61 SNP hits
p < 0.0000001: 117 SNP hits
-[Log10(p)]
Genomic position / SNPs
1
3,750k
7,500k
11,250k
15,000k
4
5
6
7
8
9
10

Figure 2.

Figure 2. Random associations across varying numbers of genetically related, case- or control-specific paired samples. Each Manhattan plot is illustrating the outputs from 60 independent replicates of case-control association analyses. The analyses were based on simulated data including 15 million SNPs (MAF ≥ 0.01) and 20,000 samples; 10,000 case samples which were genetically unrelated, i.e., genetic similarity ≤ 0.3%, to 10,000 controls samples. The elevated numbers of random association signals in these datasets reflect increased false positive rates resulting from random allocation of case and control samples. A) Analyzing genetically unrelated samples. Genetic similarities among cases, controls, and case-control samples were ≤ 0.3%. B) Analyzing genetically unrelated samples. Genetic similarities among cases, controls, and case-control samples were ≤ 0.3%, except 50 paired samples from the control group exhibiting 24.8%–25.1% genetic similarity between paired samples. C) Analyzing genetically unrelated samples. Genetic similarities among cases, controls, and case-control samples were ≤ 0.3%, except 100 paired samples from the control group exhibiting 24.8%–25.1% genetic similarity between paired samples. D) Analyzing genetically unrelated samples. Genetic similarities among cases, controls, and case-control samples were ≤ 0.3%, except 250 paired samples from the control group exhibiting 24.8%–25.1% genetic similarity between paired samples. E) Analyzing genetically unrelated samples. Genetic similarities among cases, controls, and case-control samples were ≤ 0.3%, except 500 paired samples from the control group exhibiting 24.8%–25.1% genetic similarity between paired samples. F) Analyzing genetically unrelated samples. Genetic similarities among cases, controls, and case-control samples were ≤ 0.3%, except 1,000 paired samples from the control group exhibiting 24.8%–25.1% genetic similarity between paired samples. MAF: Minor allele frequency.

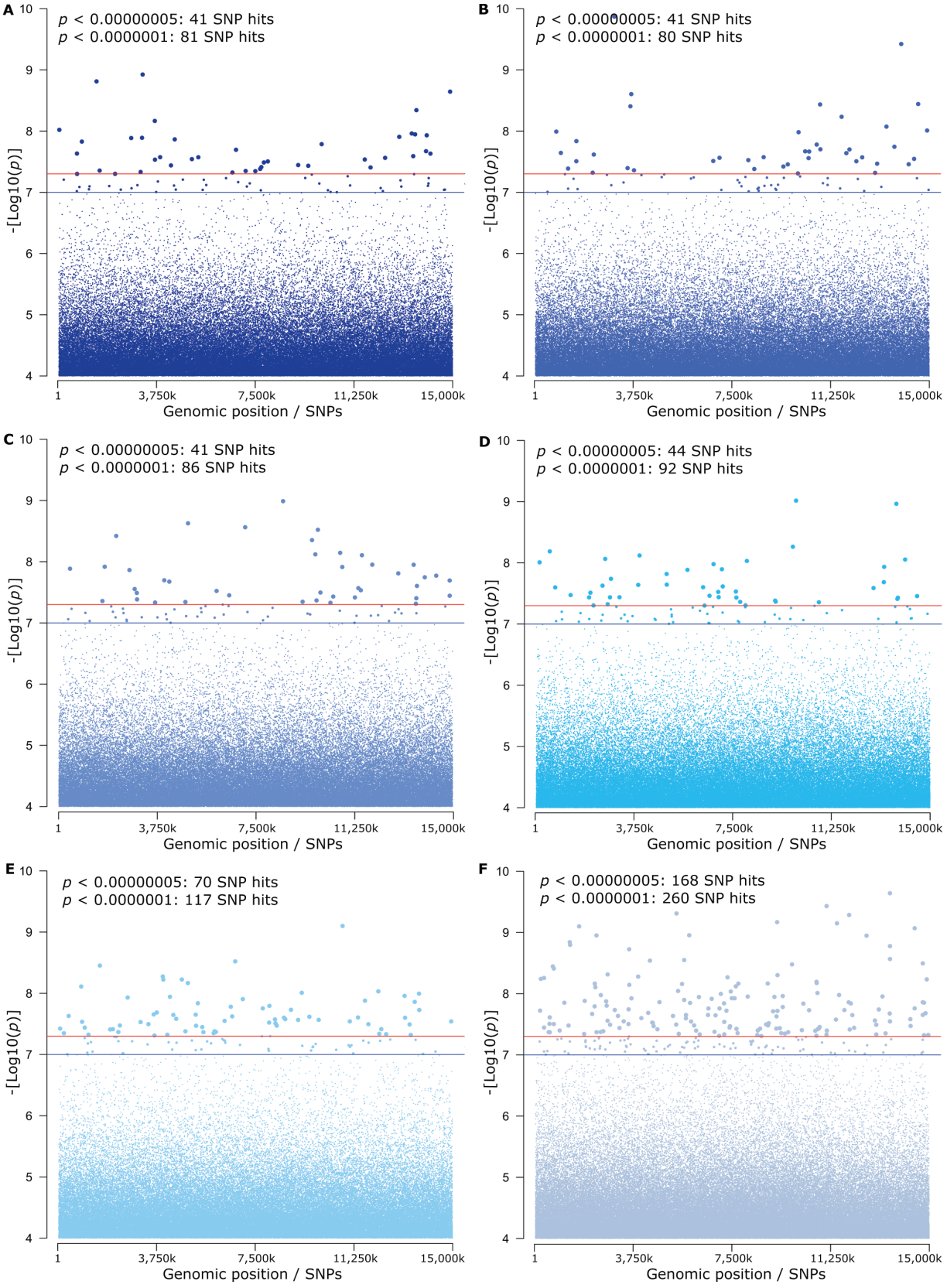
A
p < 0.00000005: 41 SNP hits
p < 0.0000001: 81 SNP hits
B
p < 0.00000005: 41 SNP hits
p < 0.0000001: 80 SNP hits
C
p < 0.00000005: 41 SNP hits
p < 0.0000001: 86 SNP hits
D
p < 0.00000005: 44 SNP hits
p < 0.0000001: 92 SNP hits
E
p < 0.00000005: 70 SNP hits
p < 0.0000001: 117 SNP hits
F
p < 0.00000005: 168 SNP hits
p < 0.0000001: 260 SNP hits
-[Log10(p)]
10
9
8
7
6
5
4
1
3,750k
7,500k
11,250k
15,000k
Genomic position / SNPs

Figure 3.

Figure 3. Random associations across varying sizes of genetically related case- or control-specific sub-communities. Each Manhattan plot is illustrating the outputs from 60 independent replicates of case-control association analyses. The analyses were based on simulated data including 15 million SNPs (MAF ≥ 0.01) and 20,000 samples; 10,000 case samples which were genetically unrelated, i.e., genetic similarity ≤ 0.3%, to 10,000 controls samples. The elevated numbers of random association signals in these datasets, relative to the dataset without relatedness (Figure 2A), reflect increased false positive rates resulting from random allocation of case and control samples. A) Analyzing genetically unrelated samples. Genetic similarities among cases, controls, and case-control samples were ≤ 0.3%, except for a sub-community of 10 control samples, each sharing 24.8% to 25.1% genetic similarity with all other samples within the sub-community. B) Analyzing genetically unrelated samples. Genetic similarities among cases, controls, and case-control samples were ≤ 0.3%, except for a sub-community of 15 control samples, each sharing 24.8% to 25.1% genetic similarity with all other samples within the sub-community. C) Analyzing genetically unrelated samples. Genetic similarities among cases, controls, and case-control samples were ≤ 0.3%, except for a sub-community of 20 control samples, each sharing 24.8% to 25.1% genetic similarity with all other samples within the sub-community. D) Analyzing genetically unrelated samples. Genetic similarities among cases, controls, and case-control samples were ≤ 0.3%, except for a sub-community of 25 control samples, each sharing 24.8% to 25.1% genetic similarity with all other samples within the sub-community. E) Analyzing genetically unrelated samples. Genetic similarities among cases, controls, and case-control samples were ≤ 0.3%, except for a sub-community of 35 control samples, each sharing 24.8% to 25.1% genetic similarity with all other samples within the sub-community. F) Analyzing genetically unrelated samples. Genetic similarities among cases, controls, and case-control samples were ≤ 0.3%, except for a sub-community of 50 control samples, each sharing 24.8% to 25.1% genetic similarity with all other samples within the sub-community. MAF: Minor allele frequency.

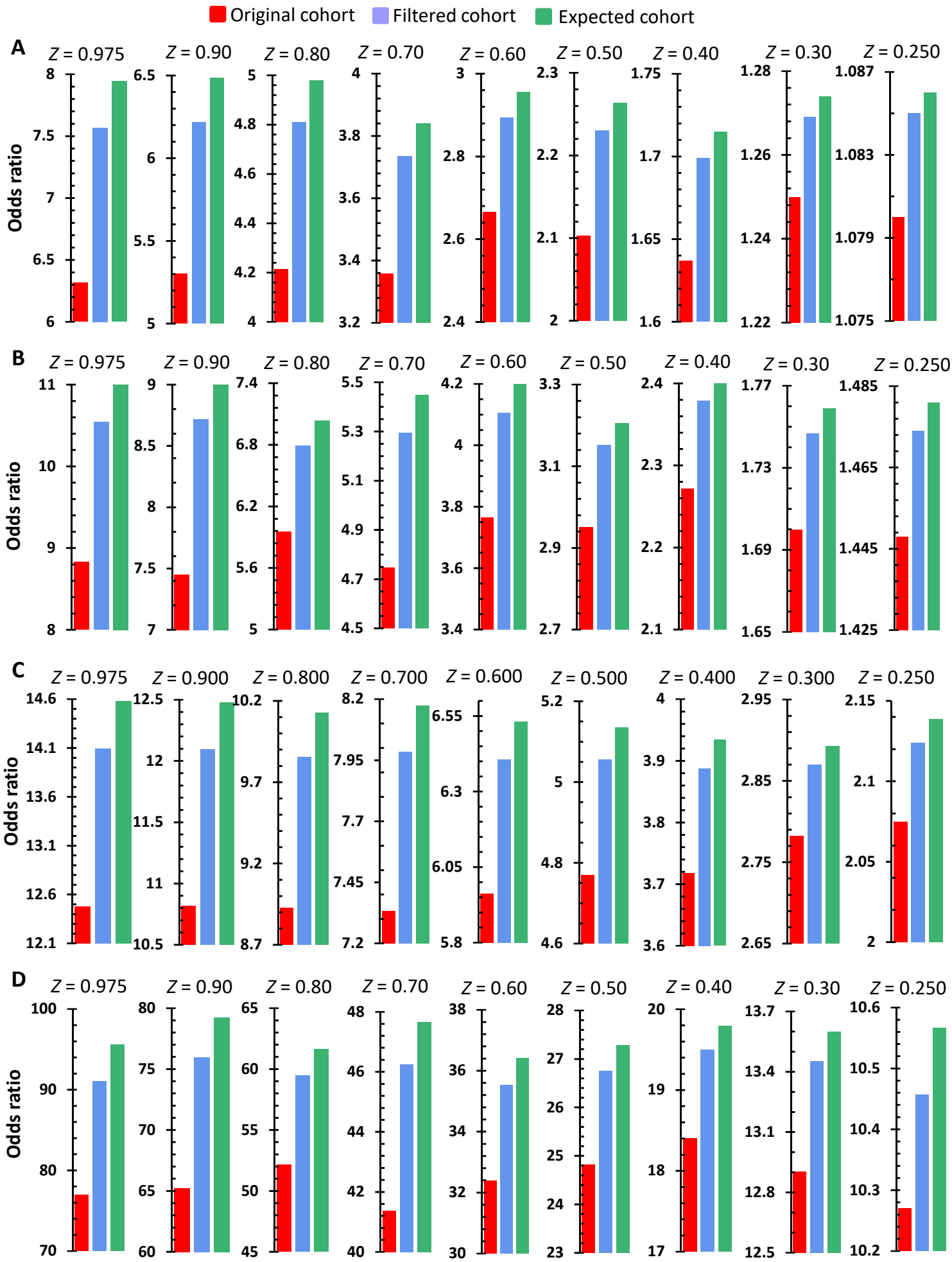
Original cohort
Filtered cohort
Expected cohort
A
B
C
D
Odds ratio
Z = 0.975
Z = 0.90
Z = 0.80
Z = 0.70
Z = 0.60
Z = 0.50
Z = 0.40
Z = 0.30
Z = 0.250
Z = 0.900
Z = 0.800
Z = 0.700
Z = 0.600
Z = 0.500
Z = 0.400
Z = 0.300

