PLOS ONE

RESEARCH ARTICLE

The impact of a fine-scale population stratification on rare variant association test results

Elodie Persyn[1,2], **Richard Redon**[1,3☯], **Lise Bellanger**[4☯*], **Christian Dina**[1☯*]

1 INSERM, CNRS, UNIV Nantes, l'institut du thorax, Nantes, France, **2** Department of Medical and Molecular Genetics, King's College London, London, United Kingdom, **3** CHU Nantes, l'institut du thorax, Nantes, France, **4** Laboratoire de Mathématiques Jean Leray, Nantes, France

☯ These authors contributed equally to this work.
* lise.bellanger@univ-nantes.fr (LB); christian.dina@univ-nantes.fr (CD)

Citation: Persyn E, Redon R, Bellanger L, Dina C (2018) The impact of a fine-scale population stratification on rare variant association test results. PLoS ONE 13(12): e0207677. https://doi.org/10.1371/journal.pone.0207677

Editor: David Meyre, McMaster University, CANADA

Received: July 27, 2018

Accepted: November 5, 2018

Published: December 6, 2018

Data Availability Statement: Data simulation is described in the Methods section of the manuscript.

Funding: This work was supported by a grant from The French Regional Council of Pays de la Loire (RFI VaCaRMe: Recherche, Formation et Innovation, Vaincre les maladies Cardiovaculaires, Respiratoires et Métaboliques) (CD, LB, and RR and funded Dr. Elodie Persyn). The website for this project is available at the following URL address http://www.vacarme-project.org/. Agence Nationale de la Recherche (ANR-15-CE17-0008-01 to RR):

Abstract

Population stratification is a well-known confounding factor in both common and rare variant association analyses. Rare variants tend to be more geographically clustered than common variants, because of their more recent origin. However, it is not yet clear if population stratification at a very fine scale (neighboring administrative regions within a country) would lead to statistical bias in rare variant analyses. As the inclusion of convenience controls from external studies is indeed a common procedure, in order to increase the power to detect genetic associations, this problem is important. We studied through simulation the impact of a fine scale population structure on different rare variant association strategies, assessing type I error and power. We showed that principal component analysis (PCA) based methods of adjustment for population stratification adequately corrected type I error inflation at the largest geographical scales, but not at finest scales. We also showed in our simulations that adding controls obviously increased power, but at a considerably lower level when controls were drawn from another population.

Introduction

Association studies have identified many common variants associated with a wide spectrum of diseases. With the advances in sequencing technologies, it became possible to perform case-control studies on rare variants, which may also play an important role in disease susceptibility. Due to their low frequency in the populations, the statistical analysis presents challenges as these variants must be tested by groups (commonly defined as genes). Many statistical methods have been developed or adapted to test the association between a group of rare variants and a disease status [1]. Each of these methods assumes a different statistical hypothesis and is able to detect association signals depending on the underlying disease mechanisms which are unknown. A first category of tests, called burden tests, consists in aggregating rare variant counts across the gene to sum up the genetic information per individual [2–7]. Another main category of tests, called variance-component tests (or joint tests) [7–10], assumes a heterogeneous group of variants with different effect sizes, and tests the variance of genetic effects.

Grant Recipient was Richard Redon and funded scientist was Dr Elodie Persyn.

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

Population stratification, i.e. the observation of differences in allele frequencies between populations, due to different ancestries, has been shown to be a confounding factor that could lead to many false positive results in rare variant association studies [11–18]. Rare variants are thought to be more recent than common variants. Therefore they are more likely to be geographically localized and contribute to a fine scale [2]genetic structure. The impact of such a fine genetic structure on association tests results is still poorly investigated. For instance, two studies [16,18] showed that, without population structure adjustment, the analysis of rare variants in simulated European populations could lead to inflation of gene-based test results. However, it has already been shown that a population structure could be identified at even lower geographical scales such as the Western French population, using common variants [19]. Therefore, it is important to know if such geographical structure could lead to false positive results. Indeed, in order to increase the ability to detect disease genes and reduce costs while sequencing more cases as a priority, the use of controls from reference databases is very common in genetic epidemiology studies. However, these controls may be from a different population ancestry than cases and thus create problems of confusion.

Many rare variant association methods exist and their performance varies depending on the genetic scenario. These methods are also influenced differently by population stratification. A higher inflation in variance-component tests (SKAT [9] and C-alpha [8]) than burden tests has been reported by [18], when comparing simulated European populations. Another study [15], also showed that the C-alpha test [8] presents a higher inflation than a burden test, with population stratification. To go further, [18] showed that depending on the joint allelic distribution in two populations, both burden tests or variance-component tests could show higher inflation.

Standard correction approaches, such as adjusting a model for principal components (PCs) representing the genetic structure among individuals, are able to reduce the inflation of rare variant association results but may also fail in specific scenarios [12]. As rare variants may present a different geographical pattern than common variants, the frequency of variants to include in the estimation of the components has also been discussed [12,17]. From these studies, the computation of PCs from common variants is more? efficient, compared to PCs from rare variants, to adjust for a world-wide population structure.

In this paper, we aim to answer two main questions on various rare variant association test strategies: (1) the impact of additional controls with different levels of fine-scale population structure, and (2) the efficiency of PCA correction methods. The impact on association results was assessed in terms of type I error and power through simulations under different genetic scenarios. In our simulations we considered two populations. Cases were drawn from one population and controls were drawn from both populations in varying proportions depending on the simulation scenario. Different geographical levels were set by varying the migration rate between the two populations. We explored these simulated genetic scenarios, to better relate it to real stratification patterns. These analyses would help to better select external controls when performing association analyses on rare variants; or at least warn against potential bias.

Materials and methods

Notations

Rare variant association methods test the association between the disease status of N individuals and their genotype information for a group of P rare variants. Let $\mathbf{X}$ be the matrix of genotypes with $X_{ij} \in \{0,1,2\}$ the count of minor alleles for the i-th individual and j-th variant. Let $\boldsymbol{Y}$ be the vector of phenotypes with $Y_i = 1$ if the i-th individual is a case, otherwise $Y_i = 0$. In equation notations, 0 and 1 superscripts denote respectively unaffected and affected persons.

Rare variant association tests

We compared different association strategies, commonly used in rare variant association studies, such as burden tests and variance component tests. We also took into account the widely used KBAC test [20] which considers multi-locus genotypes. Finally, we applied what we call "position tests", DoEstRare [21] and PODKAT [22], which are recent approaches taking into account rare variant positions. The different association tests compared in this work are presented in the **Table 1**. For each category of tests, we either selected the most used or/and the most recently developed.

Burden tests. A first category, called burden tests, consists in aggregating rare allele counts across the gene. We used the following burden tests: cohort allelic sum test (CAST) [2,23], sum test (Sum), weighted sum test (wSum) [3], and the adaptive sum test (aSum) [6]. These methods compute a genetic score per individual and test the association between this score with the disease status. We used different burden tests, with subtle hypothesis differences that may impact their statistical behavior with population stratification. In order to better compare them, we formulated burden tests with a logistic regression model:

$$logit(P(Y_i = 1)) = \alpha_0 + \beta S_i$$

with S_i a genetic score for the individual i which is a function of rare allele counts X_{ij},$j \in \{1,\ldots, P\}$, and β the regression coefficient. The CAST test is said to be a collapsing strategy as the genetic score indicates if an individual carries at least one rare mutation. This score is:

$$S_{CAST_i} = I\left(\sum_{j=1}^{P} X_{ij} \geq 1\right)$$

with $I(.)$ the indicator function. For tests other than CAST, the genetic score can be written as a weighted sum of allele counts:

$$S_i = \sum_{j=1}^{P} w_j X_{ij}$$

with w_j the weight for the variant j. In the Sum test, each rare variant presents the same weight $w_j = 1, j \in \{1,\ldots,P\}$. In the wSum test, weights $w_j = \frac{1}{\sqrt{N*\widehat{MAF}_j^{\,0}*\left(1-\widehat{MAF}_j^{\,0}\right)}}, j \in \{1,\ldots,P\}$ are a function of $\widehat{MAF}_j^{\,0} = \frac{\sum_{i=1}^{N^0} X_{ij}+1}{2N^0+2}$, the estimated MAF in controls. In the aSum test, $w_j = 1$ if the j-th variant is considered deleterious, and $w_j = -1$ if the j-th variant is considered protective. A single-marker is previously applied to know if a rare variant is classified as protective.

