Genome-Wide Association Study Heterogeneous Cohort Homogenization via Subject Weight Knock-Down

André X. C. N. Valente[1,2,3]*, Joseph Zischkau[4], Joo Heon Shin[5], Yuan Gao[5], Abhijit Sarkar[4]

1 Systems Biology Group, Biocant – Biotechnology Innovation Center, Cantanhede, Portugal, **2** Center for Neuroscience and Cell Biology, University of Coimbra, Coimbra, Portugal, **3** Center for the Study of Biocomplexity, Virginia Commonwealth University, Richmond, Virginia, United States of America, **4** Physics Department and Vitreous State Laboratory, Catholic University of America, Washington DC, United States of America, **5** Lieber Institute for Brain Development, Johns Hopkins Medical Campus, Baltimore, Maryland, United States of America

Abstract

Population structure can be a source of both false-positive and false-negative findings in a genome-wide association study. This article proposes an approach that helps to reduce the false-positives. It consists of homogenizing the diseased/healthy phenotype ratio across the cohort, by decreasing the statistical weight of selected individuals. After homogenization, the cohort is statistically handled as if originating from a single well-mixed population. The method was applied to homogenize a Parkinson's disease genome-wide association study cohort.

Citation: Valente AXCN, Zischkau J, Shin JH, Gao Y, Sarkar A (2012) Genome-Wide Association Study Heterogeneous Cohort Homogenization via Subject Weight Knock-Down. PLoS ONE 7(10): e48653. doi:10.1371/journal.pone.0048653

Editor: Frank Emmert-Streib, Queen's University Belfast, United Kingdom

Received March 13, 2012; **Accepted** September 28, 2012; **Published** October 29, 2012

Funding: The authors have no support or funding to report.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: andre.valente@biocant.pt

Introduction

The genome-wide association study (GWAS) is nowadays routinely used to discover genetic susceptibilities to disease [1,2,3]. In these studies, a critical step is the handling of population structure in the analyzed cohort [4,5]. If not correctly accounted for, population structure can result in both false-positive and false-negative phenotype-genotype associations. We briefly review some of the main approaches currently utilized to identify and correct for the presence of population structure in a cohort [6,7,8]. Genomic control is a computationally fast and easy to implement method [9,10,11]. It prescribes a reduction by a factor λ in the cohort size utilized in test statistical calculations. This compensates for statistical significance inflation due to the presence of population structure. The method assumes that only the few strongest statistical associations reflect genuine phenotype-genotype associations and thus it estimates the λ factor based on the remaining bulk of the test statistics distribution. Dadd et al. [12] discuss refinements and variations on the genomic control approach. An example is the use of multiple rather than a single adjustment factor [13]. A different approach is to first capture ancestry by changing the cohort data to the principal component coordinates of a space defined by a set of markers assumed to be independent of the trait under analysis [14,15]. The first few principal components can then be utilized as regression covariates in the subsequent association analysis [16,17]. Other population structure correction approaches based on the calculated principal components have also been proposed [18,19,20,21]. As an alternative to principal component analysis, population structure can also be captured by the multidimensional-scaling (MDS) statistical technique [20,22,23]. Li et al. [24], report a method that combines MDS with a phylogeny constructed from SNP genotypes. Spectral graph theory provides yet a different way to capture genetic ancestry. Two implementations of this approach are Spectral-GEM [25] and LAP-STRUCT [26]. Structured association methods first assign to individuals probabilities of membership in given subpopulations [27,28]. Association testing is then conditional on these subpopulation membership probabilities [29]. STRUCTURE/STRAT [30] and ADMIXMAP [31] are standard software packages that implement this method. Structured association approaches tend to be computationally intensive, but the GWAS analysis package Plink [32] includes a simplified, efficient version of structured association. Finally, linear mixed models [33,34] have been also successfully applied to address population structure. Wu reports a performance comparison of some of the above approaches [35].

To assist in reducing the specific case of false-positives, this article suggests the additional avenue of homogenizing the ratio between the two GWAS phenotypes (e.g., diseased and healthy) throughout the cohort. The homogenization is performed within a principal component coordinates space and is accomplished by knocking-down the statistical weight of selected individuals. After homogenization, the cohort is statistically handled as if originating from a single well-mixed population. First, under the idealization of exactly two distinct populations, we recall the biases introduced by population structure in a GWAS. We then present our homogenization approach for the practical case where the cohort population structure has a continuous character. The method is described alongside its application to the homogenization of a Parkinson's disease GWAS cohort [36]. Finally, the method is tested using simulated, synthetic data.

Analysis

Two populations case

Consider a population of individuals classified into two genotypes (A and ~A) and likewise classified into two phenotypes (diseased and healthy). The genotype-phenotype population odds ratio (OR) [37] quantifies the degree of correlation between genotype and phenotype intrinsic to the population. A cohort sampled from the population provides an estimate of the OR. One of the four degrees of freedom (DOFs) of the sampled cohort's 2×2 contingency table (Figure 1-a) can thus be assigned to the OR estimate. Call it the *OR DOF*. The remaining three DOFs then reflect how the cohort was sampled from the population. These three *sampling DOFs* may be expressed as:

p_1 - number of patients sampled,

p_2 - number of healthy subjects sampled and

p_3 - sampling partiality towards A subjects versus towards ~A subjects.

Any selection for the three sampling DOFs is valid, as they are independent of the OR DOF. Still, correctness of the OR estimate requires every patient and healthy subject to be sampled under the same partiality value p_3. However, population differences in the relative prevalence of A and ~A subjects are conducive to different p_3 sampling partialities. Therefore casually combining samples from different populations is problematic, as illustrated by the following two examples. Consider two populations X and Y, where in both the OR is unity (Figure 1-b). Let population X have a preponderance of genotype A and population Y have a preponderance of genotype ~A, thus inducing different sampling partialities p_3. Additionally, regarding DOFs p_1 and p_2, suppose in population X mostly patients were sampled (i.e., $p_1 >> p_2$) while in population Y mostly healthy controls were sampled (i.e., $p_1 << p_2$). Then, combining the samples from X and Y in a single cohort results in a non-unity OR estimate, a false-positive genotype-phenotype correlation due population structure. As a second example, consider two populations X and Y, where in both the OR≈2 (Figure 1-c). As in the previous example, suppose p_3 is much larger in X than in Y. Let the exact same number of diseased and healthy subjects be sampled in X and in Y. Then, combining the samples from X and Y in a single cohort results in an OR estimate that is approximately unity, in spite of the OR in both the X and Y populations being approximately 2. This second example illustrates population structure concealing a genuine genotype-phenotype correlation. For both examples, we emphasize how the OR, and thus the genotype-phenotype correlation, is identical in the X and Y populations. Population structure led to the false-positive and false-negative calls solely by affecting sampling.