Figure 4.

Figure 4. Predisposed genotypes within cohorts introduce downward biases in odds ratios. A) Disease prevalence of 15% and minor allele frequency of 22.5%. B) Disease prevalence of 10% and minor allele frequency of 15%. C) Disease prevalence of 5% and minor allele frequency of 7.5%. D) Disease prevalence of 1% and minor allele frequency of 1.5%. A-D) Genotype simulations were performed as discussed in Methods and Text-Box S1. Based on the global age-distribution and an assumed even incidence rate of 12.5% of disease prevalence for each 10-year age interval, we excluded 90% of controls to simulate filtering of individuals under 65 years of age. Varying levels of association were simulated by applying different levels of Z. For example, Z values of 0.9 and 0.5 indicate that only 90% and 50% of the cases, respectively, are attributable to the causal variant. The control groups within the original cohorts had impurity at half rate of the disease prevalence rates; predisposed individuals were simulated at a half rate of the population disease prevalence as expected and illustrated in Table 1. The control groups within the filtered cohorts had residual impurity at rates of 0.094 × disease-prevalence, as explained in Table 1.

Tables:

Table 1. Simulation of predisposed individuals within control groups of cohorts under the assumptions of similar age distribution between the source populations and cohorts and, an even, i.e., age-independent, incidence rate of 12.5% of disease prevalence rate for each 10-year age interval.

Age-groups	Proportion of population[1]	Incident rate (as proportion of DP[2])	Rate of pre-symptomatic individuals / latent cases (as proportion of DP[2])	Population-proportion weighted rate of pre-symptomatic individuals / latent cases (as proportion of DP[2])
0-9	_	0.1250	0.8750	_
10-19	_	0.2500	0.7500	_
0-19	0.34	0.2500	0.8125[3]	0.8125 × 0.34 = 0.276
20-29	_	0.3750	0.6250	_
30-39	_	0.5000	0.5000	_
40-49	_	0.6250	0.3750	_
50-59	_	0.7500	0.2500	_
60-64	_	0.8125	0.1875	_
20-64	0.56	_	0.3875[4]	0.56 × 0.3875 = 0.217
65-…	0.11	1.0000	0.1875 × 0.5 = 0.094[5]	0.11 × 0.094 = 0.01
All	1	_	_	0.276 + 0.217 + 0.01 = 0.503

[1]Data is obtained from UN, Department of Economic and Social Affairs at https://population.un.org/wpp/graphs?loc=900&type=Probabilistic%20Projections&category=Population&subcategory=Age%2080%20and%20over and from Our World in Data at https://ourworldindata.org/grapher/population-by-age-group?tab=table.

[2]Disease prevalence

[3]The average of 10-year age intervals of 0-9 and 10-19.

[4]The average of 10-year age intervals of 20-29, 30-39, 40-49, 50-59, and 60-64.

[5]We assumed a 50% rate of pre-symptomatic individuals / latent cases for the population aged 65 and up.

Table 2. Additive-model based association analyses tend to produce inaccurate risk score estimates under conditions of non-additive allelic interactions.

DP[1]	SNPs	True effect[1]	True risk score			Additive-model based risk score			Logistic-regression / Association *p*-values from additive and non-additive analysis models[1]				Genotype-specific case/control ratios [Genotype (case/control ratio)]		
			RR[2]	RA[2]	AA[2]	RR[2]	RA[2]	AA[2]	Add	Dom	OD	Het	Q1	Q2-3	Q4
1%	**SNP1**	Add	0β =0	1β = 2.502	2β = 5.04	0β =0	1β = 2.502	2β = 5.04	1.0×10^{-2029}	1.4×10^{-1356}	3.9×10^{-657}	0.998	RR (0.08)	RA (1.09)	AA (2.21)
	SNP2	Add	0β =0	1β = 2.189	2β = 4.378	0β =0	1β = 2.189	2β = 4.378	3.8×10^{-1576}	2.9×10^{-1098}	3.4×10^{-496}	0.995	RR (0.09)	RA (0.82)	AA (1.61)
	SNP3	Add	0β =0	1β = 1.774	2β = 3.548	0β =0	1β = 1.774	2β = 3.548	2.1×10^{-1053}	1.8×10^{-767}	1.1×10^{-318}	1	RR (0.09)	RA (0.56)	AA (1.05)
	SNP4	Add	0β =0	1β = 1.129	2β = 2.258	0β =0	1β = 1.129	2β = 2.258	3.3×10^{-443}	2.3×10^{-337}	1.2×10^{-125}	1	RR (0.10)	RA (0.30)	AA (0.51)
	SNP5	Add	0β =0	1β = 0.656	2β = 1.312	0β =0	1β = 0.656	2β = 1.312	2.7×10^{-157}	6.7×10^{-121}	4.8×10^{-43}	1	RR (0.1)	RA (0.19)	AA (0.28)
	SNP6	Dom	0β = 0	2β = 5.116	2β = 5.116	0β = 0	1β = 2.451	2β = 4.902	9.7×10^{-886}	1.2×10^{-978}	9.7×10^{-886}	1.3×10^{-252}	RR (0.08)		RA,AA (1.05)
	SNP7	Dom	0β = 0	2β = 4.394	2β = 4.394	0β = 0	1β = 2.122	2β = 4.244	2.3×10^{-645}	1.2×10^{-751}	2.3×10^{-645}	4.8×10^{-192}	RR (0.09)		RA,AA (0.78)
	SNP8	Dom	0β = 0	2β = 3.442	2β = 3.442	0β = 0	1β = 1.677	2β = 3.354	4.4×10^{-383}	7.2×10^{-473}	4.4×10^{-383}	9.7×10^{-120}	RR (0.09)		RA,AA (0.51)
	SNP9	Dom	0β = 0	2β = 1.921	2β = 1.921	0β = 0	1β = 0.944	2β = 1.889	2.8×10^{-110}	2.9×10^{-144}	2.8×10^{-110}	6.6×10^{-37}	RR (0.10)		RA,AA (0.25)
	SNP10	Dom	0β = 0	2β = 0.704	2β = 0.704	0β = 0	1β = 0.348	2β = 0.695	9.5×10^{-15}	3.5×10^{-19}	9.5×10^{-15}	8.6×10^{-6}	RR (0.10)		RA,AA (0.14)
	SNP11	OD	0β = 0	2β = 5.118	1β = 2.559	0β = 0	1β = 2.448	2β = 4.896	5.1×10^{-332}	1.9×10^{-775}	1.2×10^{-1091}	2.3×10^{-748}	RR (0.08)	AA (0.56)	RA (1.1)
	SNP12	OD	0β = 0	2β = 4.396	1β = 2.198	0β = 0	1β = 2.120	2β = 4.240	8.7×10^{-233}	2.4×10^{-579}	1.2×10^{-787}	1.3×10^{-551}	RR (0.09)	AA (0.43)	RA (0.78)
	SNP13	OD	0β = 0	2β = 3.444	1β = 1.722	0β = 0	1β = 1.675	2β = 3.350	8.4×10^{-131}	2.1×10^{-384}	5.2×10^{-460}	2.6×10^{-326}	RR (0.09)	AA (0.30)	RA (0.51)
	SNP14	OD	0β = 0	2β = 1.924	1β = 0.962	0β = 0	1β = 0.944	2β = 1.888	1.9×10^{-34}	8.1×10^{-97}	4.1×10^{-126}	8.2×10^{-93}	RR (0.10)	AA (0.18)	RA (0.25)
	SNP15	OD	0β = 0	2β = 0.708	1β = 0.354	0β = 0	1β = 0.349	2β = 0.697	5.5×10^{-5}	3.2×10^{-12}	1.1×10^{-15}	4.8×10^{-12}	RR (0.10)	AA (0.12)	RA (0.14)
	SNP16	Het	0β = 0	1β = 2.564	0β = 0	0β = 0	1β = 2.446	2β = 4.892	1	3.2×10^{-539}	5.9×10^{-1266}	2.8×10^{-1617}	RR,AA (0.08)		RA (1.06)
	SNP17	Het	0β = 0	1β = 2.202	0β = 0	0β = 0	1β = 2.118	2β = 4.236	1	2.3×10^{-385}	6.7×10^{-944}	9.4×10^{-1200}	RR,AA (0.09)		RA (0.78)
	SNP18	Het	0β = 0	1β = 1.726	0β = 0	0β = 0	1β = 1.673	2β = 3.346	1	9.3×10^{-210}	8.8×10^{-545}	8.9×10^{-692}	RR,AA (0.09)		RA (0.51)
	SNP19	Het	0β = 0	1β = 0.966	0β = 0	0β = 0	1β = 0.943	2β = 1.886	1	5.0×10^{-53}	1.2×10^{-146}	5.7×10^{-189}	RR,AA (0.10)		RA (0.25)
	SNP20	Het	0β = 0	1β = 0.357	0β = 0	0β = 0	1β = 0.350	2β = 0.700	1	1.3×10^{-6}	8.0×10^{-17}	9.9×10^{-22}	RR,AA (0.10)		RA (0.14)
15%	**SNP1**	Add	0β =0	1β = 1.759	2β = 3.518	0β =0	1β = 1.759	2β = 3.518	1.3×10^{-1366}	2.5×10^{-723}	6.1×10^{-399}	1	RR (0.03)	RA (0.31)	AA (0.77)
	SNP2	Add	0β =0	1β = 1.637	2β = 3.274	0β =0	1β = 1.637	2β = 3.274	2.9×10^{-1215}	1.3×10^{-697}	4.3×10^{-350}	0.995	RR (0.03)	RA (0.29)	AA (0.69)
	SNP3	Add	0β =0	1β = 1.389	2β = 2.778	0β =0	1β = 1.389	2β = 2.778	8.7×10^{-921}	6.6×10^{-593}	2.7×10^{-258}	1	RR (0.04)	RA (0.25)	AA (0.54)
	SNP4	Add	0β =0	1β = 1.130	2β = 2.261	0β =0	1β = 1.130	2β = 2.261	2.6×10^{-641}	1.6×10^{-444}	2.6×10^{-175}	0.994	RR (0.06)	RA (0.21)	AA (0.42
	SNP5	Add	0β =0	1β = 0.707	2β = 1.414	0β =0	1β = 0.707	2β = 1.414	2.9×10^{-270}	2.0×10^{-199}	1.1×10^{-71}	0.993	RR (0.08)	RA (0.16)	AA (0.26)
	SNP6	Dom	0β = 0	2β = 4.549	2β = 4.549	0β = 0	1β = 1.541	2β = 3.081	7.6×10^{-496}	3.1×10^{-505}	7.6×10^{-496}	1.3×10^{-252}	RR (0.08)		RA,AA (0.32)
	SNP7	Dom	0β = 0	2β = 3.998	2β = 3.998	0β = 0	1β = 1.406	2β = 2.812	4.4×10^{-418}	4.8×10^{-457}	4.4×10^{-418}	1.5×10^{-145}	RR (0.04)		RA,AA (0.29)
	SNP8	Dom	0β = 0	2β = 3.018	2β = 3.018	0β = 0	1β = 1.127	2β = 2.254	1.1×10^{-273}	8.1×10^{-329}	1.1×10^{-273}	1.2×10^{-94}	RR (0.05)		RA,AA (0.24)
	SNP9	Dom	0β = 0	2β = 2.114	2β = 2.114	0β = 0	1β = 0.827	2β = 1.654	6.1×10^{-149}	2.5×10^{-189}	6.1×10^{-149}	1.6×10^{-51}	RR (0.07)		RA,AA (0.19)
	SNP10	Dom	0β = 0	2β = 0.740	2β = 0.740	0β = 0	1β = 0.307	2β = 0.613	2.7×10^{-21}	9.9×10^{-28}	2.7×10^{-21}	4.4×10^{-8}	RR (0.09)		RA,AA (0.13)
	SNP11	OD	0β = 0	2β = 4.527	1β = 2.264	0β = 0	1β = 1.487	2β = 2.975	1.2×10^{-173}	2.0×10^{-379}	1.4×10^{-561}	2.6×10^{-421}	RR (0.03)	AA (0.16)	RA (0.34)
	SNP12	OD	0β = 0	2β = 4.004	1β = 2.002	0β = 0	1β = 1.360	2β = 2.721	1.2×10^{-142}	1.4×10^{-350}	1.1×10^{-517}	8.8×10^{-400}	RR (0.04)	AA (0.15)	RA (0.31)
	SNP13	OD	0β = 0	2β = 3.062	1β = 1.531	0β = 0	1β = 1.097	2β = 2.195	5.4×10^{-92}	6.6×10^{-246}	7.6×10^{-342}	1.5×10^{-260}	RR (0.05)	AA (0.14)	RA (0.25)
	SNP14	OD	0β = 0	2β = 2.184	1β = 1.092	0β = 0	1β = 0.815	2β = 1.631	5.8×10^{-50}	1.3×10^{-139}	3.3×10^{-188}	2.3×10^{-142}	RR (0.07)	AA (0.13)	RA (0.20)
	SNP15	OD	0β = 0	2β = 0.842	1β = 0.421	0β = 0	1β = 0.329	2β = 0.659	1.0×10^{-8}	5.1×10^{-23}	4.1×10^{-30}	5.9×10^{-23}	RR (0.09)	AA (0.11)	RA (0.14)
	SNP16	Het	0β = 0	1β = 2.416	0β = 0	0β = 0	1β = 1.427	2β = 2.854	1	1.0×10^{-266}	9.4×10^{-726}	4.9×10^{-919}	RR,AA (0.03)		RA (0.35)
	SNP17	Het	0β = 0	1β = 2.138	0β = 0	0β = 0	1β = 1.308	2β = 2.616	1	3.4×10^{-231}	8.7×10^{-637}	3.0×10^{-812}	RR,AA (0.04)		RA (0.32)
	SNP18	Het	0β = 0	1β = 1.645	0β = 0	0β = 0	1β = 1.061	2β = 2.122	1	1.1×10^{-154}	5.1×10^{-436}	1.0×10^{-561}	RR,AA (0.05)		RA (0.27)
	SNP19	Het	0β = 0	1β = 1.189	0β = 0	0β = 0	1β = 0.797	2β = 1.594	1	5.8×10^{-85}	2.5×10^{-243}	2.2×10^{-316}	RR,AA (0.07)		RA (0.21)
	SNP20	Het	0β = 0	1β = 0.498	0β = 0	0β = 0	1β = 0.346	2β = 0.691	1	3.0×10^{-15}	3.0×10^{-42}	2.2×10^{-55}	RR,AA (0.09)		RA (0.14)