Table 1. Rare variant association tests under comparison.

Category	Description of the strategy	Methods
Burden tests	Computation of a genetic score per individual.	CAST [2], Sum test, wSum [3], aSum [6]
KBAC test	Comparison of multi-locus genotypes counts between cases and controls	KBAC [20]
Variance-component tests	Test of the variance of genetic effects.	SKAT [9], SKAT-O [10]
Position tests	Incorporation of rare variant positions in the test statistic.	PODKAT [22], DoEstRare [21]

https://doi.org/10.1371/journal.pone.0207677.t001

For all these burden tests, we test the null hypothesis H_0:$\beta = 0$ with a score test [24,25]. The score statistic U and its variance V are:

$$U = \boldsymbol{S}'(\boldsymbol{Y} - \widehat{\boldsymbol{\mu}})$$

$$V = \frac{1}{N-1}(\boldsymbol{Y} - \widehat{\boldsymbol{\mu}})'(\boldsymbol{Y} - \widehat{\boldsymbol{\mu}}) * (\boldsymbol{S} - \bar{\boldsymbol{S}})'(\boldsymbol{S} - \bar{\boldsymbol{S}})$$

$\widehat{\boldsymbol{\mu}} = logit^{-1}(\alpha_0).\mathbf{1}_N$ is the vector of estimates under the null hypothesis (expression may change with covariates in the model) with $\mathbf{1_N} = (1,\ldots,1)'$, and $\bar{\boldsymbol{S}} = \left(\frac{1}{N}\sum_{i=1}^{N} S_i\right).\mathbf{1}_N$ the vector of average scores. The test statistic is $Q = \frac{U^2}{V}$.

Variance-component association tests. Variance-component tests consider the variance of genetic effects. These tests have been developed to better identify association signals in a context of variants with different effect sizes and directions in the same gene.

We used sequence kernel association tests (SKAT) [9,10], which are based on the following logistic regression model:

$$logit(P(Y_i = 1)) = \alpha_0 + \sum_{j=1}^{P} \beta_j X_{ij}$$

with β_j,$j \in \{1,\ldots,P\}$, the regression coefficients for the genetic effects. This model is a linear mixed-effects model with random genetic effects β_j which follow an arbitrary distribution of mean 0 and variance $w_j^2\tau$. The null hypothesis H_0:$\beta_j = 0$,$j \in \{1,\ldots,P\}$ is then equivalent to H_0: $\tau = 0$. Each variant is weighted by w_j to better discriminate causal from neutral variants.

We also used the optimal version of SKAT, called SKAT-O [10], varying the correlation between genetic effects. SKAT assumes that there is no correlation between genetic effects, while SKAT-O aims to identify the optimal correlation between genetic effects. The test statistic for a given correlation parameter ρ test is:

$$Q_\rho = (\boldsymbol{Y} - \widehat{\boldsymbol{\mu}})'\mathbf{X}\mathbf{W}\mathbf{R}_\boldsymbol{\rho}\mathbf{W}\mathbf{X}'(\boldsymbol{Y} - \widehat{\boldsymbol{\mu}})$$

with $\widehat{\boldsymbol{\mu}} = logit^{-1}(\alpha_0).\mathbf{1}_{\boldsymbol{N}}$, the vector of estimates under the null hypothesis as previously defined (expression may change with covariates in the model); $\mathbf{W}$ = diag(w_j,j $\in \{1,\ldots,P\}$), the weight matrix; and the correlation matrix between genetic effects $\mathbf{R_\rho} = (1-\rho)\mathbf{I_P} + \rho\mathbf{1_P}\mathbf{1_P}'$ with $\mathbf{I_P}$ identity matrix of order p. For the SKAT test, $\rho = 0$. For the SKAT-O statistic, ρ varies between 0 and 1 with a bin of 0.1. The distribution of each test statistic under the null hypothesis is approximated and is described by [26].

KBAC test. The kernel-based adaptive cluster (KBAC) test [20] aims to better discriminate causal multi-site genotypes from noise with the use of adaptive weights. Multi-site genotypes are the vectors of individual genotypes $\boldsymbol{X_i}$,$i \in \{1,\ldots,N\}$. Let assume $L + 1$ distinct multi-genotype sites $\boldsymbol{X_0}$,$\boldsymbol{X_1}$,$\ldots$,$\boldsymbol{X_l}$,$\ldots$,$\boldsymbol{X_L}$ we observe in the dataset $\mathbf{X}$. The KBAC statistic compares the proportions of these multi-site genotypes in cases and controls, and is equal to

$$KBAC = \left(\sum_{l=0}^{L} w_l\left(\frac{n_l^1}{N^1} - \frac{n_l^0}{N^0}\right)\right)^2$$

with w_l the weight for l-th multi-site genotype; n_l^1 and n_l^0 the observed counts of the l-th multi-site genotype in cases and controls. Weights are computed adaptively with the choice of a kernel function; greater weights are attributed to genotypes that are enriched in cases. In this paper, we use the hypergeometric kernel, which is the most often used as it is suitable for small

to moderate sample sizes. The weight w_l for the l-th multi-site genotype is then defined as

$$w_l = P\left(N_l^1 \geq n_l^1\right) = \sum_{k=0}^{n_l^A} \frac{\binom{N_l}{k}\binom{N - N_l}{N^1 - k}}{\binom{N}{N^1}}$$

Position tests. We also investigated the impact of population structure on two statistical strategies which extend the rare variant tests to situations where the position of polymorphisms has an impact in the disease. These two tests are DoEstRare, developed in our team and which tests variant density in cases and controls and PODKAT which extends SKAT through integration of a position distance matrix.

The test PODKAT [22] is an extension of the test SKAT. The test statistic is similar to the SKAT statistic:

$$Q_{PODKAT} = (\boldsymbol{Y} - \widehat{\boldsymbol{\mu}})'\mathbf{XWAA'W'X'}(\boldsymbol{Y} - \widehat{\boldsymbol{\mu}})$$

with a position-dependent matrix **A** measuring proximities between variants. The proximity measure between variants j and j' is:

$$\mathrm{A}_{j,j'} = \max\left(1 - \frac{1}{w} d_{j,j'}, 0\right)$$

with $d_{j,j'}$, the physical distance between variants j and j'; and the parameter w is called maximal radius of tolerance, by default its value is 1,000 bp.

The significance of the PODKAT statistic is based on the approximation of the distribution under the null hypothesis with the Davies' method [27]. The integration of position-dependent matrix is a strategy that has also been used by [28,29].

The DoEstRare test [21] compares both position distributions on the gene and overall allele frequencies between cases and controls. Let $\widehat{f}^1$ and $\widehat{f}^0$, be the kernel density estimators of position distributions; $\widehat{p}^1$ and $\widehat{p}^0$, the weighted average allele frequencies in cases and controls. The test statistic is:

$$STAT = \int_1^{Lg} |\widehat{p}^1 * \widehat{f}^1(pos) - \widehat{p}^0 * \widehat{f}^0(pos)| dpos$$

with Lg the length of the tested region in bp.