Transformations can be applied to the contingency table. Discarding diseased samples is akin to scaling by a less than unity common factor the contingency table entries *a* and *b* (Figure 1-a). Similarly, discarding healthy samples is akin to scaling by a less than unity common factor the entries *c* and *d*. Equivalently, we may always plausibly assume a cohort resulted from sampling a selected number of A subjects and a selected number of ~A subjects, all under a given sampling partiality between diseased and healthy subjects. From this perspective, the three sampling DOFs could be expressed as:

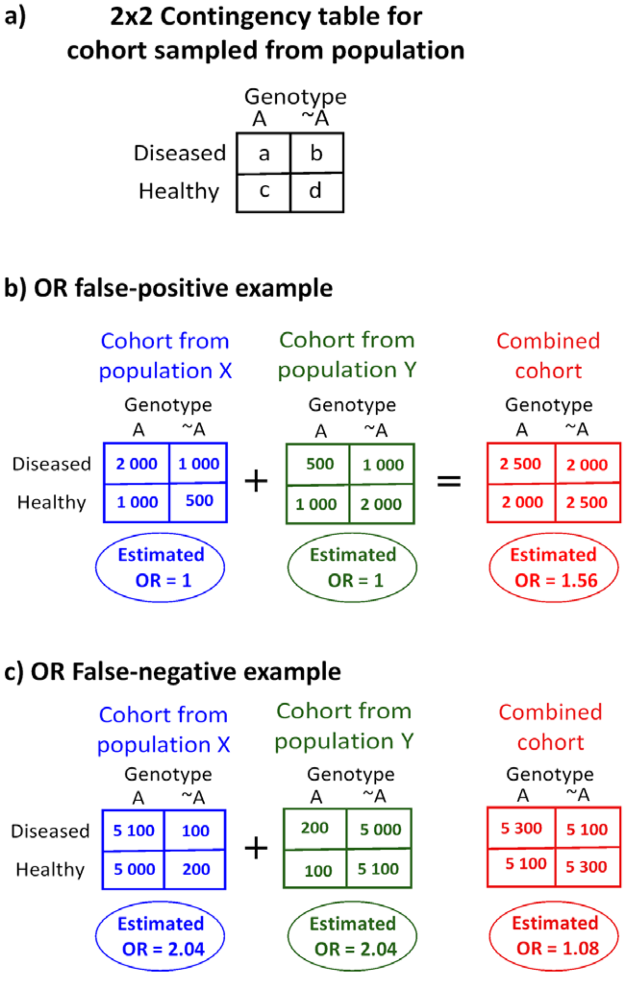

Figure 1. Odds ratio estimation biases introduced by population structure. a) The 2×2 contingency table associated with a cohort sampled from the population. Merging cohorts from distinct populations can produce both false-positive and false-negative assessments of the odds ratio (OR). Differences in the sampling process invalidate a straightforward combination of cohorts from distinct populations. **b**) False-positive example: In both population X and population Y, the OR = 1. **c**) False-negative example: In both population X and population Y, the OR≈2.
doi:10.1371/journal.pone.0048653.g001

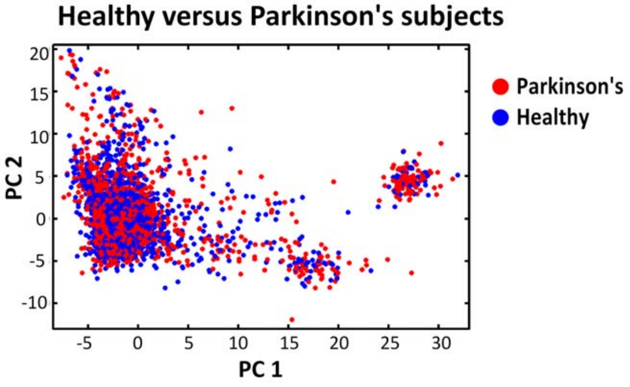

Figure 2. Distribution of patients and controls in the original cohort. The Hamza et al. [36] cohort healthy and Parkinson's individuals projected on the first 2 principal components of the SNP space. The cohort contains a total of 2000 Parkinson's patients and 1986 healthy controls.
doi:10.1371/journal.pone.0048653.g002

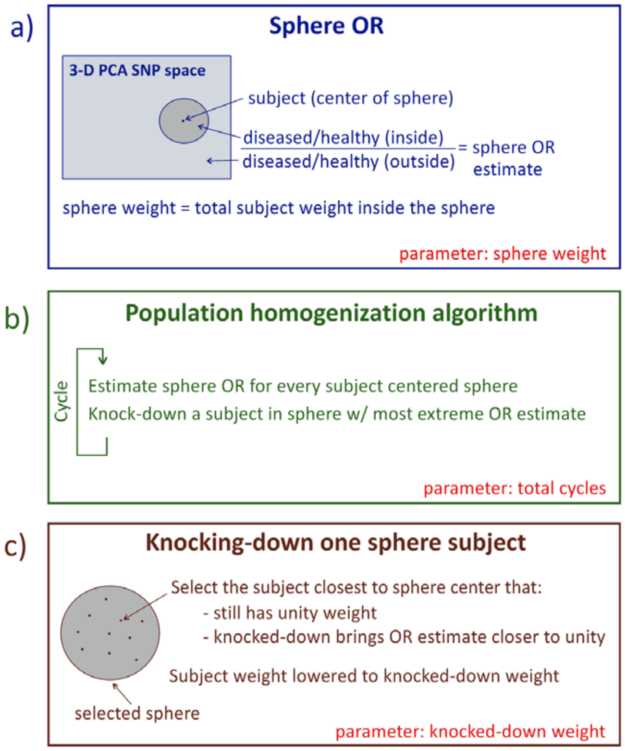
a)
Sphere OR
3-D PCA SNP space
subject (center of sphere)
diseased/healthy (inside)
diseased/healthy (outside)
= sphere OR estimate
sphere weight = total subject weight inside the sphere
parameter: sphere weight
b)
Population homogenization algorithm
Cycle
Estimate sphere OR for every subject centered sphere
Knock-down a subject in sphere w/ most extreme OR estimate
parameter: total cycles
c)
Knocking-down one sphere subject
Select the subject closest to sphere center that:
- still has unity weight
- knocked-down brings OR estimate closer to unity
Subject weight lowered to knocked-down weight
selected sphere
parameter: knocked-down weight

Figure 3. The population homogenization algorithm. a) Euclidean-distance spheres on the space defined by the first three principal component projections of the 75000 SNP space (3PC-space). The spheres are characterized by the total subject statistical weight in their interior, rather than by their spatial radius. The sphere OR is defined as the OR for an idealized genotype present in every subject in the interior of that sphere and in no subject outside of it. ORs are estimated for all spheres centered on a cohort subject and of a predefined sphere weight. **b**) Overview of the homogenization algorithm. **c**) Within the sphere with the most extreme OR estimate, one subject is selected to have its statistical weight knocked-down to a predefined knocked-down weight.
doi:10.1371/journal.pone.0048653.g003

g_1 - number of A subjects sampled,

g_2 - number of ~A subjects sampled and

g_3 - sampling partiality towards diseased subjects versus towards healthy subjects.