[1]DP: Disease prevalence. Add: Additive allelic effect; the causal allele in the homozygous state triggers the disease at twice the rate compared to its heterozygous state. Dom: Dominance allelic effect; the causal allele triggers disease at the same rate regardless of its dosage. OD: Overdominance allelic effect; the causal allele in the heterozygous state triggers the disease at twice the rate compared to its homozygous state. Het: Heterosis-like allelic effect; the causal allele in the heterozygous state triggers the disease but not at homozygous state.

[2]RR: Reference_Reference genotype, RA: Reference_Alternative genotype, AA: Alternative_Alternative genotype

Analyses were based on a simulated cohorts (n = 110,000, controls = 100,000, cases = 10,000, disease prevalence = 1% or 15%).

Prior to logistic regression, the genetic data were normalized to account for differences in within-population genotype frequencies. The identified true genetic effects are highlighted by the lowest *p*-values, shown in black.

β was calculated as the natural logarithm of the odds ratio, using odds ratios derived from the original (non-normalized) data.

Case-to-control ratio for every genotype (RR, RA, AA) was calculated and quartiles (Q1–Q4) were delineated according to the range of case-to-control ratios, i.e., from minimum to maximum, and the distribution of genotype specific ratios within these quartiles was used as an alternative approach to assess allelic effects.

Supplementary information:

Refining bias correction in genome-wide association analyses of case-control studies

Behrooz Darbani[1,2,3]*, Ole Birger Vesterager Pedersen[4,5], Sisse Rye Ostrowski[5,6], Qihua Tan[7,8], Vibeke Andersen[3,9]*

[1]Department of Clinical Research, Hospital Sønderjylland, Aabenraa, Denmark
[2]Department of Regional Health Research, University of Southern Denmark, Aabenraa, Denmark
[3]OPEN, Odense University Hospital, Odense, Denmark
[4]Department of Clinical Immunology, Zealand University Hospital, Køge, Denmark
[5]Department of Clinical Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark
[6]Department of Clinical Immunology, Copenhagen University Hospital – Rigshospitalet, Copenhagen, Denmark
[7]Research Unit for Epidemiology, Biostatistics and Biodemography, Department of Public Health, University of Southern Denmark, Odense, Denmark
[8]Unit of Human Genetics, Department of Clinical Research, University of Southern Denmark, Odense, Denmark
[9]Department of Clinical Research, University of Southern Denmark, Odense, Denmark

*Corresponding authors. BD: Department of Clinical Research, Hospital Sønderjylland, Department of Regional Health Research, University of Southern Denmark (Email: bds@rsyd.dk / behroozdarbani@gmail.com), VA: OPEN, Odense University Hospital, Department of Clinical Research, University of Southern Denmark (Email: va@rsyd.dk)