Concretely, this statistic corresponds to the area between the density function curves, each multiplied by allele frequencies in cases and controls.

The position density functions are estimated with a Gaussian kernel [30]. The $\widehat{p}^1$ and $\widehat{p}^0$ frequencies are

$$\widehat{p}^1 = \frac{1}{P}\sum_{j=1}^{P} \frac{w_j}{\sum_{j=1}^{P} w_j} \frac{m_j^1}{2N^1} \qquad \widehat{p}^0 = \frac{1}{P}\sum_{j=1}^{P} \frac{w_j}{\sum_{j=1}^{P} w_j} \frac{m_j^0}{2N^0}$$

with w_j the weight for the j-th variant. DoEstRare use a similar ponderation approach than the KBAC test. It assumes that the count of rare mutations in cases M_j^1 follows, under the null hypothesis, a binomial distribution $\mathcal{B}(2N^1, \widehat{MAF}_j^0)$ with $\widehat{MAF}_j^0$ the estimate of the minor allele frequency in controls. The weight w_j is the probability to present less than the observed count

m_j^1.

$$w_j = P(M_j^1 \leq m_j^1) = \sum_{k=0}^{m_j^1} \binom{2N^1}{k} (\widehat{MAF}_j^{\,0})^k (1 - \widehat{MAF}_j^{\,0})^{2N^1 - k}$$

Other statistical tests [31–33,28,29] take into account position information in the test statistic but they were not taken into account in our comparison.

Weighting systems. Some of the statistical association tests we presented above use a weighting system to better discriminate causal from neutral variants. wSum and SKAT tests use of a function of the MAF estimate in the dataset. Because allele frequencies differ between populations, the computation of the MAF in the dataset will depend on the geographical origin of cases and controls. As we wanted to assess the impact of different weighting systems based on MAF estimation in the context of population stratification, we considered three weighting systems for both Sum and SKAT tests:

1. an unweighted version with $w_j = 1, j \in \{1,\ldots,P\}$ (labeled Sum, SKAT and SKATO);
2. weights following a beta distribution $w_j = Beta(\widehat{MAF}_j, a_1 = 1, a_2 = 25)$, as proposed by [9], with $\widehat{MAF}_j$ the MAF estimation in both cases and controls (labeled wSum_betaMAFtot, wSKAT_betaMAFtot and wSKATO_beta_MATtot);
3. weights $w_j = \frac{1}{\sqrt{N*\widehat{MAF}_j*\left(1-\widehat{MAF}_j\right)}}$, as proposed by [3], but with the MAF estimation in both cases and controls (labeled wSum_MAFtot, wSKAT_MAFtot, wSKATO_MAFtot).

   We also considered, for the Sum test, a fourth weighting system $w_j = \frac{1}{\sqrt{N*\widehat{MAF}_j^{\,0}*\left(1-\widehat{MAF}_j^{\,0}\right)}}$, from [3], with $\widehat{MAF}_j^{\,0}$ the MAF estimation in controls (labeled wSum_MAFctrl). This weighting system cannot be used in the SKAT R Package, as it would introduce a bias without a proper permutation procedure.

Simulation workflow

We used the program *cosi* [34], based on a coalescent model, to simulate genetic data. The coalescent model's demographical parameters are derived from the bestfit model which was found to best explain present worldwide genetic diversity [34]. We added two European sub-populations A and B, which split from the original European population 80 generations ago (**Fig 1A**). Each sub-population includes 10,000 haplotypes (5,000 individuals). The geographical proximity between these two populations is linked to the migration rate parameter. This migration rate parameter varies between 0, 0.001, 0.01, 0.025, 0.05 and 0.1, to investigate the impact of population structure at different geographical scales.

From the two sub-populations, we sampled 1,000 cases and 1,000 controls according different genetic scenarios (**Fig 1B**). In all scenarios, cases are from the same population A, and controls from populations A and B. The proportion of controls coming from the population B varies between 25%, 50%, 75% and 100%. We also simulated a scenario without population stratification with all cases and controls from population A.

These scenarios varying the population structure are simulated under the H_0 (no genetic association) and H_1 (genetic association) hypotheses in order to assess respectively type I error and power.

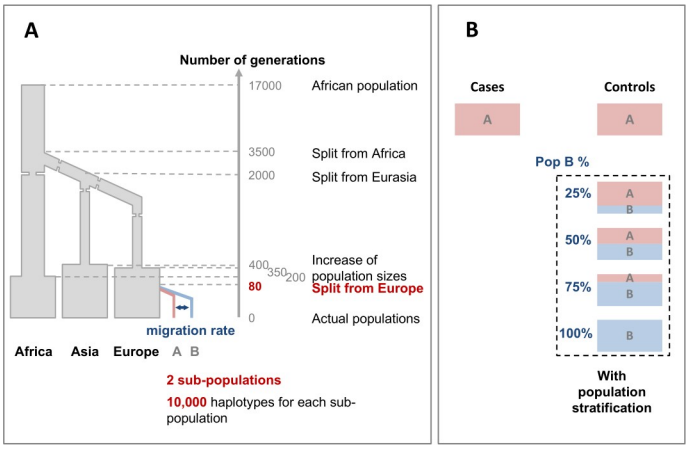

Fig 1. Simulation of a population stratification in case-control data. A. Simulated demographical model with the *cosi* program. Modifications from the bestfit model designed by [34] are indicated in red. The migration parameter we varied is in dark blue. B. Geographical origin of cases and controls. In a first scenario cases and controls are from the same population A. In scenarios with a population stratification, the percentage of controls from population B we varied is indicated in dark blue.

https://doi.org/10.1371/journal.pone.0207677.g001

Under the null hypothesis, cases and controls are sampled from the two populations A and B without regarding the genetic information.

Under the alternative hypothesis, the disease status is sampled with the probability $P(Y_i = 1|\boldsymbol{X_i})$ determined by the following logistic regression model:

$$logit(P(Y_i = 1|\boldsymbol{X_i})) = \alpha_0 + \boldsymbol{\beta}'\boldsymbol{X_i}$$

with $\boldsymbol{\beta}'$ the vector of genetic regression coefficients. We considered in this model, that 50% of rare variants are deleterious with an odd ratio $OR = 1.5$. The individual sampling process is repeated until obtaining 1,000 cases and 1,000 controls with a given percentage of controls from population B.

Rare variants are defined according to the MAF in the total population A (10,000 haplotypes) as cases are sampled from this population A.

We performed 10,000 replicates to assess type I errors and 1,000 replicates to assess power.

Exploratory analysis of simulated data

We explored our simulated data to assess how close are the populations A and B, by (1) performing PCA, and (2) computing the fixation index F_{ST} [35] that measures the population differentiation due to genetic structure. We applied these methods to the genetic data concatenating all common genetic variants from the 10,000 gene replicates. Genetic data include pruned common variants with a $MAF \geq 5\%$ and $r^2 < 0.2$ in the total population A, for a sampling of 1,000 individuals in each population A and B.

We performed the PCA with the *smartpca* program from EIGENSOFT package version 6.1.4 [36,37]. We computed the fixation index F_{ST} between populations A and B using the R function *calc_wcFst_spop_pairs* from the github repository https://github.com/ekfchan/evachan.org-Rscripts, which implements the method of [35].

Rare variant association analysis and population stratification correction

Analyses of the association of rare variants are carried out using the statistical tests previously described, on data simulated according to different scenarios mentioned above. Significance is

assessed with an adaptive permutation procedure [38] for all statistical tests except SKATs and PODKAT, which rely on a approximated distribution of the test statistic under the null hypothesis. The adaptive permutation procedure enables to save computational time in comparison with the standard permutation procedure. Parameters are the significance threshold $\alpha = 0.01$ and the precision value $c = 0.2$. Power and type I error have been assessed considering $\alpha = 0.05$.

