locStra: Fast analysis of regional/global stratification in whole genome sequencing (WGS) studies

Georg Hahn*, Sharon M. Lutz*, Julian Hecker†, Dmitry Prokopenko‡, Michael Cho†, Edwin K. Silverman†, Scott Weiss† and Christoph Lange*

Abstract

locStra is an *R*-package for the analysis of regional and global population stratification in whole genome sequencing studies, where regional stratification refers to the substructure defined by the loci in the region. Population substructure can be assessed based on the genetic covariance matrix, the genomic relationship matrix, and the unweighted/weighted genetic Jaccard similarity matrix. Using a sliding window approach, the regional similarity matrices are compared to the global ones, based on user-defined window sizes and metrics, e.g. correlation between regional and global eigenvectors. An algorithm for the specification of the window size is provided. As the implementation fully exploits sparse matrix algebra and is written in C++, the analysis is highly efficient. Even on single cores, for realistic study sizes (several thousand subjects, several million RVs per subject), the runtime for the genome-wide computation of all regional similarity matrices does typically not exceed one hour, enabling an unprecedented investigation of regional stratification across the entire genome. The package is applied to three WGS studies, illustrating the varying patterns of regional substructure across the genome and its effects on association testing.

1 Introduction

Genetic association studies are a popular mapping tool; however, they can be vulnerable to confounding due to population substructure (Laird and Lange, 2010). Numerous methods have been proposed to address this issue (Devlin and Roeder, 1999; Pritchard et al., 2000). Popular approaches rely on the genetic covariance matrix of the genotype data: EIGENSTRAT, STRATSCORE, multidimensional scaling, etc. (Price et al., 2006; Patterson et al., 2006; Lee et al., 2012), or on the genomic relationship matrix (Yang et al., 2011). For populations with recent admixture where each subject contains different proportions of the ancestral genomes, local ancestry-approaches have been suggested (Sankararaman et al., 2008).

While there is strong evidence for regional stratification (Price et al., 2009; Martin et al., 2018; Zhong et al., 2019), the matrix-based approaches are typically computed only globally. Although,

*Department of Biostatistics, T.H. Chan School of Public Health, Harvard University, Boston, MA 02115, USA
†Department of Medicine, Brigham and Women's Hospital, Harvard University, Boston MA 02115, USA
‡Massachusetts General Hospital, Harvard University, Boston, MA 02114, USA

for the validity of the matrix-based approaches, it is only required that the selected loci are not in linkage disequilibrium (LD) (Laird and Lange, 2010) and there are no theoretical constraints as to whether the loci are selected genome-wide or from a specific region, the computational burden has generally been prohibitive to use existing implementations for a genome-wide analysis of regional stratification. As most matrix-based approaches are designed for common, uncorrelated variant data, i.e. loci that are not LD, many genomic regions do not contain a sufficient number of such loci for the regional computation of matrix-based approaches.

With the arrival of whole genome sequencing data, an abundance of data on densely spaced rare variants (RVs) that are mostly not in LD became generally available. As RVs can be more informative about recent admixture (Bodmer and Bonilla, 2008; Kryukov et al., 2009; Keinan and Clark, 2012), approaches based on Jaccard similarity matrices that utilize RV/WGS data have been developed (Prokopenko et al., 2016; Schlauch et al., 2017). However, the computational bottleneck has remained.

We developed *locStra*, an *R*-package implementing four approaches to assess population stratification in RVs at the regional and global level using (1) the genetic covariance matrix, (2) the genomic relationship matrix, (3) the unweighted and (4) weighted Jaccard similarity matrices. Written in C++, all similarity matrices are algebraically transformed so that the computations are executed on sparse data structures. The sparse matrix structure is maintained throughout all computations to maximize computational efficiency. Using sliding windows (Morrison et al., 2013; Yazdani et al., 2015) of user-specified length, *locStra* enables the fast analysis of regional stratification at the genome-wide level. An algorithm for the selection of the window sizes is proposed.

Applications of *locStra* to three WGS studies illustrate the importance of the ability to investigate regional substructure and to adjust for it in genetic association testing. We also evaluate the differences between the four similarity matrices and the computational features of *locStra* in terms of runtime. *locStra* makes substantive research into regional stratification generally feasible in WGS studies at a computational cost that is not even prohibitive on a single CPU system.

2 Implementation

The core implementation of *locStra* is based on fully sparse matrix algebra in C++, using *RcppEigen* of Bates and Eddelbuettel (2013). The package is available on *The Comprehensive R Archive Network*.

Four functions provide C++ implementations of standard approaches to population stratification which are made available through wrapper functions in *R*: *covMatrix* computes the genetic covariance matrix (Price et al., 2006), *grMatrix* computes the genomic relationship (Yang et al., 2011), *jaccardMatrix* computes the Jaccard similarity matrix (Prokopenko et al., 2016), and *sMatrix* implements the weighted Jaccard matrix (Schlauch et al., 2017). The unweighted and weighted Jaccard indices, traditionally a similarity index for sets, are two recently proposed approaches for the analysis of rare variant data which were shown to provide a higher resolution than the afore-

mentioned approaches (Prokopenko et al., 2016). The entries of the Jaccard matrix measure the set-theoretic similarity of the genomic data of all pairs of subjects, and can be computed efficiently using only binary operations. All functions have a boolean argument *dense* to switch between C++ implementations for dense and sparse matrix algebra, depending on the type of the input data (default is *dense=False*).

The main function is *fullscan*. It has five arguments, and allows for a flexible specification of the regional population stratification scan:

- The first input is the matrix containing the genotype data. The input matrix contains the data for one individual per column.
- The second argument is a two-column matrix (called *windows*) that contains the window specification of the scan. The sliding windows can easily be generated by an auxiliary function which is provided. The window size can be chosen arbitrarily.
- The third argument, *matrixFunction*, handles the processing of each window and takes one input argument, e.g. *covMatrix*, *grMatrix*, etc.
- The third input is the specification of a *summaryFunction* for the processed data before comparison. This can be any function that is compatible with the output of *matrixFunction*, e.g. the function *powerMethod* computing the largest eigenvector which is provided.
- The fifth input argument is the function *comparisonFunction* that compares summaries, e.g. the native *R* correlation function *cor* applied to the first regional and the first global eigenvector.