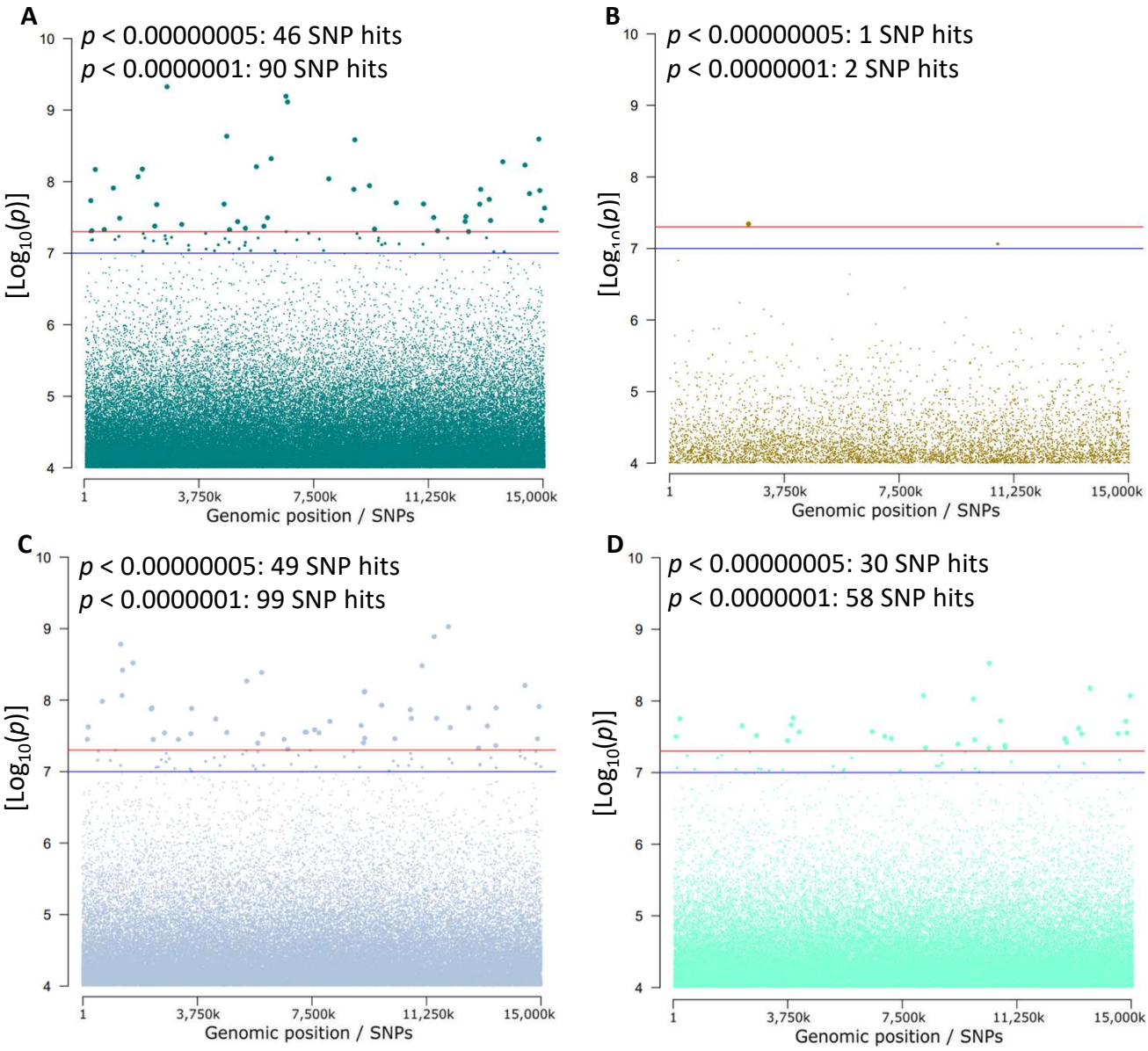

Figure S1. Random associations across varying numbers of genetically related, case- or control-group specific paired samples as well as relatedness between case and control groups. The generalized linear mixed model of fastGWA-GLMM was applied for association analyses. Each Manhattan plot is illustrating the outputs from 60 independent replicates of case-control association analyses. The analyses were based on simulated data including 15 million SNPs (MAF ≥ 0.01) and 20,000 samples. Random association signals in these datasets reflect false positive rates resulting from random allocation of case and control samples. Moreover, a smaller number of random associations observed in scenarios with case- or control-specific relatedness (D)—compared to scenarios without such relatedness and also without relatedness shared across case and control groups (A)—indicates an increased risk of false negatives. A) Analyzing genetically unrelated samples. Genetic similarities among cases, controls, and case-control samples were ≤ 0.3%. B) Analyzing genetically related samples. Genetic similarities between case and control samples were 24.8%–25.1%. Genetic similarities among cases and among controls were ≤ 0.3%. C) Analyzing genetically unrelated samples. Genetic similarities among cases, controls, and case-control samples were ≤ 0.3%, except for a sub-community of 50 control samples, each sharing 24.8% to 25.1% genetic similarity with all other samples within the sub-community. D) Analyzing genetically unrelated samples. Genetic similarities among cases, controls, and case-control samples were ≤ 0.3%, except 1,000 paired samples from the control group exhibiting 24.8%–25.1% genetic similarity between paired samples. MAF: Minor allele frequency.

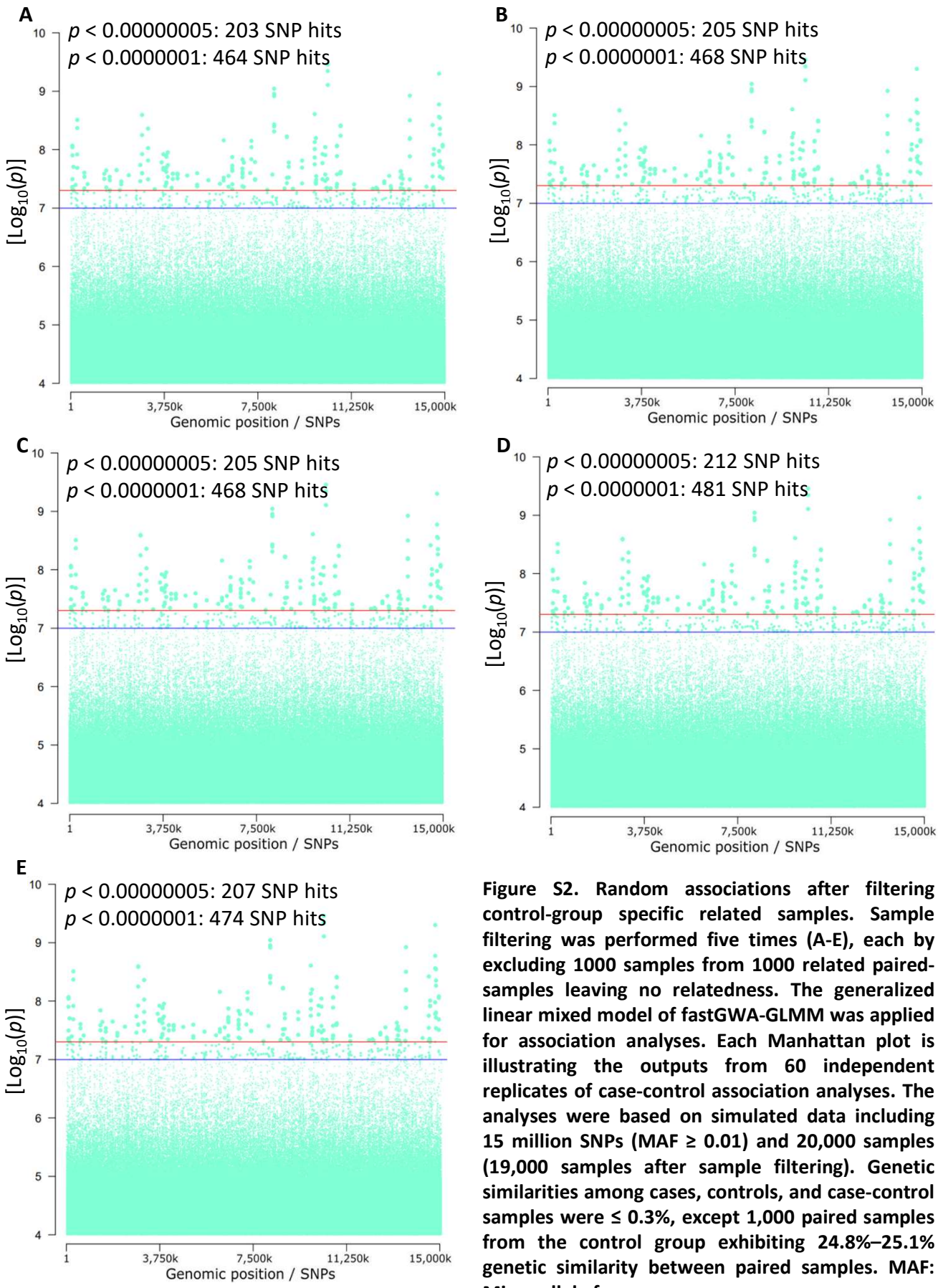

Figure S2. Random associations after filtering control-group specific related samples. Sample filtering was performed five times (A-E), each by excluding 1000 samples from 1000 related paired-samples leaving no relatedness. The generalized linear mixed model of fastGWA-GLMM was applied for association analyses. Each Manhattan plot is illustrating the outputs from 60 independent replicates of case-control association analyses. The analyses were based on simulated data including 15 million SNPs (MAF ≥ 0.01) and 20,000 samples (19,000 samples after sample filtering). Genetic similarities among cases, controls, and case-control samples were ≤ 0.3%, except 1,000 paired samples from the control group exhibiting 24.8%–25.1% genetic similarity between paired samples. MAF: Minor allele frequency.

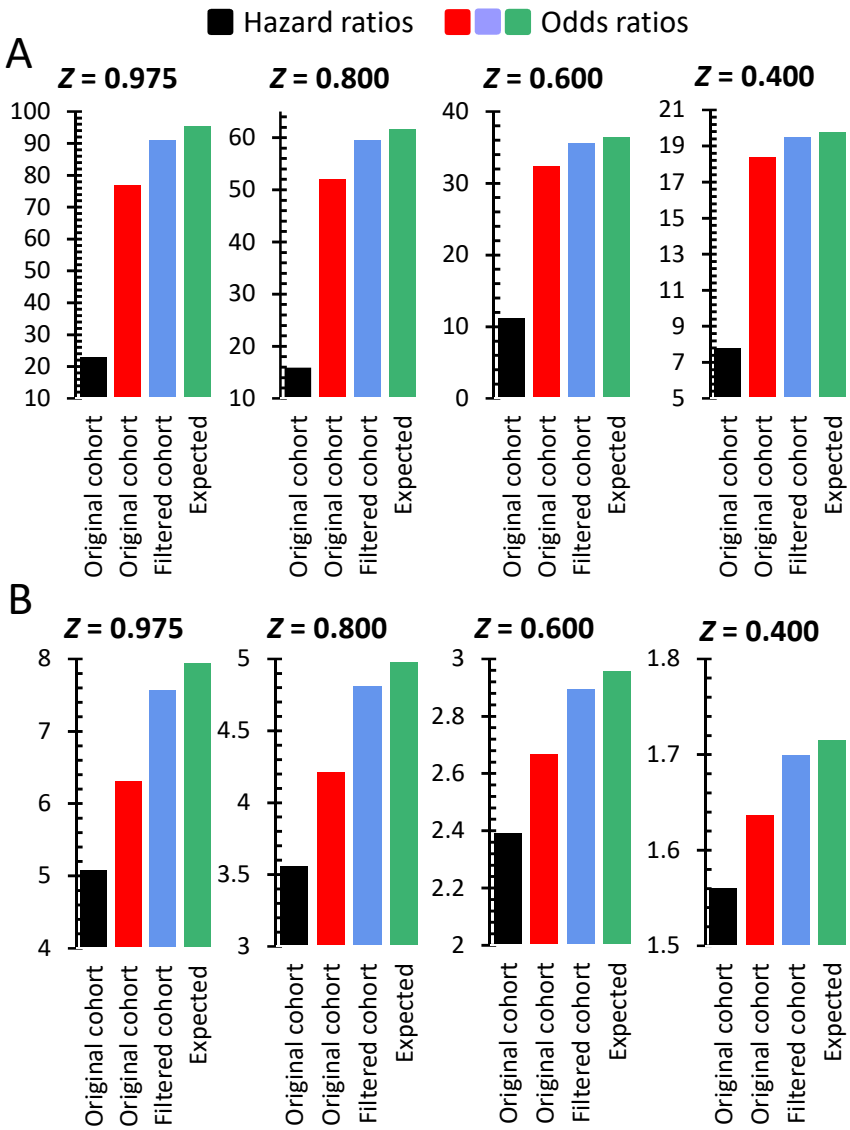

Figure S3. The presence of predisposed individuals within cohorts introduces downward bias in odds ratio estimates. A) Disease prevalence of 1% and minor allele frequency of 1.5%. B) Disease prevalence of 15% and minor allele frequency of 22.5%. A,B) Genotype simulations were performed as discussed in Methods and Text-Box S1. Based on the global average age-distribution and an even incidence rate of 12.5% of disease prevalence for each 10-year age interval, we excluded 90% of controls to simulate filtering of individuals under 65 years of age. Varying levels of association were simulated by applying different levels of Z. For example, Z values of 0.8 and 0.4 indicate that only 80% and 40% of the cases, respectively, are attributable to the causal variant. The control groups within the original cohorts had impurity at half rate of the disease prevalence rates; predisposed individuals were simulated at a half rate of the population disease prevalence as expected and illustrated in Table 1. The control groups within the filtered cohorts had residual impurity at rates of 0.094 × disease-prevalence, as explained in Table 1. Age was used as time-to-outcome in Cox regression.