Statistical tests are performed with and without correcting for population structure. The most common correction method is the integration of covariates, such as PCs computed from the genetic data, in a logistic regression model [37]. As it is described before in this paper, most of statistical tests, with the exception of KBAC and DoEstRare, are presented under the form of a logistic regression model and can be adjusted with this method.

$$\mathcal{M}_1 \; logit(P(Y_i = 1|\boldsymbol{X}_i, \boldsymbol{Z}_i)) = \alpha_0 + \boldsymbol{\alpha}'\boldsymbol{Z}_i + \boldsymbol{\beta}'\boldsymbol{X}_i$$

with $\boldsymbol{Z_i}$ the vector of covariates for the i-th individual. We label this method "PCA model correction".

Note: As it is mentioned by authors of KBAC, this test can be adjusted with covariates in a logistic regression model but is not implemented in the R package.

Another correction method, proposed by [39], using a permutation procedure taking into account covariates, can be applied to a larger range of association tests. First, the null model is adjusted for covariates:

$$\mathcal{M}_0 \; logit(P(Y_i = 1|, \boldsymbol{Z}_i)) = \alpha_0 + \boldsymbol{\alpha}'\boldsymbol{Z}_i$$

Then, from this model, odds of disease conditional on covariates $\theta_i = \exp(P(Y_i = 1|,\boldsymbol{Z}_i)), i \in \{1,\ldots,N\}$ are computed. Finally, individuals are sampled according to a Fisher's noncentral hypergeometric distribution with disease odds θ_i as individual weights, to obtain permutated data with similar population stratification. We label this method "PCA permutation correction". For significance assessment, we adapted the adaptive permutation procedure we used [38], to take into account PCs, according to the description of [39], on all statistical tests including SKATs and PODKAT. Because the permutation procedure is running very slowly for SKATs and PODKAT, we assessed only type I errors with this correction method for a number of 5,000 replicates instead of 10,000.

These two correction methods rely on covariates, reflecting the geographical origin of individuals. We considered the two first components of the PCA performed with *smartpca* program from EIGENSOFT package version 6.1.4 [36,37]. PCA was performed on pruned common variants ($MAF \geq 5\%$ and $r^2 < 0.2$ in the total population A), for each case-control sampling according to the genetic scenario. We also performed type I error analyses with 5 PCs and 10 PCs on a subset of 5,000 replicates instead of 10,000 to assess the consequences of the number of PCs setting.

Results and discussion

Simulation of a fine geographical scale population structure

We simulated genetic information for 10,000 artificial genes for two populations A and B varying the migration rate parameter between 0, 0.001, 0.01, 0.025, 0.05 and 0.1. Numbers of SNV and allele frequency distributions for 1,000 individuals are almost the same in populations A and B (**Table A** and **Table B** in **S1 Table**). Depending on the migration rate, the number of SNV varies between 702,332 and 846,053 (**Table A** in **S1 Table**), and the percentage of rare variants with a MAF< = 1% varies between 53.5% and 61.7% (**Table B** in **S1 Table**) in population A. In order to assess the geographical differentiation level of populations A and B for each

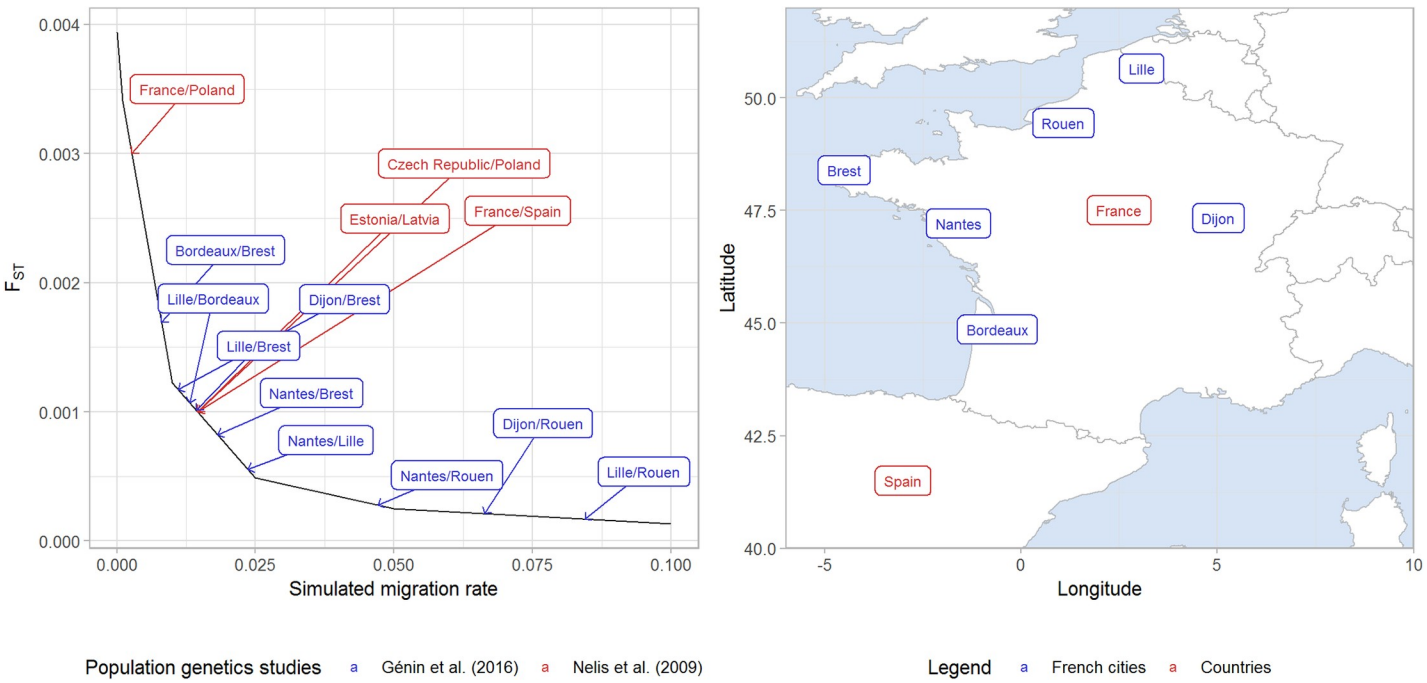

Fig 2. Comparison of F_{ST} values between simulations and real population genetic studies. FST values obtained from simulations are plotted in function of the migration rate parameter. Pairwise FST values from two real population genetic studies, [40] and [41], are added respectively in red and blue.

https://doi.org/10.1371/journal.pone.0207677.g002

migration rate, we computed pairwise F_{ST} values and performed PCA on pruned common variants. In the Fig 2 (see **Table C** in S1 Table for F_{ST} values), we related these values to estimates from [40] and from [41], which are respectively genetic studies on European populations and the French population. In [40], neighbor European countries present F_{ST} values close to 0.001 (France/Spain: 0.001; Czech Republic/Poland: 0.001; Estonia/Latvia: 0.001). Far European countries show higher F_{ST} values (France/Latvia: 0.008, Latvia/Spain: 0.01). Our simulations with a migration rate of 0.01 conduct to a F_{ST} value close to 0.001 ($F_{ST}(0.01)$ = 0.001226), which would correspond to a situation with neighbor countries. Simulations with migration rates of 0 and 0.001 correspond to situations with more distant countries. Scenarios that correspond to fine-scale population structure, are scenarios with a migration rate higher than 0.01. In [41], results are from the French Exome project (FREX)., in which controls are recruited from 6 French centers (Bordeaux, Brest, Dijon, Lille, Nantes, and Rouen cities) and had their exome sequenced. The scenario with a migration rate of 0.001 shows a F_{ST} value also close to the estimates for distant French regions. Scenarios with a migration rate of 0.025, 0.05 or 0.1 would correspond to situations with geographically close French regions, which means a very fine-scale population structure. Of course, this "inference" is based on only the F_{ST} indicator, and other parameters should be taken into account such as allele frequency distributions.