The supplementary material contains all details on the functions computing similarity matrices, on the main and auxiliary functions, and an algorithm for the selection of the window size.

3 Usage and Data Analysis Examples

The supplementary material includes implementation details of *locStra* and an example using the 1,000 Genomes Project (The 1000 Genomes Project Consortium, 2015). Using LD-pruned RVs ($< 1\%$) in the European super population (503 subjects, ca. 5 million RVs) and a sliding window approach of $120,000$ RVs (as suggested by the window selection algorithm), we computed the correlations between the first eigenvector of all regional similarity matrices with the corresponding first eigenvector of the global similarity matrix. This was done for all four different types of similarity matrices. For example, on chromosome 16 (Figure 1), for all four similarity matrices, the correlations between the first regional and first global eigenvectors are very small, except for a small genomic region where all correlations approach almost 1. This suggest that the substructure, as it is captured by similarity matrices, is regionally very different form the global substructure in terms of the eigenvectors. We also examine the varying patterns of regional substructure in a Childhood Asthma WGS study from Costa Rica and demonstrate the benefits of regional substructure

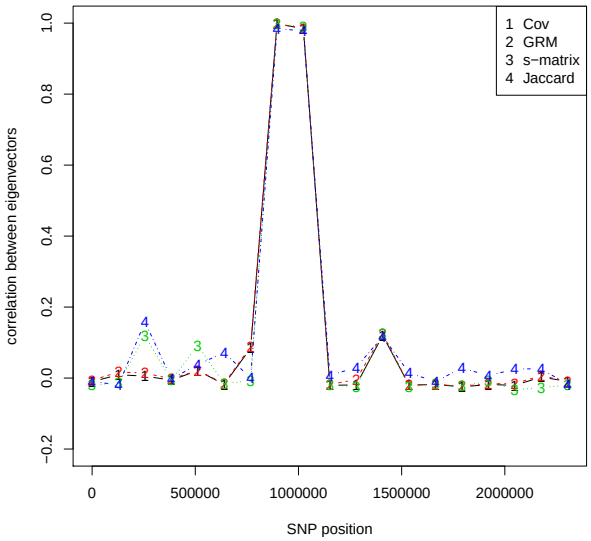

Figure 1: Results for chromosome 16 of the EUR super population of the 1,000 Genome Project. For sliding windows of size 128,000, correlations between global and regional first eigenvectors of the covariance matrix, GRM matrix, s-matrix, and Jaccard similarity matrix.

adjustment in genetic association testing exemplarily using a WGS study for Chronic obstructive pulmonary disease (COPD).

4 Conclusion

The *R*-package *locStra* is the first software packages that enables a comprehensive genome-wide analysis of regional stratification based on similarity matrices in WGS studies. Given a runtime of around 500 seconds for the genome-wide analysis of all sliding windows in the EUR super population of the 1,000 Genome Project (one Intel QuadCore i5-7200 CPU with 2.5 GHz and 8 GiB of RAM), *locStra* provides the community with the general ability to investigate regional stratification patterns at a genome-wide level in WGS studies and to adjust the association analysis for such patterns.

Appendix

We start with a detailed description of the software functionality provided by the locStra package (Section A). We then present implementation details on the sparse calculations carried out in our *R* package (Section B), including their theoretical runtimes. In Section C, we present a data analysis demonstrating (1) regional and global population stratification for certain chromosomes in populations of the 1,000 Genome Project (The 1000 Genomes Project Consortium, 2015), (2) a runtime analysis for all four similarity approaches (covariance matrix, GRM matrix, s-matrix,

Jaccard matrix), (3) a runtime comparison of locStra to *PLINK2* (Chang et al., 2015; Purcell and Chang, 2019), and (4) an approach to select suitable window sizes for population stratification. Section D highlights another regional analysis for the Costa Rica population isolate. Section E shows an application of regional stratification in which we demonstrate that it is beneficial to correct a linear regression using both global and regional PCAs.

A Software description

The locStra package makes a total of seven functions available.

A.1 Dense and sparse matrix implementations

Four functions provide C++ implementations of standard approaches to population stratification, both for dense and sparse matrix algebra. The code handles dense and sparse input matrices separately since either version can be inefficient if used for matrices of the wrong type. All following functions have a boolean argument *dense* to select which C++ implementation is to be used. The default is *dense=False*.

1. The function *covMatrix* computes the genetic covariance matrix. The input is allowed to be any real valued matrix.

2. The function *grMatrix* computes the genomic relationship matrix (GRM) as defined in Yang et al. (2011). The input must be a binary matrix. Both the classic and robust versions (Wang et al., 2017) of the GRM are supported, and can be switched using the boolean flag *robust*. The default is *robust=True*.

3. The function *jaccardMatrix* computes the Jaccard similarity matrix (Prokopenko et al., 2016). The input must be a binary matrix.

4. The function *sMatrix* implements the weighted Jaccard matrix (Schlauch, 2016). In addition to the boolean *dense* argument, the function *sMatrix* also has a boolean argument *phased* to indicate if the input data is phased (default is *phased=False*). The last argument is the integer *minVariants* which is a cutoff value for the minimal number of variants to consider (default is *minVariants=5*).

A.2 Main function

The main function of the package is *fullscan*. It has five arguments and allows for a flexible specification of the regional population stratification scan of the data through its generic structure.

- The first input is the (sparse) matrix containing the sequencing data. The input matrix is assumed to be oriented to contain the data for each individual per column.

- The second argument is a two-column matrix (called *windows*) that contains the window specification of the scan. The two entries per row are the start and end positions of each window. The matrix of sliding windows can easily be generated with the auxiliary function *makeWindows* (Section A.3).

- The third argument, *matrixFunction*, handles the processing of each sliding window. The function takes one input argument (often a matrix). Any function can in principle be used. Typical choices are *covMatrix*, *grMatrix*, *jaccardMatrix*, or *sMatrix*.