Text-Box S1. Simulation of genotypes within population (XX_P), genotypes of healthy controls after excluding cases from the population (XX_{PH}), case genotypes within the cohort (XX_C), control genotypes within the cohort (XX_H), and control genotypes within the cohort after removing individuals under 65 years of age (XX_H). According to Table 1, individuals younger than 65 years accounted for 90% of the original cohort. Varying levels of association were simulated by adjusting the value of Z, which can range from 1 (representing the full rate of disease prevalence) to 0 (representing no contribution to disease prevalence). For example, Z values of 0.9 and 0.5 indicate that 90% and 50% of the cases, respectively, are attributable to the causal variant. Here, a simulation at 1% disease prevalence and 1.5% causal allele frequency is shown. Similar simulations were also performed at 5% disease prevalence and 7.5% causal allele frequency, 10% disease prevalence and 15% causal allele frequency as well as at 15% disease prevalence and 22.5% causal allele frequency.

Defined source population
(*Disease prevalence = 1%*: 99% healthy + 1% patients)

Genotype frequencies:
(*Ref. allele A & causal allele C*)
$AA_P = (0.985 \times 0.985)$
$AC_P = 2(0.985 \times 0.015)$
$CC_P = (0.015 \times 0.015)$
Patient and non-patient frequencies:
$Healthy_P = 0.99$ (predisposed genotypes = 0.005)
$Patients_P = 0.01$

1. Calculating genotypes for cases witihin cohort
(n = 10000: 0.0% controls + 100% cases, *proportion of cases being associated with allele C = Z*)
Cases with no genetic root:
$AA_C = [AA_P \div (AA_P + 0.5AC_P)] \times n(1\text{-}Z)$
$AC_{C1} = [0.5AC_P \div (AA_P + 0.5AC_P)] \times n(1\text{-}Z)$
Cases with genetic root:
$AC_{C2} = [0.5AC_P \div (CC_P + 0.5AC_P)] \times nZ$
$CC_C = [CC_P \div (CC_P + 0.5AC_P)] \times nZ$

2. Calculating genotypes for pure controls
(n = 100000: 100% controls + 0.0% cases)
$AA_{PH} = [(100000 \times AA_P) - (0.1AA_C)] \times 10 \div 9.9$
$AC_{PH} = [(100000 \times 0.5AC_P) - (0.1AC_{C1})] \times 10 \div 9.9 + [(100000 \times 0.5AC_P) - (0.1AC_{C2})] \times 10 \div 9.9$
$CC_{PH} = [(100000 \times CC_P) - (0.1CC_C)] \times 10 \div 9.9$

3. Calculating genotypes for controls within cohort (n = 100000: 99.5% controls + 0.5% cases)
$AA_H = 0.995AA_{PH} + 0.05AA_C$
$AC_H = 0.995AC_{PH} + 0.05AC_{C1} + 0.05AC_{C2}$
$CC_H = 0.995CC_{PH} + 0.05CC_C$

4. Calculating genotypes for filtered controls (age ≥ 65 years) within cohort (n = 10000: 99.99% controls and 0.01% cases)
$AA_H = 0.0999AA_{PH} + 0.001AA_C$
$AC_H = 0.0999AC_{PH} + 0.001AC_{C1} + 0.001AC_{C2}$
$CC_H = 0.0999CC_{PH} + 0.001CC_C$

Text-Box S2. Normalization of genetic data for differences of genotype frequencies within the population prior to examining the allelic interaction models for each SNP. This normalization is required for qualitative traits to be analyzed through logistic regression. Genotype frequencies within population (XX_P), case-samples specific genotype frequencies within the cohort (XX_C), normalized case-samples specific genotype frequencies within the cohort (XX_{CN}), control-samples specific genotype frequencies within the cohort (XX_H), and normalized control-samples specific genotype frequencies within the cohort (XX_{HN}).

Population:

Ref. Allele frequency = R, Alt. Allele frequency = A,

$RR_P = R \times R$, $RA_P = 2 \times R \times A$, $AA_P = A \times A$

Cohort:

Observed genotype frequencies among controls within the cohort

RR_H, RA_H, AA_H

Normalized genotype frequencies among controls within the cohort (Assuming R > A):

$RR_{HN} = RR_H$

$RA_{HN} = RA_H \times (RR_P \div RA_P)$

$AA_{HN} = AA_H \times (RR_P \div AA_P)$

Observed genotype frequencies among cases within the cohort

RR_c, RA_c, AA_c

Normalized genotype frequencies among cases within the cohort (Assuming R > A):

$RR_{CN} = RR_C$

$RA_{CN} = RA_C \times (RR_P \div RA_P)$

$AA_{CN} = AA_C \times (RR_P \div AA_P)$

Table S1. Additive-model based association analyses tend to produce biased risk-score estimates under conditions of non-additive allelic interactions and a pre-normalization step is needed to account for the differences in within-population genotype frequencies and thereby, for the detection of genetic effects through regression line fit.