We also perform a PCA analysis on simulated data, to see whether it is possible to distinguish the populations A and B from individual genetic profiles. The representations of the two first components (S1 Fig) show that populations are clearly distinct with a migration rate$\leq$0.01. The overlaps of genetic profiles between populations A and B are respectively very small, moderate, and nearly total for scenarios with a migration rate equal to 0.025, 0.05 and 0.1. In the PCA analysis of FREX data from Génin et al. (2016), not presented here, French sub-populations showed moderate to high overlaps. However, we cannot compare their PCA

results with ours as sampling sizes are very different: around 100 individuals per French subpopulation while 1,000 individuals per simulated population.

Inflation of type I errors and efficiency of correction methods

Cases and controls, for the rare variant association analysis, are sampled from the two simulated populations A and B. We conducted simulations under H_0 to assess type I errors. For 2,000 individuals from the population A, the average number of analyzed SNV, across the 10,000 replicates, varies between 30.8 (se: 7.3) and 32.5 (se: 7.5) depending on the migration rate (respectively 0.01 and 0.1). Without population stratification (case and all controls coming from population A), type I errors at significance level $\alpha = 5\%$ are correct with the exception of CAST which seems conservative (**S2 Table** and **S2 Fig**).

We then analyzed simulated datasets with mixed ancestry in controls, first without any correction method, in order to estimate the increase of type I errors due to population stratification. Type I errors show inflation even in the presence of a very fine scale population structure, i.e. with a migration rate of 0.05 or 0.1 (**Fig 3**). However, this inflation remains negligible when 25% or less of controls are ascertained from population B, which means that cases and controls are still quite homogenous in terms of geographical origin.

We note obvious type I error differences between rare variant association tests. CAST and Sum burden tests are the least sensitive to population stratification in almost all scenarios, while the wSum_MAFctrl and aSum burden tests, which are just variations of the previous test, present high values of inflation. Variance-component tests and KBAC present high type I errors compared to other tests. Finally, tests integrating variant positions present intermediate values. These results are consistent with observations made by [18] and [15] where variance-component tests also presented higher inflation than some burden tests. From all these observations, it is understandable that variance-component tests and the aSum test present higher type I error values as they are adapted to test a group of rare variants with opposed effects (protective/risk variants). Indeed, both populations A and B include population-specific rare variants, in the sense that some rare variants are more frequent in one population. This creates a situation where rare variants seem to display opposed effects, when cases and controls are sampled from two populations. However, it has been discussed by [18] the possibility of burden tests being more sensitive than variance-component tests if global count of rare variants differ between populations, and may be influenced by demographical conditions such as population growth. We did not consider different growth rates in population A and B, also explaining why variance-component tests are more sensitive.

Weighting allele contribution according to MAF is common practice in rare variants tests, based on the assumption that the probability of functional, usually harmful, effect is increased for very rare alleles. The weighted derived version of Sum test, wSum_MAFctrl, which uses MAF the estimation in controls, presents a very high inflation of type I errors. When MAF is estimated in both controls and cases, wSum_MAFtot and wSum_betaMAFtot also present a higher inflation, but a lot lower in comparison with wSum_MAFctrl. The test wSum_MAFtot seems to present a slight increase of type I error compared to wSum_betaMAFtot. These two tests differ in the MAF weighting function, which is standard deviation or beta distribution. By using a beta distribution of parameters 1 and 25, in wSum_betaMAFtot, weights decrease less rapidly with the MAF increase and may thus buffer the effect of wrong MAF weighting. This difference induced by using beta weights is less clear for variance-component tests, wSKAT_MAFtot and wSKATO_betaMAFtot, as it is only visible with high proportions of controls from population B and low migration rates, i.e. with the highest population stratifications (see **S2 Table**). We conclude that the use of a weighting system based on the computation of MAFs, from the dataset, may provide high

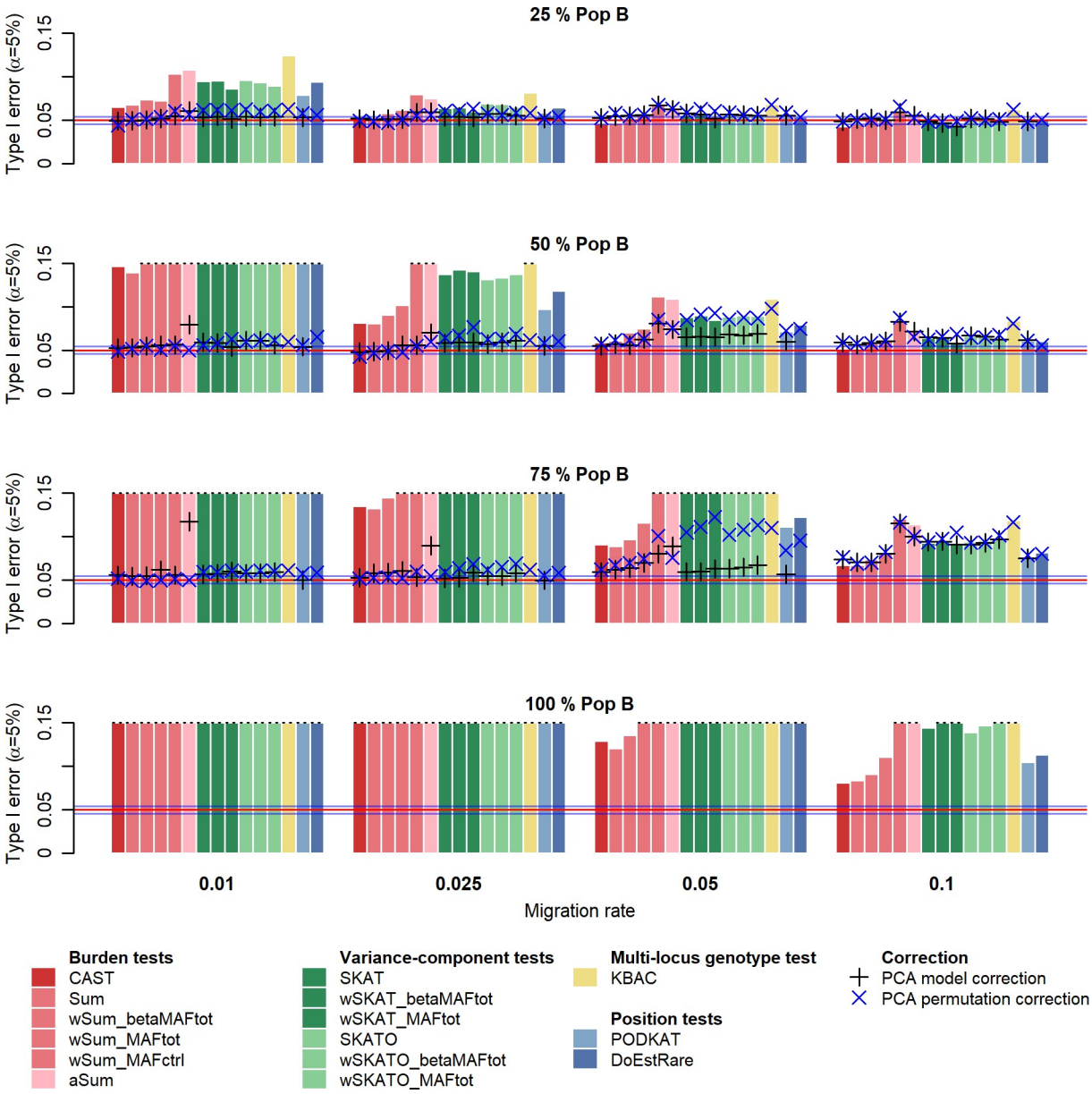

Fig 3. Type I errors at level α = 5% with a population stratification. Bars represent type I errors without correction for population stratification. The red line corresponds to α = 5% and blue lines correspond to 95% confidence interval. Confidence interval is computed assuming that the number of false positives follows a binomial distribution with parameters 10,000 and 0.05. Correction methods are performed with the first two PCs.

https://doi.org/10.1371/journal.pone.0207677.g003

numbers of false positives when allele frequencies differ between cases and controls due to stratification. We can also extend this interpretation to the KBAC and the DoEstRare tests, as weighting systems depend on observed allele counts in controls.