- Next, the modular structure of *fullscan* requires the specification of a *summaryFunction* for the processed data before comparison. This can be any function of one input argument that is compatible with the output of *matrixFunction*. The computation of the largest eigenvector via function *powerMethod* (Section A.3) is an intuitive choice.

- *fullscan* uses its fifth input argument, the function *comparisonFunction*, to compare summaries (e.g., first eigenvectors) on a regional and a global level. The *comparisonFunction* has two arguments as input, both of which need to be compatible with the output of the function *summaryFunction*. One example is the native *R* correlation function *cor* for two vectors.

The output of *fullscan* is a two column matrix with global and regional comparison values per row, where each row corresponds to a row (and thus a window) in matrix *windows* in the same order.

A.3 Auxiliary functions

Two functions provide additional functionality:

1. The function *makeWindows* generates a two-column matrix of non-overlapping or overlapping windows for the main function *fullscan*. The function takes as its arguments the length of the data, the window size and an offset. If the offset is set equal to the window size, non-overlapping windows are obtained. If the offset is less than the window size, sliding windows of given window size and offset are obtained.

2. The function *powerMethod* provides a C++ implementation of the power method for fast iterative computation of the largest eigenvector (von Mises and Pollaczek-Geiringer, 1929). The function can be used as *summaryFunction* in the main function *fullscan*.

A.4 Other comparison measures

The modular structure of locStra allows to specify (1) the similarity measure on the genome (the *matrixFunction*; for instance, the Jaccard matrix); (2) the summary statistic for the similarity matrix as function *summaryFunction*; and (3) a comparison measure on either the similarity matrices or the summary statistic (function *comparisonFunction*). In this work we always summarize the

four similiarity matrices with the first eigenvector (as *summaryFunction*) and compare the correlation between eigenvectors (as *comparisonFunction*). However, many more sensible choices exist which include:

1. The similarity matrices can be compared directly using, for instance, the $L_{p,q}$, Frobenius, maximum or Schatten matrix norms as *comparisonFunction*. In this case the *summaryFunction* is the identity function.

2. Apart from eigenvectors, two similarity matrices can be summarized using other traditional tools such as their eigenvalues, or the condition number of the difference between both which, if large, indicates that the matrices are close in this specific sense.

3. Apart from the first eigenvector, the similarity matrices can be summarized with a linear combination of higher order eigenvectors to capture more principal components. Moreover, the eigenvectors can be weighted with their corresponding eigenvalues.

4. Apart from using vector correlation, eigenvectors and other vector-valued measures can be compared using vector norms, the angle between them, etc.

However, some measures might be more meaningful than others depending on the context of the comparison and application. We did experiment with different measures and found the correlation between the first eigenvectors to capture best the variability within each chromosome.

B Details on the implementation

This section briefly describes two important implementation details (for computing the covariance and Jaccard matrices) employed to enable fully sparse matrix algebra. The GRM matrix (Yang et al., 2011) and the s-matrix (Schlauch, 2016) were computed as described in their respective publications. Throughout the section, the input data $X \in \mathbb{R}^{m\times n}$ is assumed to contain (genomic) data of length m in each of the n columns, one column per individual. The parameter m therefore represends the number of loci included in the computation of the similarity matrix and n is the number of study subjects. At the end of this section, theoretical runtimes of our implementations are given.

B.1 Covariance matrix

To compute the covariance matrix in dense algebra, let $v \in \mathbb{R}^n$ be the column means and let $Y \in \mathbb{R}^{m\times n}$ be the matrix consisting of the rows of X with their mean substracted. Then

$$\operatorname{cov}(X) = \frac{1}{m-1} Y^\top Y.$$

In sparse algebra, the matrix X cannot be normalised as in the dense case by simply substracting the column means, since this would result in a dense matrix which easily exceeds available memory.

method	dense	sparse
covariance matrix	$O(mn^2)$	$O(smn^2 + n^2)$
unweighted Jaccard	$O(mn^2)$	$O(smn^2 + n^2)$
weighted Jaccard	$O(mn^2)$	$O(smn^2 + mn)$
GRM matrix	$O(mn^2)$	$O(smn^2 + n^2)$

Table 1: Theoretical runtimes of the four matrix approaches to compute similarity measures for both dense and sparse implementations. The runtimes are given in the parameters $m \in \mathbb{N}$ and $n \in \mathbb{N}$ of the input data $X \in \mathbb{R}^{m \times n}$ as well as the matrix sparsity parameter $s \in [0, 1]$.

To always stay within sparse algebra, the computation is split up suitably. To be precise, let v denote the column means as above, and $w \in \mathbb{R}^n$ be the column sums, then

$$\operatorname{cov}(X) = \frac{1}{m-1}\left(X^\top X - wv^\top - vw^\top + mvv^\top\right).$$

This formula has the advantage that the computation of $X^\top X$ can be carried out using only one sparse matrix multiplication involving the sparse input matrix, and the remaining three outer vector products result in $n \times n$ matrices, thus never exceeding the size of the output covariance matrix.

B.2 Jaccard similarity matrix

The entry (i, j) of the Jaccard matrix $\operatorname{jac}(X)$ is given as

$$\operatorname{jac}(X)_{ij} = \frac{|\{k : X_{ik} \wedge X_{jk}\}|}{|\{k : X_{ik} \vee X_{jk}\}|},$$

where the matrix X is binary.

A naïve approach to compute the entries of the Jaccard matrix loops over all entries of the Jaccard matrix and calculates the binary *and* as well as binary *or* operations on all combinations of two columns of X. Though having the same theoretical runtime, this naïve approach turned out to be slower in practice than the following technique which uses only one (sparse) matrix-matrix multiplication which is typically highly optimized in sparse matrix algebra packages.

Let $w \in \mathbb{R}^n$ be the column sums of X as before. Compute $Y = X^\top X$ via sparse matrix multiplication. The resulting matrix $Y \in \mathbb{R}^{n \times n}$ is dense. Compute a second matrix $Z \in \mathbb{R}^{n \times n}$ by adding w to all rows and all columns of $-Y$. Then, $\operatorname{jac}(X) = Y/Z$, where the division operation is performed componentwise. This approach is computationally very fast since it relies solely on one sparse matrix multiplication, and a few more operations on the matrices Y and Z which are already of the same size as the dense Jaccard output matrix.