DP[1]	SNPs	True effect[-1]	True risk score			Additive-model based risk score			Logistic-regression / Association *p*-values from additive and non-additive analysis models[1] after normalization				Logistic-regression / Association *p*-values from additive and non-additive analysis models[1] before normalization			
			RR[2]	RA[2]	AA[2]	RR[2]	RA[2]	AA[2]	Add	Dom	OD	Het	Add	Dom	OD	Het
1%[3]	SNP1	Add	0β =0	1β = 2.502	2β = 5.04	0β =0	1β = 2.502	2β = 5.04	1.0×10^{-2029}	1.4×10^{-1356}	3.9×10^{-657}	0.998	2.1×10^{-1276}	1.7×10^{-1234}	2.9×10^{-1288}	1.4×10^{-1275}
	SNP2	Add	0β =0	1β = 2.189	2β = 4.378	0β =0	1β = 2.189	2β = 4.378	3.8×10^{-1576}	2.9×10^{-1098}	3.4×10^{-496}	0.995	2.3×10^{-859}	1.1×10^{-871}	2.2×10^{-867}	3.8×10^{-861}
	SNP3	Add	0β =0	1β = 1.774	2β = 3.548	0β =0	1β = 1.774	2β = 3.548	2.1×10^{-1053}	1.8×10^{-767}	1.1×10^{-318}	1	6.4×10^{-459}	5.7×10^{-465}	1.3×10^{-462}	3.8×10^{-459}
	SNP4	Add	0β =0	1β = 1.129	2β = 2.258	0β =0	1β = 1.129	2β = 2.258	3.3×10^{-443}	2.3×10^{-337}	1.2×10^{-125}	1	9.9×10^{-126}	4.1×10^{-126}	1.9×10^{-125}	2.1×10^{-124}
	SNP5	Add	0β =0	1β = 0.656	2β = 1.312	0β =0	1β = 0.656	2β = 1.312	2.7×10^{-157}	6.7×10^{-121}	4.8×10^{-43}	1	4.2×10^{-31}	5.7×10^{-31}	9.1×10^{-31}	1.7×10^{-30}
	SNP6	Dom	0β = 0	2β = 5.116	2β = 5.116	0β = 0	1β = 2.451	2β = 4.902	9.7×10^{-886}	1.2×10^{-978}	9.7×10^{-886}	1.3×10^{-252}	9.7×10^{-886}	1.2×10^{-978}	9.7×10^{-886}	1.3×10^{-252}
	SNP7	Dom	0β = 0	2β = 4.394	2β = 4.394	0β = 0	1β = 2.122	2β = 4.244	2.3×10^{-645}	1.2×10^{-751}	2.3×10^{-645}	4.8×10^{-192}	2.3×10^{-645}	1.2×10^{-751}	2.3×10^{-645}	4.8×10^{-192}
	SNP8	Dom	0β = 0	2β = 3.442	2β = 3.442	0β = 0	1β = 1.677	2β = 3.354	4.4×10^{-383}	7.2×10^{-473}	4.4×10^{-383}	9.7×10^{-120}	4.4×10^{-383}	7.2×10^{-473}	4.4×10^{-383}	9.7×10^{-120}
	SNP9	Dom	0β = 0	2β = 1.921	2β = 1.921	0β = 0	1β = 0.944	2β = 1.889	2.8×10^{-110}	2.9×10^{-144}	2.8×10^{-110}	6.6×10^{-37}	2.8×10^{-110}	2.9×10^{-144}	2.8×10^{-110}	6.6×10^{-37}
	SNP10	Dom	0β = 0	2β = 0.704	2β = 0.704	0β = 0	1β = 0.348	2β = 0.695	9.5×10^{-15}	3.5×10^{-19}	9.5×10^{-15}	8.6×10^{-6}	9.5×10^{-15}	3.5×10^{-19}	9.5×10^{-15}	8.6×10^{-6}
	SNP11	OD	0β = 0	2β = 5.118	1β = 2.559	0β = 0	1β = 2.448	2β = 4.896	5.1×10^{-1212}	1.9×10^{-775}	1.2×10^{-1091}	2.3×10^{-249}	2.7×10^{-1211}	4.5×10^{-1226}	1.1×10^{-1225}	4.7×10^{-1224}
	SNP12	OD	0β = 0	2β = 4.396	1β = 2.198	0β = 0	1β = 2.120	2β = 4.240	8.7×10^{-233}	2.4×10^{-579}	1.2×10^{-787}	1.3×10^{-551}	2.3×10^{-788}	6.4×10^{-800}	6.4×10^{-800}	8.3×10^{-799}
	SNP13	OD	0β = 0	2β = 3.444	1β = 1.722	0β = 0	1β = 1.675	2β = 3.350	8.4×10^{-131}	2.1×10^{-384}	5.2×10^{-460}	2.6×10^{-326}	2.4×10^{-390}	1.6×10^{-396}	1.2×10^{-396}	4.9×10^{-396}
	SNP14	OD	0β = 0	2β = 1.924	1β = 0.962	0β = 0	1β = 0.944	2β = 1.888	1.9×10^{-34}	8.1×10^{-97}	4.1×10^{-126}	8.2×10^{-93}	3.9×10^{-78}	6.5×10^{-79}	6.2×10^{-79}	9.2×10^{-79}
	SNP15	OD	0β = 0	2β = 0.708	1β = 0.354	0β = 0	1β = 0.349	2β = 0.697	5.5×10^{-5}	3.2×10^{-12}	1.1×10^{-15}	4.8×10^{-12}	6.7×10^{-8}	5.4×10^{-8}	5.1×10^{-8}	5.1×10^{-8}
	SNP16	Het	0β = 0	1β = 2.564	0β = 0	0β = 0	1β = 2.446	2β = 4.892	1	3.2×10^{-539}	5.9×10^{-1266}	2.8×10^{-1617}	3.2×10^{-1213}	6.5×10^{-1227}	4.6×10^{-1228}	2.8×10^{-1228}
	SNP17	Het	0β = 0	1β = 2.202	0β = 0	0β = 0	1β = 2.118	2β = 4.236	1	2.3×10^{-385}	6.7×10^{-944}	9.4×10^{-1200}	2.1×10^{-789}	9.5×10^{-801}	7.1×10^{-802}	3.9×10^{-802}
	SNP18	Het	0β = 0	1β = 1.726	0β = 0	0β = 0	1β = 1.673	2β = 3.346	1	9.3×10^{-210}	8.8×10^{-545}	8.9×10^{-692}	4.7×10^{-390}	1.1×10^{-396}	1.7×10^{-397}	1.1×10^{-397}
	SNP19	Het	0β = 0	1β = 0.966	0β = 0	0β = 0	1β = 0.943	2β = 1.886	1	5.0×10^{-53}	1.2×10^{-146}	5.7×10^{-189}	9.3×10^{-78}	6.5×10^{-79}	3.8×10^{-79}	3.2×10^{-79}
	SNP20	Het	0β = 0	1β = 0.357	0β = 0	0β = 0	1β = 0.350	2β = 0.700	1	1.3×10^{-6}	8.0×10^{-17}	9.9×10^{-22}	4.7×10^{-8}	3.7×10^{-8}	3.6×10^{-8}	3.5×10^{-8}
15%[4]	SNP1	Add	0β =0	1β = 1.759	2β = 3.518	0β =0	1β = 1.759	2β = 3.518	1.3×10^{-1366}	2.5×10^{-723}	6.1×10^{-189}	1	4.7×10^{-2236}	3.5×10^{-2086}	5.0×10^{-1969}	1.7×10^{-1884}
	SNP2	Add	0β =0	1β = 1.637	2β = 3.274	0β =0	1β = 1.637	2β = 3.274	2.9×10^{-1215}	1.3×10^{-697}	4.3×10^{-350}	0.995	1.4×10^{-2001}	1.1×10^{-1884}	3.7×10^{-1743}	1.7×10^{-1489}
	SNP3	Add	0β =0	1β = 1.389	2β = 2.778	0β =0	1β = 1.389	2β = 2.778	8.7×10^{-921}	6.6×10^{-593}	2.7×10^{-258}	1	3.6×10^{-1487}	1.8×10^{-1403}	2.3×10^{-1267}	6.2×10^{-1062}
	SNP4	Add	0β =0	1β = 1.130	2β = 2.261	0β =0	1β = 1.130	2β = 2.261	2.6×10^{-641}	1.6×10^{-444}	2.6×10^{-175}	0.994	4.4×10^{-977}	1.3×10^{-915}	1.7×10^{-816}	2.6×10^{-676}
	SNP5	Add	0β =0	1β = 0.707	2β = 1.414	0β =0	1β = 0.707	2β = 1.414	2.9×10^{-270}	2.0×10^{-199}	1.1×10^{-71}	0.993	3.3×10^{-350}	2.8×10^{-324}	1.8×10^{-286}	4.8×10^{-235}
	SNP6	Dom	0β = 0	2β = 4.549	2β = 4.549	0β = 0	1β = 1.541	2β = 3.081	7.6×10^{-496}	3.1×10^{-505}	7.6×10^{-496}	1.3×10^{-252}	3.3×10^{-1807}	1.8×10^{-1910}	6.1×10^{-1902}	3.9×10^{-1792}
	SNP7	Dom	0β = 0	2β = 3.998	2β = 3.998	0β = 0	1β = 1.406	2β = 2.812	4.4×10^{-418}	4.8×10^{-457}	4.4×10^{-418}	1.5×10^{-145}	9.9×10^{-1526}	1.4×10^{-1636}	4.6×10^{-1615}	7.7×10^{-1509}
	SNP8	Dom	0β = 0	2β = 3.018	2β = 3.018	0β = 0	1β = 1.127	2β = 2.254	1.1×10^{-273}	8.1×10^{-329}	1.1×10^{-273}	1.2×10^{-94}	1.2×10^{-972}	9.5×10^{-1055}	1.8×10^{-1031}	1.3×10^{-953}
	SNP9	Dom	0β = 0	2β = 2.114	2β = 2.114	0β = 0	1β = 0.827	2β = 1.654	6.1×10^{-149}	2.5×10^{-189}	6.1×10^{-149}	1.6×10^{-51}	4.8×10^{-496}	4.9×10^{-539}	2.2×10^{-525}	2.1×10^{-483}
	SNP10	Dom	0β = 0	2β = 0.740	2β = 0.740	0β = 0	1β = 0.307	2β = 0.613	2.7×10^{-21}	9.9×10^{-28}	2.7×10^{-21}	4.4×10^{-8}	8.0×10^{-58}	1.1×10^{-62}	4.9×10^{-61}	4.6×10^{-56}
	SNP11	OD	0β = 0	2β = 4.527	1β = 2.264	0β = 0	1β = 1.487	2β = 2.975	1.2×10^{-173}	2.0×10^{-379}	1.4×10^{-561}	2.6×10^{-441}	4.3×10^{-1698}	8.4×10^{-1930}	7.8×10^{-1994}	2.4×10^{-1364}
	SNP12	OD	0β = 0	2β = 4.004	1β = 2.002	0β = 0	1β = 1.360	2β = 2.721	1.2×10^{-142}	1.4×10^{-350}	1.1×10^{-517}	8.8×10^{-400}	2.7×10^{-1433}	7.3×10^{-1663}	1.7×10^{-1711}	3.7×10^{-1677}
	SNP13	OD	0β = 0	2β = 3.062	1β = 1.531	0β = 0	1β = 1.097	2β = 2.195	5.4×10^{-92}	6.6×10^{-246}	7.6×10^{-342}	1.5×10^{-260}	4.3×10^{-920}	2.3×10^{-1089}	8.0×10^{-1118}	4.1×10^{-1089}
	SNP14	OD	0β = 0	2β = 2.184	1β = 1.092	0β = 0	1β = 0.815	2β = 1.631	5.8×10^{-50}	1.3×10^{-139}	3.3×10^{-188}	2.3×10^{-142}	3.2×10^{-481}	2.9×10^{-574}	4.5×10^{-589}	1.4×10^{-572}
	SNP15	OD	0β = 0	2β = 0.842	1β = 0.421	0β = 0	1β = 0.329	2β = 0.659	1.0×10^{-8}	5.1×10^{-23}	4.1×10^{-30}	5.9×10^{-23}	2.9×10^{-67}	1.9×10^{-80}	1.5×10^{-82}	3.6×10^{-80}
	SNP16	Het	0β = 0	1β = 2.416	0β = 0	0β = 0	1β = 1.427	2β = 2.854	1	1.0×10^{-266}	9.4×10^{-726}	4.9×10^{-919}	2.6×10^{-1574}	9.2×10^{-1949}	1.7×10^{-2087}	2.4×10^{-2138}
	SNP17	Het	0β = 0	1β = 2.138	0β = 0	0β = 0	1β = 1.308	2β = 2.616	1	3.4×10^{-231}	8.7×10^{-637}	3.0×10^{-812}	9.0×10^{-1328}	1.7×10^{-1688}	4.0×10^{-1813}	4.1×10^{-1857}
	SNP18	Het	0β = 0	1β = 1.645	0β = 0	0β = 0	1β = 1.061	2β = 2.122	1	1.1×10^{-154}	5.1×10^{-436}	1.0×10^{-561}	1.0×10^{-857}	1.0×10^{-1123}	9.1×10^{-1212}	7.7×10^{-1242}
	SNP19	Het	0β = 0	1β = 1.189	0β = 0	0β = 0	1β = 0.797	2β = 1.594	1	5.8×10^{-85}	2.5×10^{-243}	2.2×10^{-316}	4.9×10^{-458}	5.7×10^{-609}	3.8×10^{-659}	2.6×10^{-676}
	SNP20	Het	0β = 0	1β = 0.498	0β = 0	0β = 0	1β = 0.346	2β = 0.691	1	3.0×10^{-15}	3.0×10^{-42}	2.2×10^{-55}	3.4×10^{-74}	1.7×10^{-99}	4.0×10^{-108}	3.5×10^{-111}
15%[5]	SNP1	Add	0β =0	1β = 1.759	2β = 3.518	0β =0	1β = 1.759	2β = 3.518	8.4×10^{-1147}	9.7×10^{-693}	6.5×10^{-355}	0.959	1.5×10^{-2055}	1.2×10^{-1178}	1.3×10^{-1055}	4.5×10^{-843}
	SNP2	Add	0β =0	1β = 1.637	2β = 3.274	0β =0	1β = 1.637	2β = 3.274	4.7×10^{-1030}	3.4×10^{-664}	3.9×10^{-312}	0.989	1.1×10^{-915}	4.7×10^{-1016}	3.7×10^{-907}	3.2×10^{-723}
	SNP3	Add	0β =0	1β = 1.389	2β = 2.778	0β =0	1β = 1.389	2β = 2.778	2.1×10^{-792}	9.9×10^{-558}	1.3×10^{-231}	0.969	2.9×10^{-654}	3.2×10^{-711}	4.7×10^{-634}	2.4×10^{-506}
	SNP4	Add	0β =0	1β = 1.130	2β = 2.261	0β =0	1β = 1.130	2β = 2.261	8.1×10^{-557}	5.5×10^{-413}	3.9×10^{-158}	0.983	6.0×10^{-427}	4.9×10^{-451}	4.7×10^{-402}	1.1×10^{-322}
	SNP5	Add	0β =0	1β = 0.707	2β = 1.414	0β =0	1β = 0.707	2β = 1.414	9.1×10^{-241}	8.9×10^{-184}	3.6×10^{-60}	0.965	4.6×10^{-162}	7.8×10^{-162}	6.2×10^{-144}	1.4×10^{-116}
	SNP6	Dom	0β = 0	2β = 4.549	2β = 4.549	0β = 0	1β = 1.541	2β = 3.081	3.3×10^{-443}	7.3×10^{-483}	3.3×10^{-443}	1.6×10^{-145}	2.1×10^{-904}	7.0×10^{-1035}	6.2×10^{-993}	1.8×10^{-893}
	SNP7	Dom	0β = 0	2β = 3.998	2β = 3.998	0β = 0	1β = 1.406	2β = 2.812	1.0×10^{-373}	9.4×10^{-433}	1.8×10^{-374}	2.2×10^{-123}	6.2×10^{-742}	8.4×10^{-850}	5.4×10^{-816}	6.4×10^{-734}
	SNP8	Dom	0β = 0	2β = 3.018	2β = 3.018	0β = 0	1β = 1.127	2β = 2.254	4.9×10^{-246}	1.2×10^{-307}	2.6×10^{-246}	1.1×10^{-81}	4.8×10^{-459}	1.7×10^{-522}	2.1×10^{-503}	4.5×10^{-454}
	SNP9	Dom	0β = 0	2β = 2.114	2β = 2.114	0β = 0	1β = 0.827	2β = 1.654	7.2×10^{-135}	2.2×10^{-175}	1.1×10^{-134}	3.1×10^{-45}	2.6×10^{-236}	2.6×10^{-265}	1.3×10^{-256}	1.1×10^{-232}
	SNP10	Dom	0β = 0	2β = 0.740	2β = 0.740	0β = 0	1β = 0.307	2β = 0.613	1.9×10^{-19}	1.6×10^{-25}	1.5×10^{-19}	1.8×10^{-7}	1.4×10^{-30}	2.4×10^{-33}	1.9×10^{-32}	9.4×10^{-30}
	SNP11	OD	0β = 0	2β = 4.527	1β = 2.264	0β = 0	1β = 1.487	2β = 2.975	1.1×10^{-158}	3.9×10^{-377}	1.4×10^{-547}	5.9×10^{-409}	2.1×10^{-902}	1.4×10^{-1040}	1.3×10^{-1046}	2.0×10^{-1000}
	SNP12	OD	0β = 0	2β = 4.004	1β = 2.002	0β = 0	1β = 1.360	2β = 2.721	2.2×10^{-131}	8.2×10^{-333}	3.6×10^{-465}	1.4×10^{-346}	7.0×10^{-739}	2.5×10^{-866}	1.3×10^{-865}	6.1×10^{-827}
	SNP13	OD	0β = 0	2β = 3.062	1β = 1.531	0β = 0	1β = 1.097	2β = 2.195	5.3×10^{-85}	8.3×10^{-231}	9.1×10^{-308}	9.4×10^{-228}	8.4×10^{-457}	1.9×10^{-540}	7.7×10^{-543}	1.4×10^{-519}
	SNP14	OD	0β = 0	2β = 2.184	1β = 1.092	0β = 0	1β = 0.815	2β = 1.631	1.9×10^{-46}	5.2×10^{-130}	4.6×10^{-170}	1.0×10^{-125}	3.6×10^{-238}	1.1×10^{-282}	1.3×10^{-285}	2.1×10^{-274}
	SNP15	OD	0β = 0	2β = 0.842	1β = 0.421	0β = 0	1β = 0.329	2β = 0.659	4.1×10^{-8}	2.8×10^{-21}	1.5×10^{-27}	6.9×10^{-21}	9.3×10^{-36}	2.5×10^{-42}	3.6×10^{-43}	8.6×10^{-42}
	SNP16	Het	0β = 0	1β = 2.416	0β = 0	0β = 0	1β = 1.427	2β = 2.854	0.978	5.5×10^{-256}	7.2×10^{-660}	9.9×10^{-848}	6.0×10^{-978}	5.9×10^{-1054}	1.1×10^{-1109}	9.4×10^{-1121}
	SNP17	Het	0β = 0	1β = 2.138	0β = 0	0β = 0	1β = 1.308	2β = 2.616	0.988	3.2×10^{-220}	2.1×10^{-577}	2.6×10^{-743}	1.5×10^{-734}	1.3×10^{-882}	1.9×10^{-919}	8.1×10^{-933}
	SNP18	Het	0β = 0	1β = 1.595	0β = 0	0β = 0	1β = 1.027	2β = 2.053	0.982	8.9×10^{-146}	8.9×10^{-394}	1.5×10^{-507}	1.3×10^{-452}	2.2×10^{-558}	6.8×10^{-586}	3.2×10^{-595}
	SNP19	Het	0β = 0	1β = 1.189	0β = 0	0β = 0	1β = 0.797	2β = 1.594	0.951	1.6×10^{-79}	6.8×10^{-220}	2.9×10^{-284}	3.4×10^{-237}	2.1×10^{-299}	8.5×10^{-317}	1.1×10^{-322}
	SNP20	Het	0β = 0	1β = 0.498	0β = 0	0β = 0	1β = 0.346	2β = 0.691	0.953	2.9×10^{-14}	1.5×10^{-38}	4.6×10^{-50}	4.9×10^{-40}	6.2×10^{-52}	9.9×10^{-56}	4.7×10^{-57}