In practice, a correction method can and should be applied to avoid statistical biases from the population stratification. We applied two correction methods, the "PCA model correction" and the "PCA permutation correction", integrating the information from the two first PCs on common variants from the 10,000 replicates, to reduce the inflation of type I errors. In the scenarios with 100% of controls from population B, these two correction methods were not

applied as PCs may totally explain the phenotype, not allowing testing for genetic effects. These correction methods perform well in scenarios with the most differentiated populations (migration rate of 0.01); but they do not totally reduce the inflation caused by a very fine scale population structure (migration rate of 0.05 or 0.1). The "PCA permutation correction" is advantageous with association tests which are not based on a logistic regression model, but seems to be less efficient in the scenario with a migration rate of 0.05. It suggests that the "PCA permutation correction method" gradually lose more in efficiency than the "PCA model correction method" with finer population stratification.

We integrated the information from the two first PCs, as it was sufficient to correct with the largest population structures. As in practice the choice of the number of PCs to integrate in the models may be arbitrary and adapted depending on the scenario, we here performed analyses by integrating 2, 5 or 10 PCs (**S2 Table** and **S3 Fig**). By increasing the number of PCs in the models, the type I error inflation does not decrease in the context of very fine scale population structures (migration rate of 0.05 or 0.1), hence our choice to keep only two PCs. We even note an increase of type I errors in some scenarios for burden tests, whose significance is assessed by an adaptive permutation procedure. It is maybe due to an over-adjustment of the logistic regression model with a high number of PCs [42].

Our analyses were conducted considering a small sample size of 1,000 cases and 1,000 controls. By increasing the sample size, type I errors may increase greatly (see S4 Fig). A fine population structure in the data would have even more impact in large sample sequencing studies.

Type I errors might also differ considering other demographical models resulting in very different site frequency spectrum [43–45]. In our simulations, we used a derived model from [34], in which the population expansion events are instantaneous. However, this model is very simplistic and does not reflect demographical growth observations.

In this study, type I errors have been assessed considering $\alpha = 0.05$. We realize that the actual significance threshold being used in genetic studies is much lower after correcting multiple testing ($\alpha = 2.5e\text{-}6$ when considering 20,000 genes in an exome-sequencing study). Test statistics may behave differently at very low significance levels but because a large subset of our tests is using time-consuming permutations, whose number could not be increased within the scope of the present study.

Balance between type I error and power under population stratification

The purpose of using external controls, i.e. from population B in our scenarios, is to increase the power to detect deleterious genes in association studies when it is impossible to sequence larger sample size of controls. We observed previously that stratification correction methods enable to reduce statistical biases with the largest population structures but fail with the finest ones. For this reason, we aim to assess the impact of the "PCA model correction" on the power of rare variant association tests. We simulated a simple scenario under H_1, with half of rare variants being deleterious with an OR of 1.5. In the **Fig 4**, with the most structured simulated populations, i.e. with high percentages of controls from population B and low migration rates, we can observe an obvious loss of power for every statistical test, compared to the analysis without population stratification (see **S3 Table** for power values). In these scenarios, deleterious variants are likely to be under population structure, i.e. presenting different allele frequencies in populations A and B. The adjustment for PCs also removes, from rare variant association tests, a portion of the disease genetic component. For the finest population structures, we note a small increase of power with the use of controls from population B, due to statistical biases not fully (or at all) corrected.

We also estimated powers after removing from the analysis controls from the population B. For example, in the scenario with 75% controls from population B, only 25% controls are

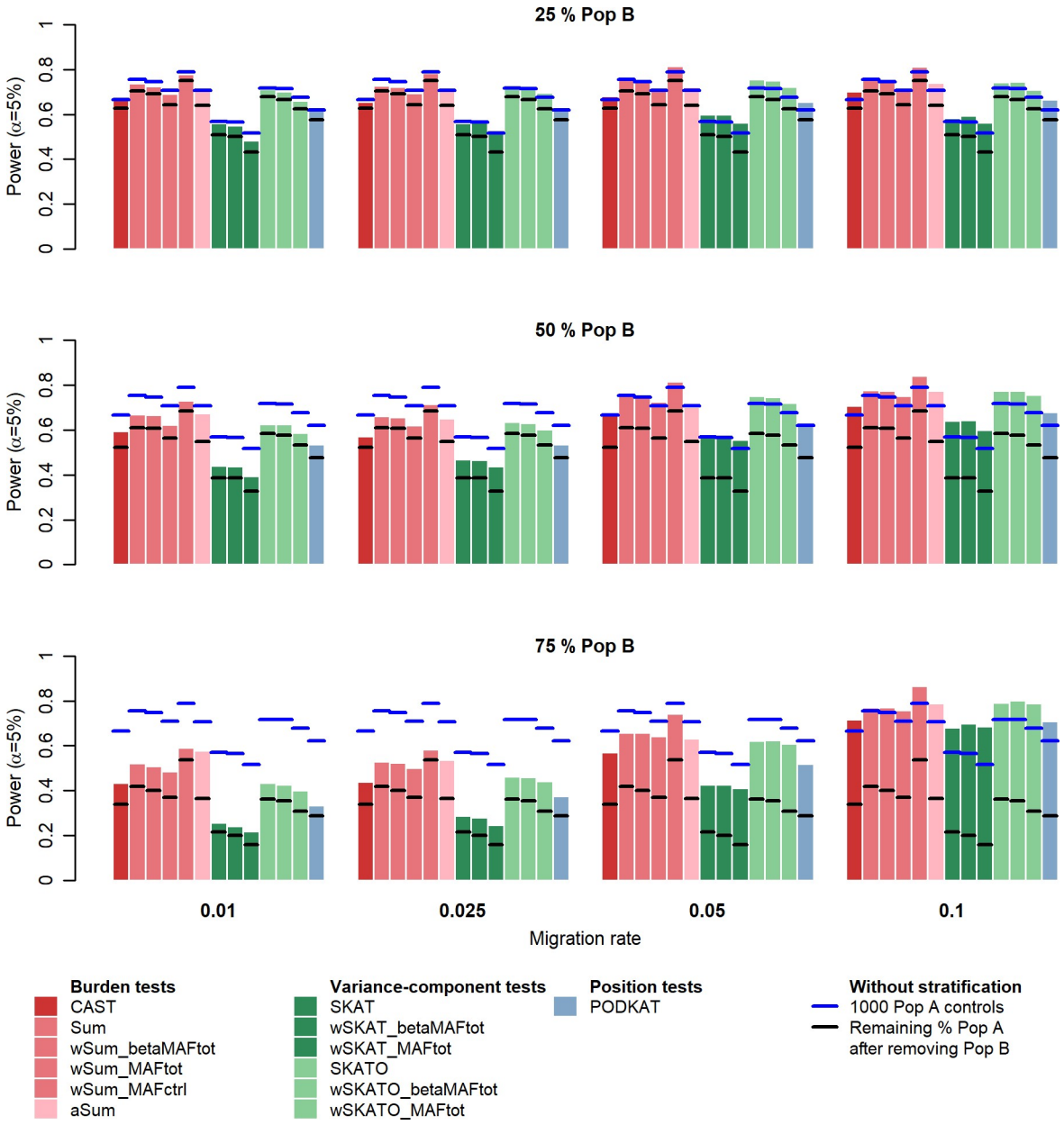

Fig 4. Powers at α = 5% after PCA model correction. PCA model correction was performed on all scenarios except when 100% of controls are from population B (geographical covariates may totally predict the phenotype).

https://doi.org/10.1371/journal.pone.0207677.g004

remaining in the analysis. Obviously, the power is decreased in these analyses as the number of controls is much less. More interestingly, with large scale population structures, we observe that the power is not much higher when adding controls from population B. As we discussed previously, the PCs capture the disease genetic component which is confounded with the genetic population structure. The power is, as expected, greatly increased when adding controls from a very close population. However, this advantage is balanced by the increase of type I error which is not fully corrected by adjusting the model.