Now, discarding A samples is akin to scaling by a less than unity common factor the entries a and c. Similarly, discarding healthy samples is akin to scaling by a less than unity common factor the entries b and d.

Let us call p_3 *matched cohorts* to those cohorts that could plausibly originate from a sampling of diseased and healthy subjects under the same partiality p_3. Similarly, let us call g_3 *matched cohorts* to those cohorts that could plausibly originate from a sampling of A and ~A subjects under the same partiality g_3. Only p_3 or g_3 matched cohorts are suitable to being merged. Fortunately, the above-mentioned transformations can be utilized to produce matched cohorts. For instance, the transformations can always be used to equalize the corresponding marginal ratios in the two contingency tables (i.e., making the net A/~A ratio equal on the two tables and making the net diseased/healthy ratio equal on the two tables). Doing so produces cohorts that are both p_3 and g_3 matched.

In a GWAS, there is typically a vast number of genotype based population splits to consider and each requires its own cohort-matching transformation. Furthermore, population structure generally has a more complex continuous character than the two distinct populations assumed so far. Thus, computational-time wise cohort-matching for every genetic trait may not be feasible. Instead, we suggest equalizing the net diseased/healthy subject marginal ratio on the two cohorts. This can be achieved by appropriately discarding diseased or healthy samples (without regards to their genotype). Note that this equalization is independent of the genotype split under consideration. Now, if the OR is unity in the two populations, then the matching of the diseased/healthy subject ratio guarantees that the cohorts are g_3 matched. This follows from an unity OR indicating that the diseased/healthy ratio in the population is identical for both genotypes under consideration. Note that the method is robust: in the limit where the individual population ORs approach unity and the cohort marginal diseased/healthy ratios approach each other, the cohorts approach being g_3 matched. The limitation of the approach is that, for non-unity ORs, matching the diseased/healthy ratio no longer guarantees that the cohorts are g_3 matched. This is clear in the Figure 1-c example, where the diseased/healthy ratio is unity in both cohorts and yet, the merged cohort produces a false-negative outcome. A careful examination of the example will show that the sampling is partial towards healthy subjects in population X relative to the sampling in population Y. Let us call *phenotype marginal matching* to this procedure of matching the diseased/healthy ratios in the cohorts. In summary, phenotype marginal matching reduces the false-positives (Figure 1-b case) but not the false-negatives (Figure 1-c case). However, the computational faster time, due to a single homogenization serving for all genotype pairs, makes it a practical choice. Finally, note that throughout we have assumed that the OR is the same in both populations. However, the OR could also be population dependent. As an example, consider a SNP that is a risk factor only under the net genetic background or dietary habits characteristic of a particular population. Such cases always benefit from an individualized analysis of the population cohorts, as their merging unavoidably produces some form of averaged OR estimate.

A commonly applied test to the cohort contingency table is the Fisher exact test [38]. Importantly, note that the test is valid on a contingency table from merged cohorts that were phenotype marginal matched, since the unity OR assumption of the test also guarantees that the cohorts are g_3 matched.

Continuous population structure case

So far we have assumed two well-defined, distinct populations. We now describe our phenotype marginal matching approach for the practical case where the population structure has a continuous character. The method is presented alongside its application to homogenization of the Hamza et al. Parkinson's GWAS cohort [36,39]. This cohort of 2000 Parkinson's disease patients and 1986 controls was genotyped at on the order of 10^6 SNPs. Subjects were recruited in North America and reported European ancestry. We first identified the population structure as Hamza et al. [36], via principal component analysis [16,36]. The relative overall location of individuals in SNP space (Euclidean distance wise) reflects the cohort population structure. However, there are regions of the genome with a particular high density of genotyped, highly correlated SNPs. This results in a few such groups of SNPs overwhelmingly determining the clustering of subjects in SNP space. Following standard procedure [36], this was avoided by utilizing only a reduced set of 75 000 SNPs, selected for their comparatively lower correlation. The SNPs were selected using the program Plink [32], with a 50-SNP sliding window, shifting 5 SNPs with each move and recursively removing SNPs with