[1]DP: Disease prevalence. Add: Additive allelic effect; the causal allele in the homozygous state triggers the disease at twice the rate compared to its heterozygous state. Dom: Dominance allelic effect; the causal allele triggers disease at the same rate regardless of its dosage. OD: Overdominance allelic effect; the causal allele in the heterozygous state triggers the disease at twice the rate compared to its homozygous state. Het: Heterosis-like allelic effect; the causal allele in the heterozygous state triggers the disease but not at homozygous state.

[2]RR: Reference_Reference genotype, RA: Reference_Alternative genotype, AA: Alternative_Alternative genotype

[3]Analyses were based on a simulated cohorts (n = 110000, controls = 100000, cases = 10000, disease prevalence = 1%).

[4]Analyses were based on a simulated cohorts (n = 110000, controls = 100000, cases = 10000, disease prevalence = 15%).

[5]Analyses were based on a simulated cohorts (n = 20000, controls = 10000, cases = 10000, disease prevalence = 15%).

Logistic regression was performed before and after normalization for differences in within-population genotype frequencies. The identified false and true genetic effects are highlighted by the lowest *p*-values, shown in red and black, respectively.

β was calculated as the natural logarithm of the odds ratio, using odds ratios derived from the original (non-normalized) data.

Table S2. The Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) were not capable of distinguishing between additive and overdominance effects.

DP[1]	SNPs	True effect[1]	True risk score			Additive-model based risk score			AIC values derived from logistic regression using additive and non-additive analysis models[1]				BIC values derived from logistic regression using additive and non-additive analysis models[1]			
			RR[2]	RA[2]	AA[2]	RR[2]	RA[2]	AA[2]	Add	Dom	OD	Het	Add	Dom	OD	Het
1%	SNP1	Add	0β =0	1β = 2.502	2β = 5.04	0β =0	1β = 2.502	2β = 5.04	67981.2	69339.9	67981.2	81124.3	68008.1	69357.8	68008.1	81142.2
	SNP2	Add	0β =0	1β = 2.189	2β = 4.378	0β =0	1β = 2.189	2β = 4.378	61295.7	62412.6	61295.7	71165	61322.2	62430.3	61322.2	71183.1
	SNP3	Add	0β =0	1β = 1.774	2β = 3.548	0β =0	1β = 1.774	2β = 3.548	52160	52964.2	52160	58426.4	52186.1	52981.6	52186.1	58443.8
	SNP4	Add	0β =0	1β = 1.129	2β = 2.258	0β =0	1β = 1.129	2β = 2.258	38598.4	38973.2	38598.4	40977.3	38624	38990.3	38624	40994.3
	SNP5	Add	0β =0	1β = 0.656	2β = 1.312	0β =0	1β = 0.656	2β = 1.312	29817.2	29959.4	29817.2	30585.6	29842.6	29976.3	29842.6	30602.5
	SNP6	Dom	0β = 0	2β = 5.116	2β = 5.116	0β = 0	1β = 2.451	2β = 4.902	58896.3	58894.3	58896.3	65476.7	589227	58911.9	58922.7	65494.3
	SNP7	Dom	0β = 0	2β = 4.394	2β = 4.394	0β = 0	1β = 2.122	2β = 4.244	52176.2	52174.2	52176.2	56718	52202.3	52191.6	52202.3	56735.4
	SNP8	Dom	0β = 0	2β = 3.442	2β = 3.442	0β = 0	1β = 1.677	2β = 3.354	43396.6	43394.4	43396.6	45876	43422.4	43411.8	43422.4	45893.2
	SNP9	Dom	0β = 0	2β = 1.921	2β = 1.921	0β = 0	1β = 0.944	2β = 1.889	30959.9	30957.9	30959.9	31562.2	30985.3	30974.9	30985.3	31579.1
	SNP10	Dom	0β = 0	2β = 0.704	2β = 0.704	0β = 0	1β = 0.348	2β = 0.695	23229.8	23227.8	23229.8	23291.8	232555	23244.6	23255	23308.6
	SNP11	OD	0β = 0	2β = 5.118	1β = 2.559	0β = 0	1β = 2.448	2β = 4.896	51776.6	52606.7	51776.6	54803.7	51802.7	52621.1	51802.7	54821.1
	SNP12	OD	0β = 0	2β = 4.396	1β = 2.198	0β = 0	1β = 2.120	2β = 4.240	45902.4	46530	45902.4	47854.6	45928.3	46547.2	45928.3	47871.9
	SNP13	OD	0β = 0	2β = 3.444	1β = 1.722	0β = 0	1β = 1.675	2β = 3.350	38470	38857.8	38470	39434.5	38495.6	38874.9	38495.6	39451.6
	SNP14	OD	0β = 0	2β = 1.924	1β = 0.962	0β = 0	1β = 0.944	2β = 1.888	28375.1	28490.1	28375.1	28569.4	28400.4	28507	28400.4	28586.3
	SNP15	OD	0β = 0	2β = 0.708	1β = 0.354	0β = 0	1β = 0.349	2β = 0.697	22372.4	22385.1	22372.4	22388.2	22397.6	22401.9	22397.6	22405
	SNP16	Het	0β = 0	1β = 2.564	0β = 0	0β = 0	1β = 2.446	2β = 4.892	38025.9	44643.6	38025.9	38023.9	38051.7	44660.8	38051.7	38041.1
	SNP17	Het	0β = 0	1β = 2.202	0β = 0	0β = 0	1β = 2.118	2β = 4.236	35044.9	39615.3	35044.9	35042.9	35070.6	39632.4	35070.6	35060
	SNP18	Het	0β = 0	1β = 1.726	0β = 0	0β = 0	1β = 1.673	2β = 3.346	30796.1	33220.2	30796.1	30794.1	30821.6	33237.2	30821.6	30811.1
	SNP19	Het	0β = 0	1β = 0.966	0β = 0	0β = 0	1β = 0.943	2β = 1.886	25160.3	25770.6	25160.3	25158.3	25185.5	25787.5	25185.5	25175.1
	SNP20	Het	0β = 0	1β = 0.357	0β = 0	0β = 0	1β = 0.350	2β = 0.700	21441.4	21505.5	21441.4	21439.4	21466.6	21522.3	21466.6	21456.2
15%	SNP1	Add	0β =0	1β = 1.759	2β = 3.518	0β =0	1β = 1.759	2β = 3.518	44316	45788.5	44316	53088.8	44342.4	45806.1	44342.4	53106.4
	SNP2	Add	0β =0	1β = 1.637	2β = 3.274	0β =0	1β = 1.637	2β = 3.274	44055	45339.8	44055	51617.5	44081.4	45357.4	44081.4	51617.5
	SNP3	Add	0β =0	1β = 1.389	2β = 2.778	0β =0	1β = 1.389	2β = 2.778	43047.4	43990.1	43047.4	48432.7	43073.8	44007.6	43073.8	48450.2
	SNP4	Add	0β =0	1β = 1.130	2β = 2.261	0β =0	1β = 1.130	2β = 2.261	41377.5	42018.3	41377.5	44898.5	41403.8	42035.9	41403.8	44916.1
	SNP5	Add	0β =0	1β = 0.707	2β = 1.414	0β =0	1β = 0.707	2β = 1.414	37507.6	3777.6	37507.6	38851.8	37533.9	37791.1	37533.9	38869.3
	SNP6	Dom	0β = 0	2β = 4.549	2β = 4.549	0β = 0	1β = 1.541	2β = 3.081	40212.6	40210.6	40212.6	43523.7	40239	40228.2	40239	43541.2
	SNP7	Dom	0β = 0	2β = 3.998	2β = 3.998	0β = 0	1β = 1.406	2β = 2.812	39609.6	39607.6	39609.6	42290.4	39635.9	39625.1	39635.9	42308
	SNP8	Dom	0β = 0	2β = 3.018	2β = 3.018	0β = 0	1β = 1.127	2β = 2.254	38019.3	38017.3	38019.3	39636.3	38045.6	38034.8	38045.6	39636.3
	SNP9	Dom	0β = 0	2β = 2.114	2β = 2.114	0β = 0	1β = 0.827	2β = 1.654	35911.2	35909.2	35911.2	36717.6	35937.5	35926.8	35937.5	36735.2
	SNP10	Dom	0β = 0	2β = 0.740	2β = 0.740	0β = 0	1β = 0.307	2β = 0.613	31664.3	31662.3	31664.3	31757.3	31690.6	31679.8	31690.6	31774.8
	SNP11	OD	0β = 0	2β = 4.527	1β = 2.264	0β = 0	1β = 1.487	2β = 2.975	36315.8	36793	36315.8	37649.9	36342.2	36810.6	36342.2	37667.5
	SNP12	OD	0β = 0	2β = 4.004	1β = 2.002	0β = 0	1β = 1.360	2β = 2.721	35470.7	35978.7	35470.7	36519.6	35497.1	35996.3	35497.1	36537.2
	SNP13	OD	0β = 0	2β = 3.062	1β = 1.531	0β = 0	1β = 1.097	2β = 2.195	34617	34949.8	34617	35216.1	34643.3	34967.4	34643.3	35233.7
	SNP14	OD	0β = 0	2β = 2.184	1β = 1.092	0β = 0	1β = 0.815	2β = 1.631	33412.6	33596.5	33412.6	33698	33438.9	33614	33438.9	33715.6
	SNP15	OD	0β = 0	2β = 0.842	1β = 0.421	0β = 0	1β = 0.329	2β = 0.659	30890.5	30918.5	30890.5	30924.6	30916.8	30936.1	30916.8	30942.2
	SNP16	Het	0β = 0	1β = 2.416	0β = 0	0β = 0	1β = 1.427	2β = 2.854	27333.2	31166	27333.2	27331.2	27359.5	31183.6	27359.5	27348.7
	SNP17	Het	0β = 0	1β = 2.138	0β = 0	0β = 0	1β = 1.308	2β = 2.616	28005.6	31151.2	28005.6	28003.6	28031.9	31168.7	28031.9	28021.2
	SNP18	Het	0β = 0	1β = 1.645	0β = 0	0β = 0	1β = 1.061	2β = 2.122	29001.4	30969	29001.4	28999.4	29027.8	30986.5	29027.8	29017
	SNP19	Het	0β = 0	1β = 1.189	0β = 0	0β = 0	1β = 0.797	2β = 1.594	29565	30610	29565	29563	29591.3	30628.4	29591.3	29580.5
	SNP20	Het	0β = 0	1β = 0.498	0β = 0	0β = 0	1β = 0.346	2β = 0.691	29598.6	29772.7	29598.6	29596.6	29625	29790.3	29625	29614.2