Conclusions

Our simulation study showed that, as expected, we observed a very high inflation of type I errors in the presence of strong stratification. This inflation could, in principle, be controlled by through standard correction methods. In this work, our objective was to assess the impact at a finer geographical scale, as rare variants tend to be more localized than common variants. In this context, the inflation was still present but notably smaller. Intriguingly, the standard methods were less efficient at correcting this bias, as they did not effectively capture geographical origin from genetic data. This work underlies the importance of selecting controls with similar genetic background even at very fine geographical scales in sequencing studies. Exploratory analyses of the population structure should not be neglected and adjusting for potential bias must be done carefully.

Supporting information

S1 Fig. PCA plots in function of the migration rate. PCA was performed on the pruned dataset (MAF$\geq$5% and $r^2 \leq 0.2$ in the total population A) with 1,000 individuals from each population A and B.
(PNG)

S2 Fig. Type I errors at α = 5% without population stratification. The red line corresponds to α = 5% and blue lines correspond to 95% confidence interval. Confidence interval is computed assuming that the number of false positives follows a binomial distribution with parameters 10,000 and 0.05.
(PNG)

S3 Fig. Type I errors at α = 5% varying the number of PC to integrate in the PCA model correction method. The red line corresponds to α = 5% and blue lines correspond to 95% confidence interval. Confidence interval is computed assuming that the number of false positives follows a binomial distribution with parameters 10,000 and 0.05.
(PNG)

S4 Fig. Type I errors at α = 5% varying the number of cases and controls. The red line corresponds to α = 5% and blue lines correspond to 95% confidence interval. Confidence interval is computed assuming that the number of false positives follows a binomial distribution with parameters 10,000 and 0.05.
(PNG)

S1 Table. Supplementary tables from Table A to Table C. Table A: Number of SNV in populations A and B for 10,000 simulated genes; Table B: Variant frequency distributions in populations A and B for 10,000 simulated genes; Table C: FST values for simulated data in function of the migration rate.
(DOCX)

S2 Table. Type I error values at α = 5%.
(CSV)

S3 Table. Power values at α = 5%.
(CSV)

Acknowledgments

The authors would like to thank the Genomics and Bioinformatics Core Facility of Nantes (GenoBiRD, Biogenouest). This work was supported by the French Regional Council of Pays-

de-la-Loire (VaCaRMe program to C. D., L.B. and R.R.), and the Agence Nationale de la Recherche (ANR-15-CE17-0008-01 to R.R.).

Author Contributions

Conceptualization: Richard Redon, Lise Bellanger, Christian Dina.

Formal analysis: Elodie Persyn.

Funding acquisition: Richard Redon, Christian Dina.

Investigation: Elodie Persyn.

Methodology: Elodie Persyn, Richard Redon, Lise Bellanger, Christian Dina.

Project administration: Richard Redon, Lise Bellanger, Christian Dina.

Supervision: Richard Redon, Lise Bellanger, Christian Dina.

Visualization: Elodie Persyn.

Writing – original draft: Elodie Persyn.

Writing – review & editing: Richard Redon, Lise Bellanger, Christian Dina.

References

1. Lee S, Abecasis GR, Boehnke M, Lin X. Rare-Variant Association Analysis: Study Designs and Statistical Tests. Am J Hum Genet. 2014; 95: 5–23. https://doi.org/10.1016/j.ajhg.2014.06.009 PMID: 24995866
2. Morgenthaler S, Thilly WG. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). Mutat Res. 2007; 615: 28–56. https://doi.org/10.1016/j.mrfmmm.2006.09.003 PMID: 17101154
3. Madsen BE, Browning SR. A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet. 2009; 5: e1000384. https://doi.org/10.1371/journal.pgen.1000384 PMID: 19214210
4. Morris AP, Zeggini E. An Evaluation of Statistical Approaches to Rare Variant Analysis in Genetic Association Studies. Genet Epidemiol. 2010; 34: 188–193. https://doi.org/10.1002/gepi.20450 PMID: 19810025
5. Price AL, Kryukov GV, de Bakker PIW, Purcell SM, Staples J, Wei L-J, et al. Pooled association tests for rare variants in exon-resequencing studies. Am J Hum Genet. 2010; 86: 832–838. https://doi.org/10.1016/j.ajhg.2010.04.005 PMID: 20471002
6. Han F, Pan W. A data-adaptive sum test for disease association with multiple common or rare variants. Hum Hered. 2010; 70: 42–54. https://doi.org/10.1159/000288704 PMID: 20413981
7. Pan W, Shen X. Adaptive Tests for Association Analysis of Rare Variants. Genet Epidemiol. 2011; 35: 381–388. https://doi.org/10.1002/gepi.20586 PMID: 21520272
8. Neale BM, Rivas MA, Voight BF, Altshuler D, Devlin B, Orho-Melander M, et al. Testing for an Unusual Distribution of Rare Variants. PLoS Genet. 2011; 7: e1001322. https://doi.org/10.1371/journal.pgen.1001322 PMID: 21408211
9. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test. Am J Hum Genet. 2011; 89: 82–93. https://doi.org/10.1016/j.ajhg.2011.05.029 PMID: 21737059
10. Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA, et al. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. Am J Hum Genet. 2012; 91: 224–237. https://doi.org/10.1016/j.ajhg.2012.06.007 PMID: 22863193
11. Tintle N, Aschard H, Hu I, Nock N, Wang H, Pugh E. Inflated Type I Error Rates When Using Aggregation Methods to Analyze Rare Variants in the 1000 Genomes Project Exon Sequencing Data in Unrelated Individuals: Summary Results from Group 7 at Genetic Analysis Workshop 17. Genet Epidemiol. 2011; 35: S56–S60. https://doi.org/10.1002/gepi.20650 PMID: 22128060
12. Mathieson I, McVean G. Differential confounding of rare and common variants in spatially structured populations. Nat Genet. 2012; 44: 243–246. https://doi.org/10.1038/ng.1074 PMID: 22306651

13. Babron M-C, de Tayrac M, Rutledge DN, Zeggini E, Génin E. Rare and low frequency variant stratification in the UK population: description and impact on association tests. PloS One. 2012; 7: e46519. https://doi.org/10.1371/journal.pone.0046519 PMID: 23071581

14. Jiang Y, Epstein MP, Conneely KN. Assessing the impact of population stratification on association studies of rare variation. Hum Hered. 2013; 76: 28–35. https://doi.org/10.1159/000353270 PMID: 23921847

15. Liu Q, Nicolae DL, Chen LS. Marbled inflation from population structure in gene-based association studies with rare variants. Genet Epidemiol. 2013; 37: 286–292. https://doi.org/10.1002/gepi.21714 PMID: 23468125

16. O'Connor TD, Kiezun A, Bamshad M, Rich SS, Smith JD, Turner E, et al. Fine-Scale Patterns of Population Stratification Confound Rare Variant Association Tests. Chen L, editor. PLoS ONE. 2013; 8: e65834. https://doi.org/10.1371/journal.pone.0065834 PMID: 23861739