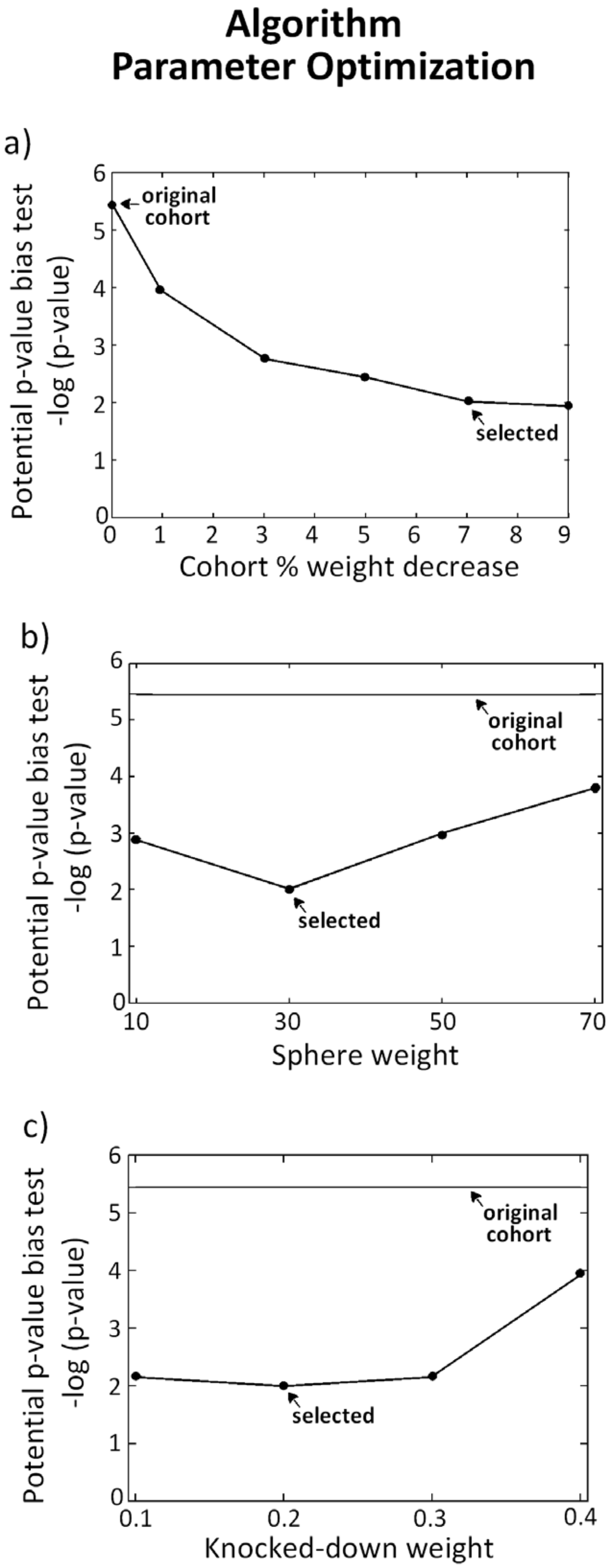

Figure 4. Algorithm parameter optimization. The algorithm was applied to homogenize the Hamza et al. cohort. The algorithm parameters were selected based on the value of the potential p-value bias test (see Table 1 and main text). Graph **a**) shows potential p-value bias versus % of cohort weight decrease. A 7% cohort weight decrease was selected, as weight decreases beyond this value produced only a marginal further decline in the potential p-value bias. **b**) The sphere weight parameter was set to 30. **c**) A 0.2 knocked-down weight was selected. In each of these graphs, the two parameters not represented are held at their selected values.
doi:10.1371/journal.pone.0048653.g004

$r^2 > 0.1$. As discusssed in Hamza et al., the population structure is well captured by the first three principal components of this 75 000 SNP space. In here, our objective is therefore to homogenize the diseased/healthy subject ratio (phenotype marginal match) throughout the subspace defined by these first three principal components (henceforth 3PC-space). The uneven distribution of healthy and Parkinson's subjects in this space is visually perceptible in Figure 2 (third principal component not shown).

The homogenization is performed by lowering the statistical weight ('statistical' subsumed henceforth) of select subjects from their original unity weight, to a smaller, but nonzero, predefined *knocked-down weight*. Henceforth, let all spatial references be with respect to 3PC-space and based on Euclidean distances. We define the *sphere OR* as the OR for an idealized genotype present nowhere other than on every subject within that spatial sphere (Figure 3-a). Such a genotype thus marks exactly the population present within the sphere. The homogenization algorithm will employ spheres centered on cohort subjects. In lieu of sphere spatial radius, a more relevant parameter is *sphere cohort weight*, the total weight of the cohort subjects inside the sphere. This is due to the pertinent statistical comparison, as described below, being between spheres with the same cohort weight, not between spheres with the same spatial size per se.

We now describe the iterative homogenization algorithm (Figure 3-b,c). The sphere OR for every sphere with a predetermined cohort weight and centered around a cohort subject is estimated (note on weight discreteness: in practice, utilize the smallest sphere centered on the subject that equals or surpasses the desired weight). The sphere with the most extreme OR estimate is selected (given the symmetry in the OR definition, by extreme it is meant furthest from unity, in the sense whereby OR = 3 and OR = 1/3 are equally distanced from unity). Next, the weight of one cohort subject within this selected sphere is knocked-down in order to bring the sphere estimated OR closer to unity. The subject closest to the sphere center, still with unity weight and of the appropriate phenotype (diseased if the OR imbalance is due to too many diseased subjects within the sphere, healthy in the reverse case) is selected to have its weight knocked-down to the

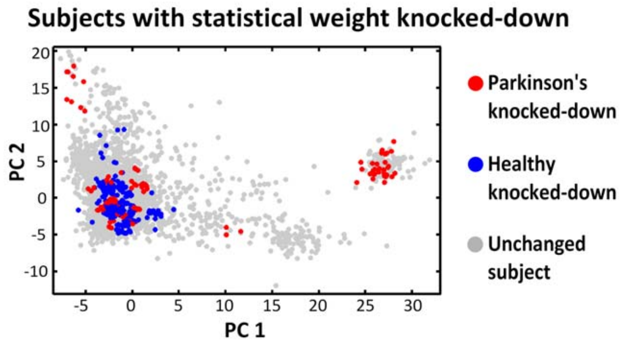

Figure 5. Statistically knocked-down individuals in the cohort homogenization process. The local homogenization of the Parkinson's/healthy marginal ratio (phenotype marginal matching) is performed by knocking-down the statistical weight of selected subjects. Overall, 120 of the 2000 Parkinson's patients and 230 of the 1986 healthy controls had their statistical weight knocked-down from 1 to 0.2. This represented a 7% net weight decrease in the cohort.
doi:10.1371/journal.pone.0048653.g005

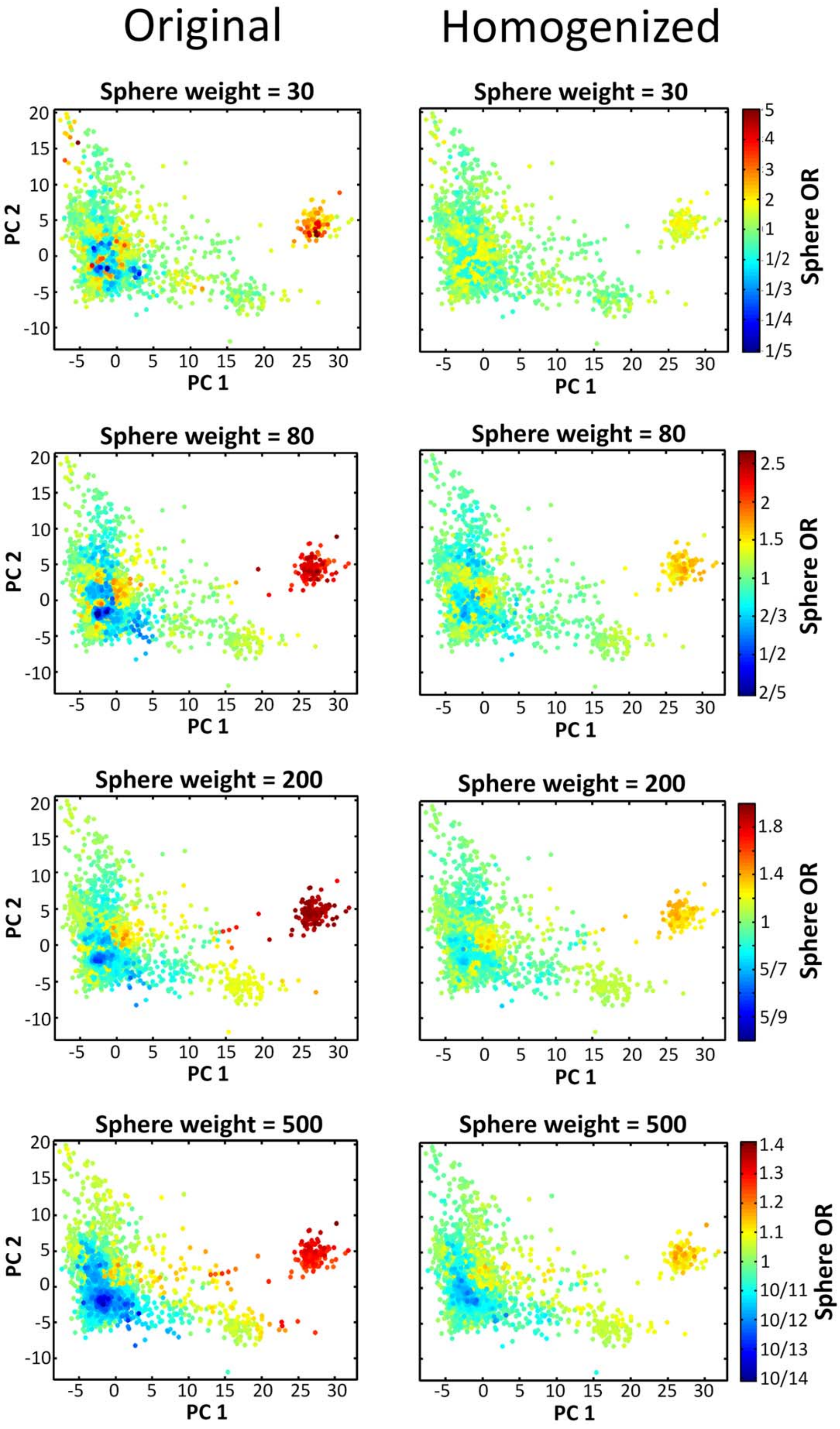
Original
Homogenized
Sphere weight = 30
Sphere weight = 30
PC 2
PC 1
Sphere OR
5
4
3
2
1
1/2
1/3
1/4
1/5
Sphere weight = 80
Sphere weight = 80
2.5
2
1.5
1
2/3
1/2
2/5
Sphere weight = 200
Sphere weight = 200
1.8
1.4
1
5/7
5/9
Sphere weight = 500
Sphere weight = 500
1.4
1.3
1.2
1.1
1
10/11
10/12
10/13
10/14
20
15
10
5
0
-5
-10
-5
0
5
10
15
20
25
30