B.3 Theoretical runtimes of dense and sparse implementations

Table 1 shows theoretical runtimes for both the dense and sparse matrix versions of the four similarity matrix approaches. It turns out that the runtimes for the dense computations of all

similarity matrices coincide, and that the sparse computations have slightly different runtimes. The following highlights the effort of the main computation steps in each case as a function of the parameters $m \in \mathbb{N}$ and $n \in \mathbb{N}$ of the input data $X \in \mathbb{R}^{m \times n}$, as well as the matrix sparsity parameter $s \in [0, 1]$ (the proportion of non-zero matrix entries).

In the dense case, computing the covariance matrix involves subtracting the column means in $O(mn)$ and multiplying $Y^\top Y$ in $O(mn^2)$. In the sparse case, the computation of $X^\top X$ requires $O(smn^2)$, and the computation of the three additional outer products requires another $O(n^2)$.

The Jaccard matrix involves computing $Y = X^\top X$ in $O(mn^2)$ in the dense case, and $O(smn^2)$ in the sparse case. Adding the column sums to the rows and columns of the resulting $n \times n$ matrix takes effort $O(n^2)$ in both the dense and sparse case.

The effort for computing the weighted Jaccard matrix (or s-matrix) stems from the computation of weights through row sums ($O(mn)$ in dense algebra and $O(smn)$ in sparse algebra), multiplying the input matrix with the weights (likewise $O(mn)$ in dense algebra and $O(smn)$ in sparse algebra), and one matrix-matrix multiplication ($O(mn^2)$ in dense algebra and $O(smn^2)$ in sparse algebra).

Computing the GRM matrix involves the calculation of population frequencies across rows ($O(mn)$ in dense algebra and $O(smn)$ in sparse algebra), one matrix-matrix multiplication ($O(mn^2)$ in dense algebra and $O(smn^2)$ in sparse algebra), multiplying the input matrix with the population frequencies ($O(mn)$ in dense algebra and $O(smn)$ in sparse algebra), and one outer product of the population frequencies in $O(n^2)$.

C Regional stratification analysis of the 1,000 Genome Project

To illustrate the practical relevance of locStra and the feasibility of regional substructure analysis at the genome-wide level, we apply locStra to all chromosomes of the 1,000 Genome Project (The 1000 Genomes Project Consortium, 2015) and take a closer look at the results for four chromosomes, precisely chromosomes 5, 7, 10, and 12. Moreover, we investigate runtimes across all chromosomes, and present an approach to select suitable window sizes for population stratification.

Before applying locStra, the raw data from the 1,000 Genome Project is prepared using PLINK2 with cutoff value 0.01 for option *--max-maf* to select rare variants. We applied LD pruning with parameters *--indep-pairwise 2000 10 0.01*. Analysis results are shown for the super population *EUR* of the 1,000 Genome Project.

All timings presented in this and the following sections were measured on one Intel QuadCore i5-7200 CPU with 2.5 GHz and 8 GiB of RAM.

C.1 Data analysis results for certain chromosomes of the 1,000 Genome Project

The analysis results for the four selected chromosomes of the EUR super population are shown in Figure 2. The regional substructure analysis reveals several notable features. Regardless of which type of similarity matrix is used for the regional substructure analysis, there are only a few genomic regions for which the regional and global substructures are similar in terms of the first

window size	locStra		PLINK2	
	global EV	full scan	global EV	full scan
1000	1.4	332.1	65.3	6343.6
10000	1.5	31.3	61.2	731.7
100000	1.5	4.9	66.7	189.1

Table 2: Runtimes in seconds between locStra and PLINK2 for the computation of the global eigenvector (global EV) and for a complete stratification scan of chromosome 1 of the 1,000 Genome Project as a function of the window size.

eigenvectors. Overall, there is substantial variability of the regional substructure across the genome, when measured via the similarity matrices. This observation also has implications for association mapping, as the association analysis is typically adjusted for the global eigenvectors to minimize potential genetic confounding. It will be subject of future research to investigate the best ways to incorporate regional substructure adjustments based on RVs in genetic association testing.

In the areas where the correlation between the regional and global first eigenvector is not high, the standard Jaccard approach is able to maintain the highest correlation values compared to the other similarity matrices. In the areas where the regional first eigenvectors of Cov, GRM and s-matrix/weighted Jaccard are highly correlated with the corresponding global first eigenvectors, the first eigenvector of the standard Jaccard approach is often almost uncorrelated with the global one. Further methodological and substantive research is required to understand the reasons for these performance differences. It is part of our ongoing research and beyond the scope of this manuscript.

C.2 Runtime of locStra for the 1,000 Genome Project Analysis

Figure 3 shows the runtime in seconds for the R function *fullscan* as a function of the window sizes. Each plot depicts the minimal and maximal runtime observed among any of the chromosomes per window size, as well as the mean runtime for a particular window size when averaged across all chromosomes. The maximum number of RVs per chromosome is 13577 and the sample size is 503 study subjects.

One can see that the mean runtime never exceeds 500 seconds for a complete scan of any chromosome. As expected, the runtime decreases for larger window sizes. For a realistic window size of e.g. 10^4 or 10^5, see Section C.4, the runtime for any method lies in the neighborhood of one minute for a full scan. Repeating the runtime analysis for the AFR super population group of the 1,000 Genome Project shows qualitative similar results.