17. Zhang Y, Guan W, Pan W. Adjustment for population stratification via principal components in association analysis of rare variants. Genet Epidemiol. 2013; 37: 99–109. https://doi.org/10.1002/gepi.21691 PMID: 23065775

18. Zawistowski M, Reppell M, Wegmann D, St Jean PL, Ehm MG, Nelson MR, et al. Analysis of rare variant population structure in Europeans explains differential stratification of gene-based tests. Eur J Hum Genet EJHG. 2014; 22: 1137–1144. https://doi.org/10.1038/ejhg.2013.297 PMID: 24398795

19. Karakachoff M, Duforet-Frebourg N, Simonet F, Le Scouarnec S, Pellen N, Lecointe S, et al. Fine-scale human genetic structure in Western France. Eur J Hum Genet EJHG. 2015; 23: 831–836. https://doi.org/10.1038/ejhg.2014.175 PMID: 25182131

20. Liu DJ, Leal SM. A Novel Adaptive Method for the Analysis of Next-Generation Sequencing Data to Detect Complex Trait Associations with Rare Variants Due to Gene Main Effects and Interactions. PLoS Genet. 2010; 6: e1001156. https://doi.org/10.1371/journal.pgen.1001156 PMID: 20976247

21. Persyn E, Karakachoff M, Le Scouarnec S, Le Clézio C, Campion D, Consortium FE, et al. DoEstRare: A statistical test to identify local enrichments in rare genomic variants associated with disease. Wang K, editor. PLOS ONE. 2017; 12: e0179364. https://doi.org/10.1371/journal.pone.0179364 PMID: 28742119

22. Bodenhofer U. PODKAT: An R Package for Association Testing Involving Rare and Private Variants. R package version 1.0.3; 2015.

23. Li B, Leal SM. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am J Hum Genet. 2008; 83: 311–321. https://doi.org/10.1016/j.ajhg.2008.06.024 PMID: 18691683

24. Chapman JM, Cooper JD, Todd JA, Clayton DG. Detecting Disease Associations due to Linkage Disequilibrium Using Haplotype Tags: A Class of Tests and the Determinants of Statistical Power. Hum Hered. 2003; 56: 18–31. https://doi.org/10.1159/000073729 PMID: 14614235

25. Clayton D, Chapman J, Cooper J. Use of unphased multilocus genotype data in indirect association studies. Genet Epidemiol. 2004; 27: 415–428. https://doi.org/10.1002/gepi.20032 PMID: 15481099

26. Lee S, Wu MC, Lin X. Optimal tests for rare variant effects in sequencing association studies. Biostat Oxf Engl. 2012; 13: 762–775. https://doi.org/10.1093/biostatistics/kxs014 PMID: 22699862

27. Davies RB. Algorithm AS 155: The Distribution of a Linear Combination of $\chi 2$ Random Variables. J R Stat Soc Ser C Appl Stat. 1980; 29: 323–333. https://doi.org/10.2307/2346911

28. Schaid DJ, Sinnwell JP, McDonnell SK, Thibodeau SN. Detecting genomic clustering of risk variants from sequence data: cases versus controls. Hum Genet. 2013; 132: 1301–1309. https://doi.org/10.1007/s00439-013-1335-y PMID: 23842950

29. Lin W-Y. Association testing of clustered rare causal variants in case-control studies. PloS One. 2014; 9: e94337. https://doi.org/10.1371/journal.pone.0094337 PMID: 24736372

30. Silverman BW. Density Estimation for Statistics and Data Analysis. CRC Press; 1986.

31. Chen Y-C, Carter H, Parla J, Kramer M, Goes FS, Pirooznia M, et al. A hybrid likelihood model for sequence-based disease association studies. PLoS Genet. 2013; 9: e1003224. https://doi.org/10.1371/journal.pgen.1003224 PMID: 23358228

32. Ionita-Laza I, Makarov V, ARRA Autism Sequencing Consortium, Buxbaum JD. Scan-statistic approach identifies clusters of rare disease variants in LRP2, a gene linked and associated with autism spectrum disorders, in three datasets. Am J Hum Genet. 2012; 90: 1002–1013. https://doi.org/10.1016/j.ajhg.2012.04.010 PMID: 22578327

33. Fier H, Won S, Prokopenko D, AlChawa T, Ludwig KU, Fimmers R, et al. "Location, Location, Location": a spatial approach for rare variant analysis and an application to a study on non-syndromic cleft lip with or without cleft palate. Bioinforma Oxf Engl. 2012; 28: 3027–3033. https://doi.org/10.1093/bioinformatics/bts568 PMID: 23044548

34. Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, Altshuler D. Calibrating a coalescent simulation of human genome sequence variation. Genome Res. 2005; 15: 1576–1583. https://doi.org/10.1101/gr.3709305 PMID: 16251467

35. Weir BS, Cockerham CC. Estimating F-Statistics for the Analysis of Population Structure. Evolution. 1984; 38: 1358–1370. https://doi.org/10.1111/j.1558-5646.1984.tb05657.x PMID: 28563791

36. Patterson N, Price AL, Reich D. Population Structure and Eigenanalysis. PLoS Genet. 2006; 2: e190. https://doi.org/10.1371/journal.pgen.0020190 PMID: 17194218

37. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006; 38: 904–909. https://doi.org/10.1038/ng1847 PMID: 16862161

38. Che R, Jack JR, Motsinger-Reif AA, Brown CC. An adaptive permutation approach for genome-wide association study: evaluation and recommendations for use. BioData Min. 2014; 7: 9. https://doi.org/10.1186/1756-0381-7-9 PMID: 24976866

39. Epstein MP, Duncan R, Jiang Y, Conneely KN, Allen AS, Satten GA. A Permutation Procedure to Correct for Confounders in Case-Control Studies, Including Tests of Rare Variation. Am J Hum Genet. 2012; 91: 215–223. https://doi.org/10.1016/j.ajhg.2012.06.004 PMID: 22818855

40. Nelis M, Esko T, Mägi R, Zimprich F, Zimprich A, Toncheva D, et al. Genetic Structure of Europeans: A View from the North–East. Fleischer RC, editor. PLoS ONE. 2009; 4: e5472. https://doi.org/10.1371/journal.pone.0005472 PMID: 19424496

41. Génin E, Dina C, Ludwig T, Quenez O, Letort S, Lindenbaum P, et al. Are population-specific panels of exomes useful to identify disease variants: Insights from the French Exome Project. Vancouver: Presented at the 69th Annual Meeting of The American Society of Human Genetics; 2016.

42. Schisterman EF, Cole SR, Platt RW. Overadjustment Bias and Unnecessary Adjustment in Epidemiologic Studies: Epidemiology. 2009; 20: 488–495. https://doi.org/10.1097/EDE.0b013e3181a819a1 PMID: 19525685

43. Coventry A, Bull-Otterson LM, Liu X, Clark AG, Maxwell TJ, Crosby J, et al. Deep resequencing reveals excess rare recent variants consistent with explosive population growth. Nat Commun. 2010; 1: 131. https://doi.org/10.1038/ncomms1130 PMID: 21119644

44. Gravel S, Henn BM, Gutenkunst RN, Indap AR, Marth GT, Clark AG, et al. Demographic history and rare allele sharing among human populations. Proc Natl Acad Sci. 2011; 108: 11983–11988. https://doi.org/10.1073/pnas.1019276108 PMID: 21730125

45. Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, Gravel S, et al. Evolution and Functional Impact of Rare Coding Variation from Deep Sequencing of Human Exomes. Science. 2012; 337: 64–69. https://doi.org/10.1126/science.1219240 PMID: 22604720