Figure 6. Estimated sphere ORs in the original and homogenized cohorts. Each subject is colored based on the estimated OR of a sphere centered on that subject and with the sphere weight stated on the top of the respective color map (see Figure 3-a for sphere OR definition). Sphere OR color scale reflects the symmetry of the OR definition. Note sphere OR color scale amplitude is smaller on color maps for larger sphere weights.
doi:10.1371/journal.pone.0048653.g006

predefined knocked-down weight. The cycle is then repeated, with every sphere OR being re-calculated, a new sphere with the most extreme OR being highlighted and, within it, a new subject being selected for knock-down. This cycle is performed a determined number of times (set by the desired final cohort weight) upon which the procedure is concluded.

The algorithm requires selecting values for three parameters: total cycles (or final decrease in cohort weight), sphere weight and knocked-down weight (Figure 3). To select these, we introduce the *potential p-value bias test*. This is a measure of the OR estimate Fisher p-value bias potential due to population structure in the cohort. Consider again a hypothetical genotype present on every individual within a given spatial sphere and on none outside that sphere. This constitutes an extreme case of a genotype associated with a specific population (in this case, the population within the sphere). This genotype produces a false-positive OR estimate if the diseased/healthy ratio is not identical inside and outside the sphere. The smallest p-value that can be generated by such hypothetical genotypes thus provide a measure of the potential for p-value bias introduced by population structure. We numerically implement this test by considering all spheres centered on cohort subjects, hence effectively covering every relevant sphere location in 3PC-space. Similarly, a spectrum of sphere weights is considered. The value of the potential p-value bias test is then the smallest of the p-values thus generated.

The algorithm was applied to the Hamza et al. cohort. Based on the potential p-value bias test, the parameters were set to cohort weight decrease = 7%, sphere weight = 30 and knocked-down weight = 0.2 (Figure 4). We underline the limited sensitivity of the potential p-value bias test results to the selected parameter values, indicating the algorithm is robust in this regard (Figure 4). Table 1 breaks down the potential p-value bias test over the range of sphere weights tested. From the original Hamza et al. cohort to the homogenized cohort under the above selected parameters, the extreme p-value found over the spectrum of tested locations and weights decreased by four orders of magnitude, from $4*10^{-6}$ in the original cohort, to $1*10^{-2}$ in the homogenized cohort. Figure 5 shows the knocked-down individuals in the homogenization process. Figure 6 compares estimated sphere ORs on the original and homogenized cohorts. The comparison is performed across a range of sphere weights.

Simulation study

Utilizing synthetic data, we compared the effect that homogenizing a cohort has on the GWAS typical true-positive, false-positive and false-negative association. We still relied on the real data Hamza et al cohort. However, the following synthetic disease/healthy distribution was now utilized (Figure 7-a): Diseased/healthy labels were assigned to individuals with a 0.5/0.5 chance in the population at large, except for individuals in the rectangular box region of Figure 7-a, where a 0.6 vs. 0.4 diseased vs. healthy chance was used when assigning the labels. In reality, this boxed region delimits approximately cohort individuals of Irish or English ancestry (based on self-reported ancestry, see Figure 2 in Valente et al. [39]). The created imbalance in the diseased/healthy distribution is apparent in Figure 7-b. The homogenizing algorithm was applied to this dataset (algorithm parameters: cohort weight decrease = 9%, sphere weight = 30 and knocked-down weight = 0.2). Figure 7-c shows the individuals knocked-down in the homogenization process. Comparison of Figure 7-d with Figure 7-b shows the effect of the homogenization procedure. Next, we created a number of synthetic genotypes to observe the effect of the homogenization procedure on the typical false-positive and false-negative association induced by population structure, as well as on the standard true-positive association (Figure 7-e).

To generate typically true-positive associations (under no population structure correction), genotype labels were assigned to all cohort individuals as follows:

- For a diseased-phenotype individual there was a 0.57 probability of a ~A genotype assignment, while for a healthy-phenotype individual this probability was 0.50 (yielding an OR = 1.32). The process was repeated five hundred times, thus generating five hundred such genotypes. The red circles in Figure 7-e correspond to three such genotypes.
- For a diseased-phenotype individual there was a 0.15 probability of a ~A genotype assignment, while for a healthy-phenotype individual this probability was 0.10 (yielding an OR = 1.59). The process was repeated five hundred times, thus generating five hundred such genotypes. The cyan circles in Figure 7-e correspond to three such genotypes.

Table 1. The potential p-value bias test.