C.3 Comparison of locStra to PLINK2 on chromosome 1 of the 1,000 Genome Project

We compare locStra to PLINK2. To this end, we first prepare the data of chromosome 1 using the same parameters as given in Section C. Since locStra and PLINK2 require different input files, we

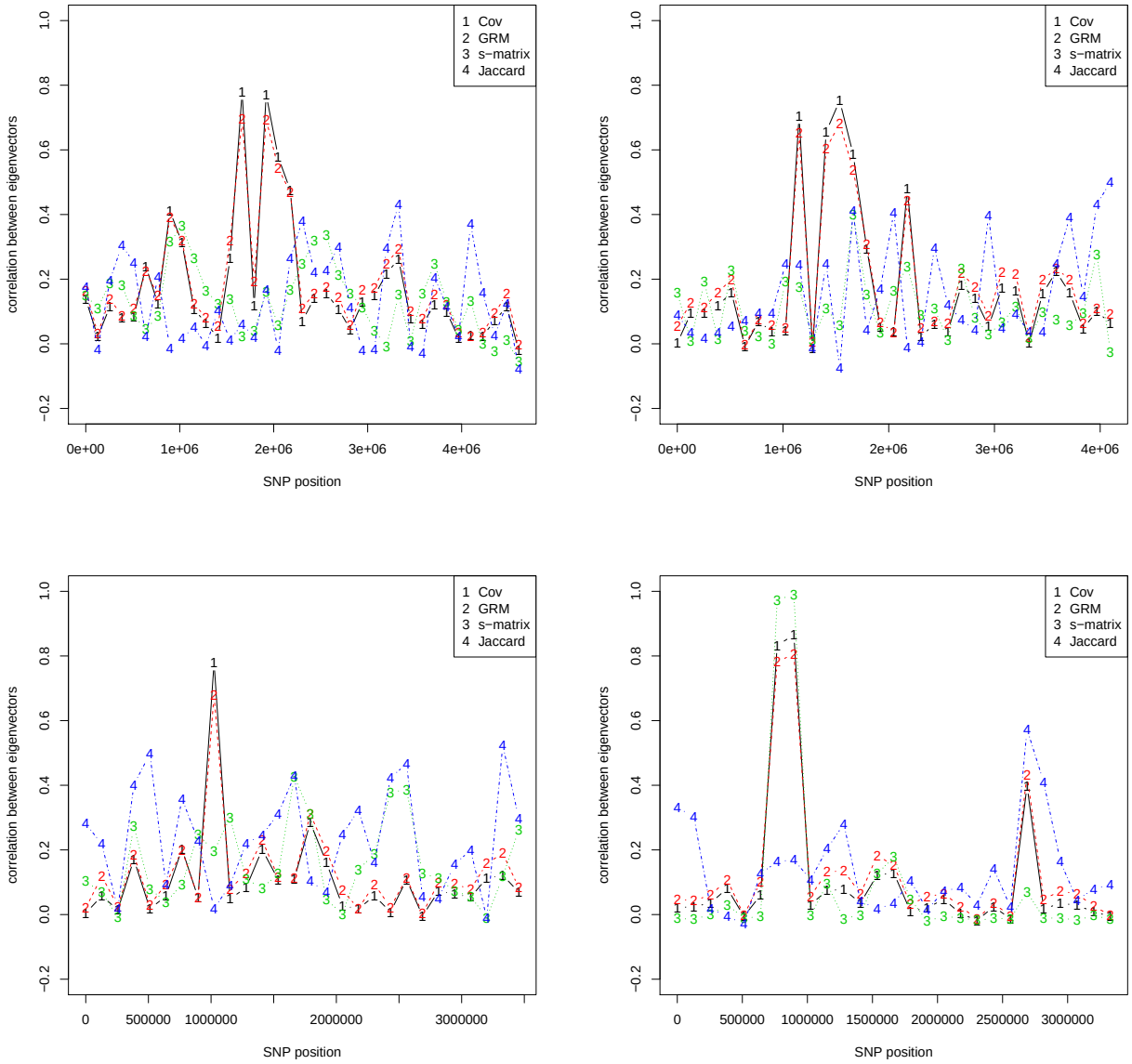

Figure 2: Super population EUR of the 1,000 Genomes Project. Correlation of regional to global eigenvectors for chromosomes 5 (top left), 7 (top right), 10 (bottom left), and 12 (bottom right). Covariance matrix, GRM matrix, s-matrix, and Jaccard matrix. Window size 128000 RVs.