[1]AIC: Akaike Information Criterion. BIC: Bayesian Information Criterion. DP: Disease prevalence. Add: Additive allelic effect; the causal allele in the homozygous state triggers the disease at twice the rate compared to its heterozygous state. Dom: Dominance allelic effect; the causal allele triggers disease at the same rate regardless of its dosage. OD: Overdominance allelic effect; the causal allele in the heterozygous state triggers the disease at twice the rate compared to its homozygous state. Het: Heterosis-like allelic effect; the causal allele in the heterozygous state triggers the disease but not at homozygous state.

[2]RR: Reference_Reference genotype, RA: Reference_Alternative genotype, AA: Alternative_Alternative genotype

Analyses were based on a simulated cohorts (n = 110000, controls = 100000, cases = 10000, disease prevalence = 1% or 15%).

Prior to logistic regression, the genetic data were normalized to account for differences in within-population genotype frequencies.

β was calculated as the natural logarithm of the odds ratio, using odds ratios derived from the original (non-normalized) data.

Table S3. Additive-model based association analyses tend to produce biased risk-score estimates under conditions of non-additive allelic interactions.

DP[1]	SNPs	True effect[1]	True risk score			Additive-model based risk score			Logistic-regression / Association *p*-values from additive and non-additive analysis models[1] after normalization				Genotype-specific case/control ratios [Genotype (case/control ratio)]		
			RR[2]	RA[2]	AA[2]	RR[2]	RA[2]	AA[2]	Add	Dom	OD	Het	Q1	Q2-3	Q4
15%[3]	SNP1	Add	0β =0	1β = 1.759	2β = 3.518	0β =0	1β = 1.759	2β = 3.518	8.4×10^{-1147}	9.7×10^{-693}	6.5×10^{-355}	0.959	RR (0.27)	RA (3.09)	AA (7.71)
	SNP2	Add	0β =0	1β = 1.637	2β = 3.274	0β =0	1β = 1.637	2β = 3.274	4.7×10^{-1030}	3.4×10^{-664}	3.9×10^{-312}	0.989	RR (0.33)	RA (2.89)	AA (6.91)
	SNP3	Add	0β =0	1β = 1.389	2β = 2.778	0β =0	1β = 1.389	2β = 2.778	2.1×10^{-792}	9.9×10^{-558}	1.3×10^{-231}	0.969	RR (0.45)	RA (2.51)	AA (5.45)
	SNP4	Add	0β =0	1β = 1.130	2β = 2.261	0β =0	1β = 1.130	2β = 2.261	8.1×10^{-557}	5.5×10^{-413}	3.9×10^{-158}	0.983	RR (0.57)	RA (2.14)	AA (4.19)
	SNP5	Add	0β =0	1β = 0.707	2β = 1.414	0β =0	1β = 0.707	2β = 1.414	9.1×10^{-241}	8.9×10^{-184}	3.6×10^{-65}	0.965	RR (0.75)	RA (1.63)	AA (2.65)
	SNP6	Dom	0β = 0	2β = 4.549	2β = 4.549	0β = 0	1β = 1.541	2β = 3.081	3.3×10^{-443}	7.3×10^{-483}	3.3×10^{-443}	1.6×10^{-145}	RR (0.33)		RA,AA (3.18)
	SNP7	Dom	0β = 0	2β = 3.998	2β = 3.998	0β = 0	1β = 1.406	2β = 2.812	1.0×10^{-373}	9.4×10^{-433}	1.8×10^{-374}	2.2×10^{-123}	RR (0.39)		RA,AA (2.19)
	SNP8	Dom	0β = 0	2β = 3.018	2β = 3.018	0β = 0	1β = 1.127	2β = 2.254	4.9×10^{-246}	1.2×10^{-307}	2.6×10^{-246}	1.1×10^{-81}	RR (0.53)		RA,AA (2.41)
	SNP9	Dom	0β = 0	2β = 2.114	2β = 2.114	0β = 0	1β = 0.827	2β = 1.654	7.2×10^{-135}	2.2×10^{-175}	1.1×10^{-134}	3.1×10^{-45}	RR (0.67)		RA,AA (1.94)
	SNP10	Dom	0β = 0	2β = 0.740	2β = 0.740	0β = 0	1β = 0.307	2β = 0.613	1.9×10^{-19}	1.6×10^{-25}	1.5×10^{-19}	1.8×10^{-7}	RR (0.89)		RA,AA (1.29)
	SNP11	OD	0β = 0	2β = 4.527	1β = 2.264	0β = 0	1β = 1.487	2β = 2.975	1.1×10^{-158}	3.9×10^{-377}	1.4×10^{-547}	5.9×10^{-409}	RR (0.32)	AA (1.67)	RA (3.36)
	SNP12	OD	0β = 0	2β = 4.004	1β = 2.002	0β = 0	1β = 1.360	2β = 2.721	2.2×10^{-131}	8.2×10^{-333}	3.6×10^{-465}	1.4×10^{-346}	RR (0.38)	AA (1.59)	RA (3.07)
	SNP13	OD	0β = 0	2β = 3.062	1β = 1.531	0β = 0	1β = 1.097	2β = 2.195	5.3×10^{-85}	8.3×10^{-231}	9.1×10^{-308}	9.4×10^{-228}	RR (0.52)	AA (1.45)	RA (2.53)
	SNP14	OD	0β = 0	2β = 2.184	1β = 1.092	0β = 0	1β = 0.815	2β = 1.631	1.9×10^{-46}	5.2×10^{-130}	4.6×10^{-170}	1.0×10^{-125}	RR (0.66)	AA (1.31)	RA (2.04)
	SNP15	OD	0β = 0	2β = 0.842	1β = 0.421	0β = 0	1β = 0.329	2β = 0.659	4.1×10^{-8}	2.8×10^{-21}	1.5×10^{-27}	6.9×10^{-21}	RR (0.88)	AA (1.11)	RA (1.35)
	SNP16	Het	0β = 0	1β = 2.416	0β = 0	0β = 0	1β = 1.427	2β = 2.854	0.978	5.5×10^{-256}	7.2×10^{-660}	9.9×10^{-848}	RR,AA (0.32)		RA (3.55)
	SNP17	Het	0β = 0	1β = 2.138	0β = 0	0β = 0	1β = 1.308	2β = 2.616	0.988	3.2×10^{-220}	2.1×10^{-577}	2.6×10^{-743}	RR,AA (0.38)		RA (3.25)
	SNP18	Het	0β = 0	1β = 1.595	0β = 0	0β = 0	1β = 1.027	2β = 2.053	0.982	8.9×10^{-146}	8.9×10^{-394}	1.5×10^{-507}	RR,AA (0.52)		RA (2.57)
	SNP19	Het	0β = 0	1β = 1.189	0β = 0	0β = 0	1β = 0.797	2β = 1.594	0.951	1.6×10^{-79}	6.8×10^{-220}	2.9×10^{-284}	RR,AA (0.65)		RA (2.14)
	SNP20	Het	0β = 0	1β = 0.498	0β = 0	0β = 0	1β = 0.346	2β = 0.691	0.953	2.9×10^{-14}	1.5×10^{-38}	4.6×10^{-50}	RR,AA (0.86)		RA (1.42)

[1]DP: Disease prevalence. Add: Additive allelic effect; the causal allele in the homozygous state triggers the disease at twice the rate compared to its heterozygous state. Dom: Dominance allelic effect; the causal allele triggers disease at the same rate regardless of its dosage. OD: Overdominance allelic effect; the causal allele in the heterozygous state triggers the disease at twice the rate compared to its homozygous state. Het: Heterosis-like allelic effect; the causal allele in the heterozygous state triggers the disease but not at homozygous state.

[2]RR: Reference_Reference genotype, RA: Reference_Alternative genotype, AA: Alternative_Alternative genotype

[3]Analyses were based on a simulated cohorts (n = 20000, controls = 10000, cases = 10000, disease prevalence = 15%).

Logistic regression was performed before and after normalization for differences in within-population genotype frequencies. The identified true genetic effects are highlighted by the lowest *p*-values, shown in black.

β was calculated as the natural logarithm of the odds ratio, using odds ratios derived from the original (non-normalized) data.

Case-to-control ratio for every genotype (RR,RA,AA) was calculated and quartiles (Q1–Q4) were delineated according to the range of case-to-control ratios, i.e., from minimum to maximum, and the distribution of genotype specific ratios within these quartiles was used as an alternative approach to assess allelic effects.