Cohort weight	**Sphere weight**	**25**	**50**	**100**	**200**	**400**	**800**	**1600**	*Potential p-value bias test*
3986	Original cohort	$2*10^{-4}$	$3*10^{-4}$	$4*10^{-6}$	$4*10^{-6}$	$3*10^{-4}$	$9*10^{-5}$	$7*10^{-5}$	*$4*10^{-6}$*
3706	Homogenized cohort	$2*10^{-2}$	$1*10^{-2}$	$2*10^{-2}$	$2*10^{-2}$	$3*10^{-2}$	$3*10^{-2}$	$3*10^{-2}$	*$1*10^{-2}$*

Consider the Fisher exact test p-value associated with an OR estimate for a hypothetical genotype present nowhere other than on every subject within a given spatial sphere. Each table entry reports the smallest such p-value, over all the subject-centered spatial spheres with the stated weight. These p-values provide a comparative measure of the potential for p-value bias due to population structure, before and after cohort homogenization. The value of the potential p-value bias test is the smallest of these p-values across all the tested sphere weights (rightmost column). Values shown pertain to application of the test to the original and homogenized (under the Figure 4 selected parameter values) Hamza et al. cohort. Note: for the homogenized cohort, the Fisher test is performed on the contingency table entries rounded to the nearest integer.
doi:10.1371/journal.pone.0048653.t001

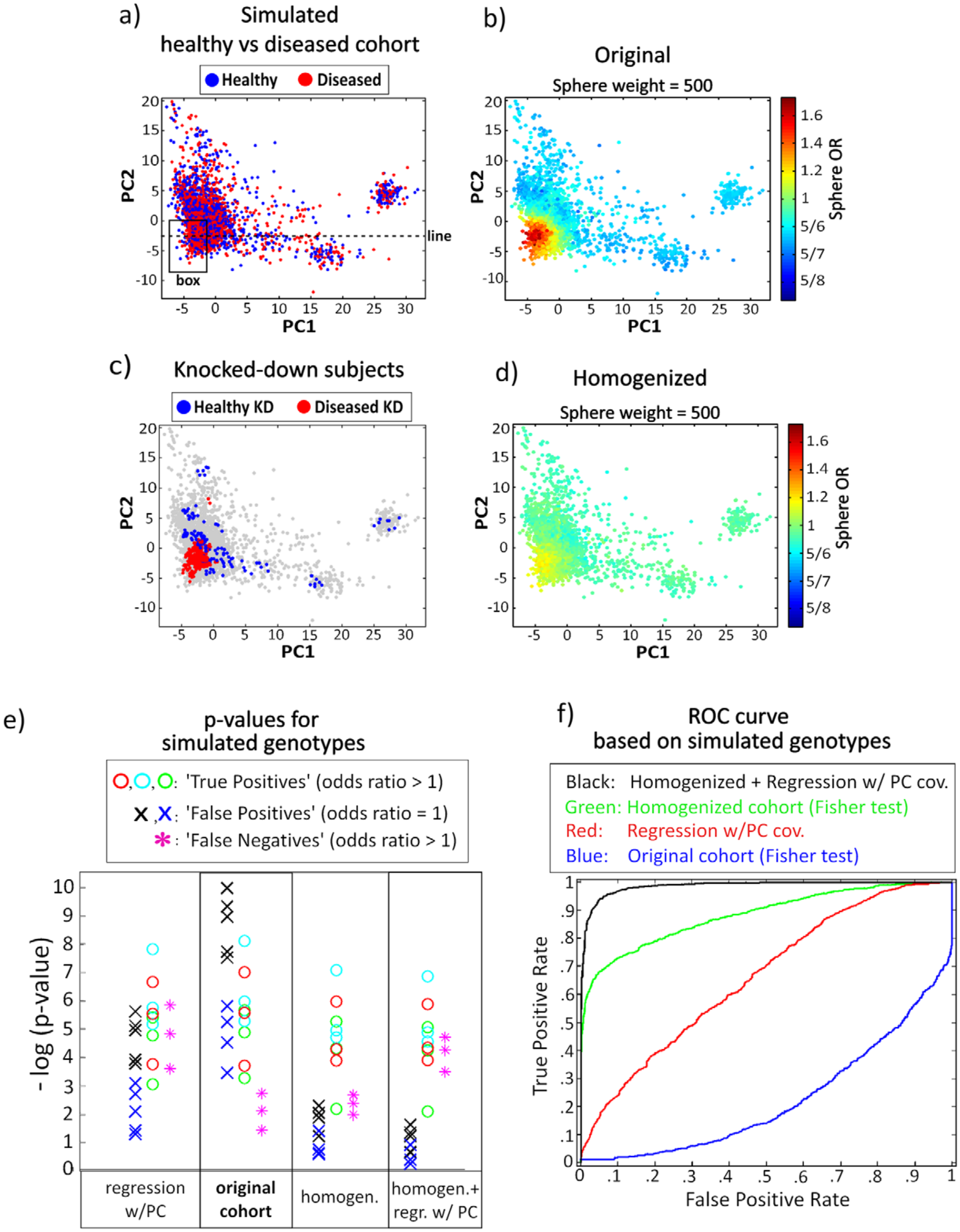

Figure 7. Simulation study. a) Synthetic diseased-healthy phenotype labels were assigned to the Hamza et al. cohort subjects. **b**) Visualization of the sphere ORs (at sphere weight = 500) in this synthetic dataset shows the imbalance in the diseased-healthy distribution. **c**) Individuals with their statistical weight knocked-down in the cohort homogenization process (final homogenized cohort weight reduction = 9%). **d**) Visualization of the sphere ORs (at sphere weight = 500), after the homogenization process. **e**) Synthetic genotype labels were assigned to individuals to produce the typical true-positive, false-positive and false-negative association (under no population structure correction). The -log(p-value) for these synthetic genotype-phenotype associations were calculated, based on four different approaches: Fisher exact test using the original cohort; Fisher exact test using the homogenized cohort; logistic regression with the 3 PCs as covariates; and logistic regression using the homogenized cohort weights and with the 3 PCs as covariates. **f**) Receiver Operating Characteristic (ROC) curves for the four different approaches. See Simulation Study Section for details.

doi:10.1371/journal.pone.0048653.g007

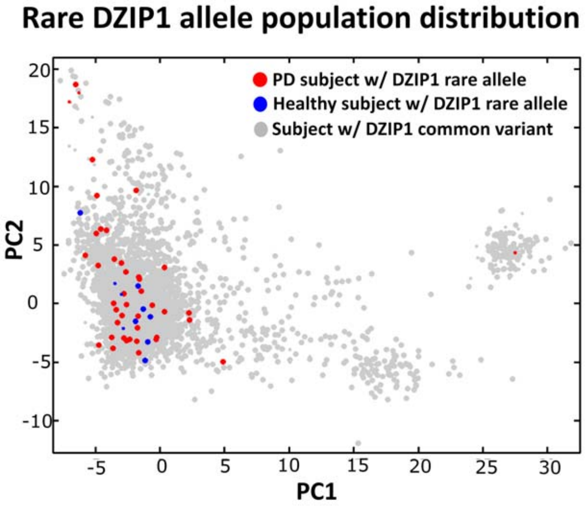

Figure 8. Interpreting the DZIP1 allele distribution in the context of population structure. Based on an analysis of the homogenized Hamza et al. cohort, a rare SNP mutation in the gene DZIP1 that increases the susceptibility to Parkinson's disease was found [39]. In this cohort, individuals carrying a copy of this rare DZIP1 allele are predominantly in a region where, as per the color maps of Figure 5, there is still a slight relative overabundance of healthy subjects (indicated by the blue hue in Figure 5). Thus, we conclude that the remaining population heterogeneity in the homogenized cohort has a propensity to diminish, rather than to inflate, the estimated OR of the DZIP1 SNP. Above, smaller dot sizes indicate subjects with knocked-down weight. Figure reproduced from Valente et al. [39].
doi:10.1371/journal.pone.0048653.g008