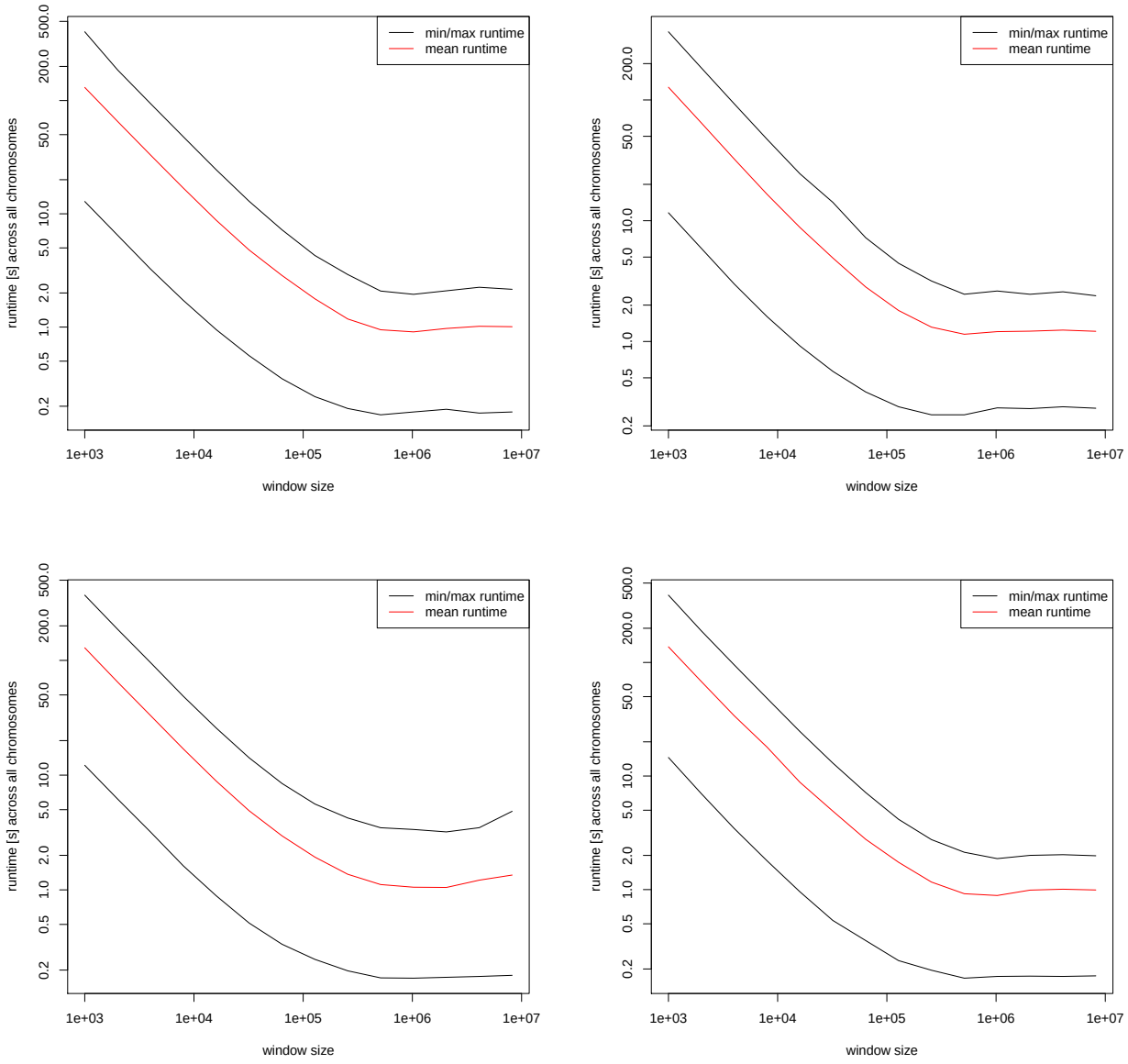

Figure 3: Super population EUR of the 1,000 Genomes Project. Runtime in seconds as a function of the window sizes across all chromosomes for the computation of the covariance matrix (top left), GRM matrix (top right), s-matrix (bottom left), and Jaccard matrix (bottom right). All plots show the minimal and maximal runtimes for any of the chromosomes, as well as the mean runtime averaged across all chromosomes. Logarithmic scale on the x- and y-axes.

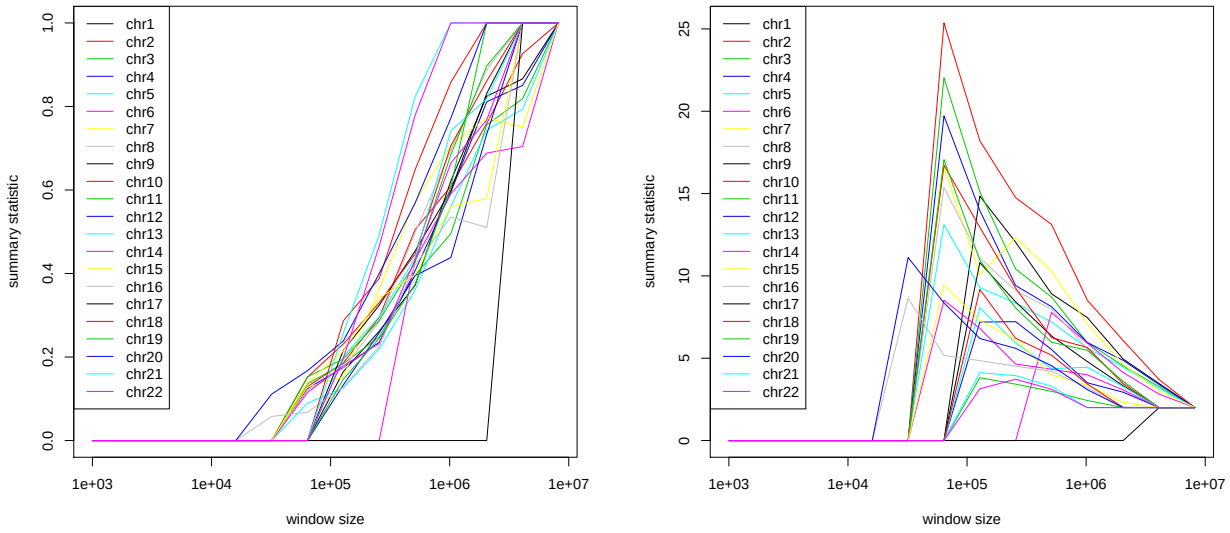

Figure 4: Super population EUR of the 1,000 Genomes Project. Left: Mean correlation across all windows as a function of the window size. Right: Mean correlation across all windows divided by the number of windows, again as a function of the window size. Input data are the correlations between global and regional eigenvectors of the Jaccard matrices for differrent window sizes.

write the curated data for chromosome 1 once in the *.bed* format for PLINK2 to read, and once convert it into a sparse matrix of class *Matrix* in *R* (saved as *.Rdata* file).

In PLINK2, the first eigenvector can be computed on the *.bed* file input with the option *--pca 1*, and in order to do a regional scan, a variant range on the data can be specified with the parameters *--from* and *--to* followed by the rs numbers. After computing the first global eigenvector or one regional eigenvector, PLINK2 writes the vector data into a file *.eigenvec*, from which the eigenvectors are read in order to compute correlations between them. In this way, a complete scan of global to regional correlations can be carried out in PLINK2.

For locStra, we load the sparse matrix input data into *R* and employ the function *fullscan* from the *R* package to carry out a complete scan.

Results for three different window sizes are given in Table 2, showing both the runtimes (in seconds) for the computation of the single global eigenvector on the full data, as well as for a complete scan (which includes the computation of the global eigenvector before starting the scan). Since the times for PLINK2 necessarily include the read/write operations for file in- and outputs, we likewise report times for locStra that include the reading time of the input data for a fair comparison. As visible from the table, locStra is at least one order of magnitude faster than PLINK2 for realistic studies (several thousand subjects, several million RVs per subject), where the speed-up seems to be more pronounced for larger window sizes.

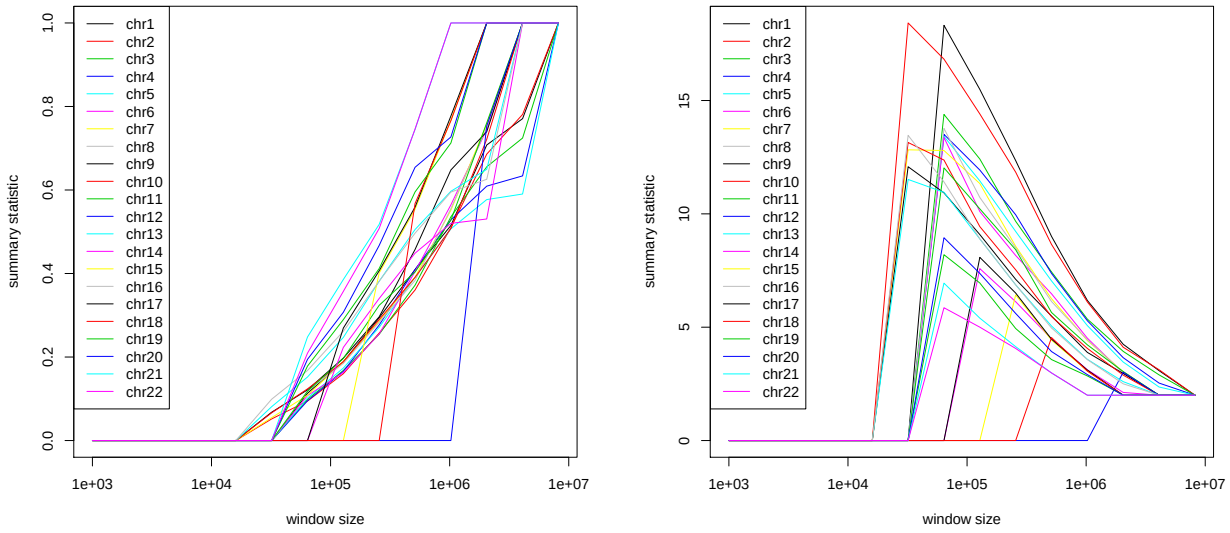

Figure 5: Super population AFR of the 1,000 Genomes Project. Setting as in Figure 4. Left: Mean correlation across all windows as a function of the window size. Right: Mean correlation across all windows divided by the number of windows, again as a function of the window size.

C.4 Selecting suitable window sizes for population stratification

An interesting question pertains to the selection of an appropriate window size for population stratification. Two quantities work against each other in the process of selecting a suitable window size: As the window size becomes larger, less windows are used in the scan of the data, and thus the correlation between regional and global eigenvectors increases as seen in Figure 4 (left). On the other hand, larger window sizes imply the usage of fewer windows across the genomic data, thus causing less data points to be calculated and results to be less meaningful.

A natural tradeoff is therefore to define a measure by multiplying the mean correlation among all windows (for a particular window size) with the number of windows generated using that size. The resulting measure is displayed in Figure 4 (right). It can be seen that for small and large window sizes, the product of mean correlation and number of windows is close to zero. For the 1,000 Genomes Project, we observe a peak in this measure at around a window size of $100,000$ RVs. Interestingly, the peak occurs at an almost identical position for all chromosomes. This is attributed to the fact that the slope in Figure 4 (left) is very similar for all chromosomes. For the analysis of data from the 1,000 Genomes Project, we thus recommend a window size of around $100,000$ RVs. We propose to use this algorithm to select a window size as an heuristic guideline.

Repeating the above analysis for the super population AFR of the 1,000 Genome Project in Figure 5 again shows qualitative similar results.

D Data analysis of Childhood Asthma Study from Costa Rica (population isolate)

We now analyze the global and the regional population substructure and their differences in terms of clustering in a dataset of a Costa Rica population isolate, as one expects stronger degrees of stratification in such a sample. The dataset includes children aged 6 to 14 and their parents (2736 subjects, 1824 of which are parents) from GACRS (Genetics of Asthma in Costa Rica Study) family-based trios recruited from a genetically homogeneous Hispanic population isolate living in the Central Valley of Costa Rica. This population has one of the highest prevalences of asthma in the world. Please see Hunninghake et al. (2007) for a detailed description of the recruitment process. The study has been sequenced as part of the TOPMED Project. The data is available through dbGaP (NHLBI TOPMed, 2019).

To avoid genetic correlations among study subjects due to family structure, we only select the parents for the analysis, and prepare their genetic data using PLINK2 with cutoff value 0.01 for option *--max-maf* to select rare variants. We applied LD pruning with parameters *--indep-pairwise 2000 10 0.01*.

To evaluate the population substructure in the data, we compute the Jaccard similarity matrix globally and regionally for one window on each chromosome. We selected here the Jaccard approach for ease of presentation. None of the qualitative conclusions that we reach below would have been different, if we had selected a different similarity matrix.

In Figure 6, we provide the plot for the first two "global" PCs of the Jaccard matrix that was computed globally based on loci from the entire genome. The plot shows clear evidence of population substructure among the parents in the Costa Rica sample. For each of the 22 autosomal chromosomes, Figure 7 contains the plot of the first two "regional" PCs of a region on the chromosome, where the Jaccard matrix was computed based on RVs from a region of 10^5 loci.

To generate Figure 6, we randomly sampled 10^5 RVs from each chromosome, and combined them into one matrix on which the Jaccard similarity measure was computed. To generate each of the subplots in Figure 7, we selected a window size of 10^5 for each chromosome (resulting in around 20 windows per chromosome depending on the size of the chromosome data), and computed the similarity matrix on the middle window. In both cases we display the first two PCs of the corresponding Jaccard similarity matrix obtained in this fashion.

The two plots, Figures 6 and 7, clearly illustrate that the regional substructure can vary substantially. While it can be very similar to the global substructure for some of the regions, often it is more extreme or fundamentally different, showing sub-clusters that are not detectable in the global components. The corresponding results for the other similarity matrix approaches support the same conclusion (data not shown).

The findings of the Costa Rica data analysis clearly demonstrate the importance of regional substructure analysis and the utility of the proposed locStra package.