– For a diseased-phenotype individual there was a 0.06 probability of a ~A genotype assignment, while for a healthy-phenotype individual this probability was 0.03 (yielding an OR = 2.06). The process was repeated five hunderd times, thus generating five hundred such genotypes. The green circles in Figure 7-e correspond to three such genotypes.

The above genotypes, spanning a range of minor allele frequencies, are homogeneously distributed across the population. Therefore, evaluation of their association with phenotype status is not distorted by population structure effects.

To generate typically false-positive associations (under no population structure correction), genotypes labels were assigned to all cohort individuals as follows:

– For an individual inside the rectangular box (Figure 7-a), there was a 0.90 probability of a ~A genotype assignment, regardless of phenotype status. For an individual outside the rectangular box, there was a 0.10 probability of a ~A genotype assignment, regardless of phenotype status. The process was repeated five hunderd times, thus generating five hundred such genotypes. The black crosses in Figure 7-e correspond to four such genotypes.
– For an individual inside the rectangular box, there was a 0.80 probability of a ~A genotype assignment, regardless of phenotype status. For an individual outside the rectangular box, there was a 0.20 probability of a ~A genotype assignment, regardless of phenotype status. The process was repeated five hunderd times, thus generating five hundred such genotypes. The blue crosses in Figure 7-e correspond to four such genotypes.

The above genotypes have no effect on the phenotype status (OR = 1). However, the ~A genotype is highly common in the rectangular box population, by comparison with its presence in the population at large. In accordance with the Figure 1-b example, the concurrent higher prevalence of both the ~A genotype and the diseased phenotype in the rectangular box population region generates a false apparent genotype-phenotype association.

To generate typically false-negative associations (under no population structure correction), genotypes labels were assigned to all cohort individuals as follows:

– For an individual below the dashed line on Figure 7-a ($PC2 < -2.68$) and a disease phenotype, there was a 0.08 probability of a ~A genotype assignment. For an individual below the dashed line on Figure 7-a and a healthy phenotype, there was a 0.05 probability of a ~A genotype assignment. For an individual above the dashed line on Figure 7-a and a disease phenotype, there was a 0.95 probability of a ~A genotype assignment. For an individual above the dashed line on Figure 7-a and a healthy phenotype, there was a 0.92 probability of a ~A genotype assignment. The process was repeated five hunderd times, thus generating five hundred such genotypes. The magenta stars in Figure 7-e correspond to three such genotypes. In reality, the dashed line roughly delimits from above individuals of Irish or Italian ancestry (based on self-reported ancestry, see Figure 2 in Valente et al. [39]).

For the above genotypes, OR = 1.65 in both the above and the below the dashed line populations. However, the ~A genotype is highly prevalent in the population above the dashed line, while the A genotype is highly prevalent in the population below the dashed line. This is enough to conceal the genotype-phenotype association. Note that the diseased/healthy ratio is roughly identical in the above and below the dashed line populations. As previously illustrated by the Figure 1-c example, an homogeneous diseased/healthy phenotype ratio across the population does not prevent these false-negatives cases.

Figure 7-e shows -log(p-value) for some of these synthetic genotypes, based on four different approaches:

– Original cohort: using the original cohort and the Fisher exact test.
– Regression w/PCs: using the original cohort and logistic regression with the first 3 PCs as covariates to account for population structure effects.
– Homogenized: using the homogenized cohort and the Fisher exact test.
– Homogenized + regression w/PCs: using the homogenized cohort and logistic regression with the first 3 PCs as covariates (the knocked-down homogenized cohort weights are employed in the logistic regression).

The true-positives, false-positives and false-negatives behave as expected when not accounting for population structure. Both logistic regression and the cohort homogenization method are able to decrease -log(p-value) of the false-positives. However, the reduction is significantly more pronounced using the cohort homogenization method. The average -log(p-value) of true-positives is not significantly decreased by the 9% smaller size of the homogenized cohort. The decrease is comparable to the one observed under logistic regression, being in this latter case attributable to the addition of the 3 PCs as covariates. As expected, false-negatives are not rescued by the homogenization method. Their -log(p-value) does significantly increased under logistic regression. Finally, the overall best results were obtained by combining the homogenization and logistic regression methods: The -log(p-value) of false-positives was reduced the most, true-positives were again only marginally affected, and false-negatives were still rescued by the logistic regression.

Using the entire set of simulated genotypes, we built Receiver Operating Characteristic (ROC) curves for each of the above four approaches, showing the attainable combination of true positive and false positive rates, depending on the selected p-value significance level. The ROC curves confirm the observations based on the few genotypes individually analyzed in Figure 7-e, namely with the ROC curves being ordered from best to worst as 1) Homogenized + regression with PCs, 2) Homogenized, 3) Regression with PCs and 4) Original cohort.

Results and Discussion

In this article we i) propose cohort homogenization as a strategy for minimizing false-positives in a GWAS, ii) present an algorithm for homogenizing a cohort and iii) introduce a measure for assessing p-value bias potential due to population structure. Note that although the homogenization method reduces false-positives, it is not able to guarantee a given false-positive rate. When applied to the Hamza et al. Parkinson's cohort, the method significantly reduced its p-value bias potential (Table 1). The intuitive character of the approach is also advantageous. For instance, in a separate work, this homogenized Parkinson's cohort was analyzed under the hypothesis-rich framework [39,40]. The main finding was a rare SNP mutation in the gene DZIP1 that increases the susceptibility to Parkinson's disease. As shown in Figure 8, this rare DZIP1 mutant occurs predominantly in a region where, as per the color maps of Figure 6, there is still a slight relative overabundance of healthy subjects (marked by the blue end of the spectrum). Therefore it follows that the remaining population heterogeneity in the homogenized cohort has a propensity to diminish, rather than to inflate, the estimated OR of the DZIP1 SNP.

The homogenization approach introduced in this article opens a diversity of interesting research directions for future exploration. Firstly, the simulation work we performed highlights that the homogenization approach could potentially be profitably used in conjunction with other methods, namely logistic regression with the principal components as covariates. We are currently working on a thorough examination of such combined approaches. Secondly, our focus on homogenizing only the diseased/healthy marginal phenotype meant that false negatives induced by population structure are not addressed by the approach. However, they would be, if *both* phenotype and genotype marginals were homogenized. This is computationally demanding, due to requiring a new homogenization for every additional genotype pair being tested. However, perhaps variations on the approach or on the algorithm can make this full marginal matching computationally tractable and practical to apply to at least a considerable selection of genotype pairs. Thirdly, we have not shown optimality of the homogenization algorithm regarding the trade-off between achieved homogenization and imposed cohort weight reduction. Thus, the development of alternative more efficient homogenization algorithms is another open research problem.

Author Contributions

Conceived and designed the experiments: AV YG AS. Performed the experiments: AV JZ JS. Analyzed the data: AV YG AS. Wrote the paper: AV YG AS.

References

1. Manolio TA (2010) Genomewide association studies and assessment of the risk of disease. New Engl J Med 363 (2): 166–176.
2. Hardy J, Singleton A (2009) Genomewide Association Studies and Human Disease. N Engl J Med 360: 1759–1768.
3. Witte JS (2010) Genome-wide association studies and beyond. Annu Rev Publ Health 31: 9–20.
4. Marchini J, Cardon LR, Phillips MS, Donnelly P (2004) The effects of human population structure on large genetic association studies. Nat Genet 36: 512–517.
5. Teo YY (2008) Common statistical issues in genome-wide association studies: a review on power, data quality control, genotype calling and population structure. Curr Opin Lipidol 19: 133–143.
6. Sillanpää MJ (2011) Overview of techniques to account for confounding due to population stratification and cryptic relatedness in genomic data association analyses. Heredity 106: 511–519.
7. Astle W, Balding DJ (2009) Population structure and cryptic relatedness in genetic association studies. Stat Sci 24 (4): 451–471.
8. Edwards TL, Gao X (2012) Methods for Detecting and Correcting for Population Stratification. Curr Protoc Hum Genet 73: 1.22.1–1.22.14.
9. Devlin B, Roeder K (1999) Genomic control for association studies. Biometrics 55 (4): 997–1004.
10. Bacanu SA, Devlin B, Roeder K (2000) The Power of Genomic Control. Am J Hum Genet 66 (6): 1933–1944.
11. Zheng G, Freidlin B, Gastwirth JL (2006) Robust genomic control for association studies. Am J Hum Genet 78: 350–356.
12. Dadd T, Weale ME, Lewis CM (2009) A critical evaluation of genomic control methods for genetic association studies. Genet Epidemiol 33: 290–298.
13. Wang K (2009) Testing for genetic association in the presence of population stratification in genome-wide association studies. Genet Epidemiol 33: 637–645.
14. Patterson N, Price A, Reich D (2006) Population structure and eigenanalysis. PLoS Genet 2: e190.
15. Zhu X, Li S, Cooper RS, Elston RC (2008) A unified association analysis approach for family and unrelated samples correcting for stratification. Am J Hum Genet 82: 352–365.
16. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, et al. (2006) Principal component analysis corrects for stratification in genome-wide association studies. Nat Genet 38 (8): 904–909.
17. Setakis E, Stirnadel H, Balding DJ (2006) Logistic regression protects against population structure in genetic association studies. Genome Res 16: 290–296.
18. Kimmel G, Jordan MI, Halperin E, Shamir R, Karp RM (2007) A randomization test for controlling population stratification in whole-genome association studies. Am J Hum Gent 81: 895–905.
19. Li Q, Wacholder S, Hunter DJ, Hoover EN, Chanock S, et al (2009) Genetic background comparison using distance-based regression, with applications in population stratification evaluation and adjustment. Genet Epidemiol 33 (5): 432–441.
20. Li Q, Yu K (2008) Improved correction for population stratification in genome-wide association studies by identifying hidden population structures. Genet Epidemiol 32 (3): 215–226.
21. Novembre J, Stephens M (2008) Interpreting principal component analyses of spatial population genetic variation. Nat Genet 40 (5): 646–649.
22. Zhu C, Yu J (2009) Nonmetric Multidimensional Scaling Corrects for Population Structure in Association Mapping With Different Sample Types. Genetics 182 (3): 875–888.
23. Miclaus K, Wolfinger R, Czika W (2009). SNP selection and multidimensional scaling to quantify population structure. Genet Epidemiol 33 (6): 488–496.
24. Li M, Reilly MP, Rader DJ, Wang LS (2010) Correcting population stratification in genetic association studies using a phylogenetic approach. Bioinformatics 26 (6): 798–806.
25. Lee AB, Luca D, Klei L, Devlin B, Roeder K (2009) Discovering genetic ancestry using spectral graph theory. Genet Epidemiol 34 (1): 51–59.
26. Zhang J, Niyogi P, McPeek MS (2009) Laplacian eigenfunctions learn population structure. PLoS ONE 4: e7928.
27. Pritchard JK, Stephens M, Rosenberg NA, Donnelly P (2000) Association mapping in structured populations. Am J Hum Genet 67: 170–181.
28. Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. Genome Res 19: 1655–1664.
29. Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155 (2): 945–959.
30. Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. Genetics 164: 1567–1587.
31. Hoggart CJ, Parra EJ, Shriver MD, Bonilla C, Kittles RA, et al (2003) Control of confounding of genetic associations in stratified populations. Am J Hum Genet 72: 1492–1504.

32. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, et al (2007) PLINK: a toolset for whole-genome association and population-based linkage analysis. American Journal of Human Genetics 81 (3): 559–575.
33. Yu J, Pressoir G, Briggs WH, Bi IV, Yamasaki M, et al (2005) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet 38: 203–208.
34. Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, et al (2008) Efficient control of population structure in model organism association mapping. Genetics 178: p. 1709–1723.
35. Wu C (2011) A comparison of association methods correcting for population stratification in case–control studies. Ann Hum Genet 75 (3): 418–427.
36. Hamza TH, Zabetian CP, Tenesa A, Laederach A, Montimurro J, et al (2010) Common genetic variation in the HLA region is associated with late-onset sporadic Parkinson's disease. Nat Genet 42: 781–785.
37. Edwards AWF (1963) The measure of association in a 2×2 table. J Roy Statist Soc Ser A 126 (1): 109–114.
38. Fisher RA (1954) Statistical Methods for Research Workers. Oliver and Boyd.
39. Valente AXCN, Shin JH, Sarkar A, Gao Y (2012) Rare coding SNP in DZIP1 gene associated with late-onset sporadic Parkinson's disease. Sci Rep 2: art. no. 256.
40. Valente AXCN (2011) Prediction in the hypothesis-rich regime. In: Valente AXCN, editor. Science and engineering in high-throughput biology including a theory on Parkinson's disease. Lulu Books. 23–38.