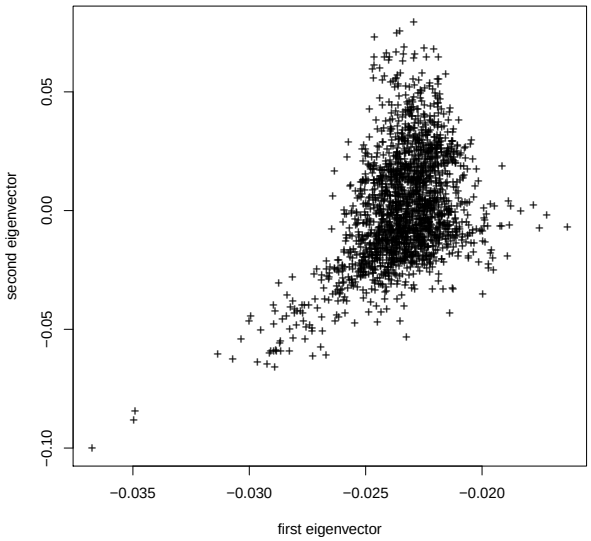

Figure 6: Costa Rica population isolate. First two PCAs for the Jaccard similarity matrix. All chromosomes combined.

E Correcting a linear regression with global and regional PCAs

To assess the effects of regional substructure on association testing, we consider an example in the COPDGene study, a case-control study of Chronic Obstructive Pulmonary Disease (COPD) in current and former smokers (Regan E.A., 2010). The study has been sequenced as part of the TOPMED Project. The data is available through dbGaP (NHLBI TOPMed, 2018).

We examine the effect of the particular SNP *rs16969968* (chromosome 15) on FEV_1 which is a well-established risk locus for COPD and cigarette smoking (Pillai et al., 2009; Lutz, 2015). It is unclear whether the regional substructure or the global substructure is more relevant for a particular locus that is tested for association. It is important to note that the inclusion of additional principal components will not have a major impact on the power of the association analysis, given the current sample size of such studies. As a consequence, the analysis plan is to evaluate three regression models:

- Model 1: Regress FEV_1 on *rs16969968* adjusting for age, height, sex, and the first 5 global PCs;
- Model 2: Regress FEV_1 on *rs16969968* adjusting for age, height, sex, and the first 5 regional PCs that are computed for the region that harbors *rs16969968*;
- Model 3: Regress FEV_1 on *rs16969968* adjusting for age, height, sex, and the first 5 regional PCs and on the first 5 global PCs.

We will assess the association p-values for *rs16969968* on FEV_1 to evaluate whether the analysis benefits from the inclusion of the regional PCs.

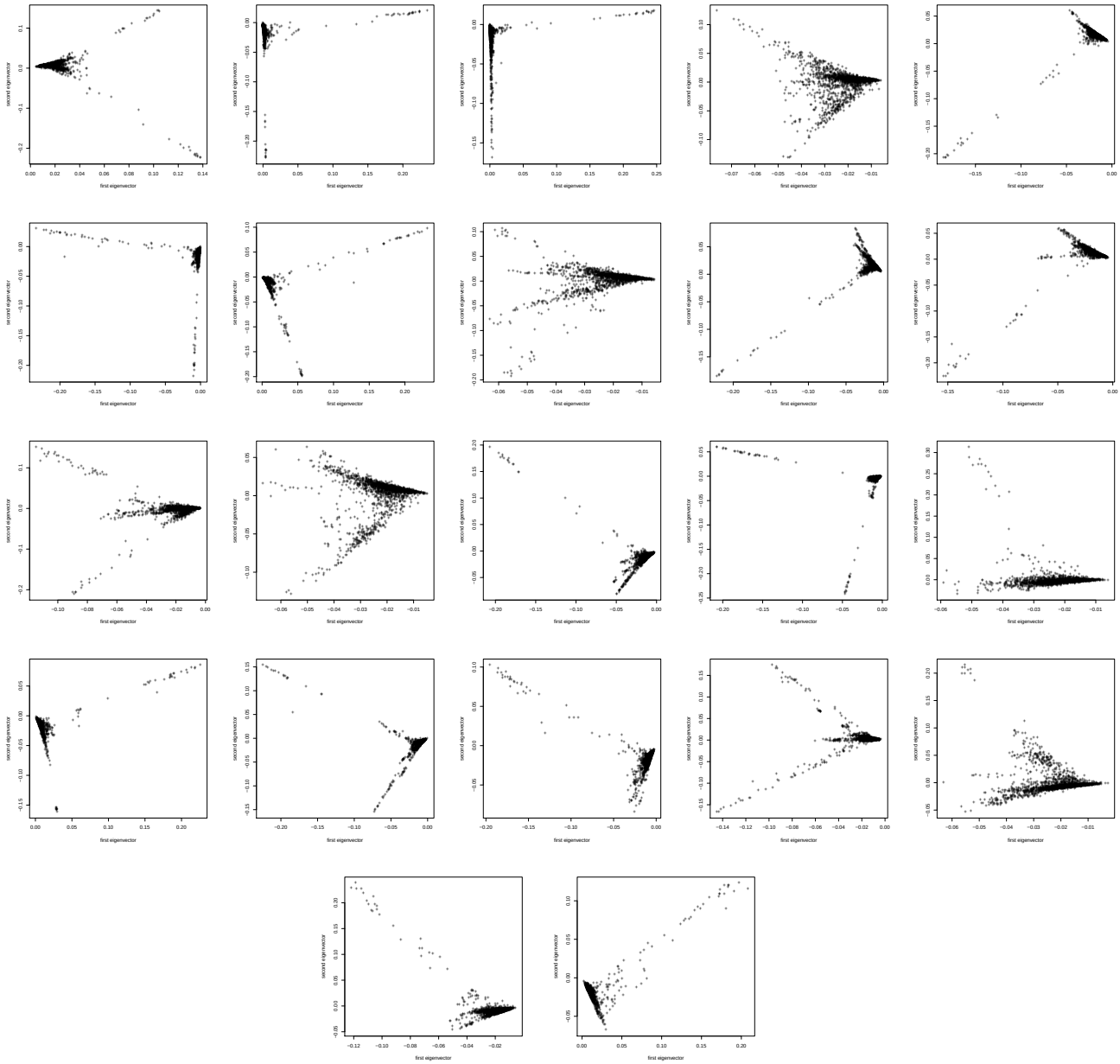

Figure 7: All 22 chromosomes of the Costa Rica population isolate. First two PCAs for the Jaccard similarity matrix computed for the middle window of a stratification scan with window size 10^5. Separate plot for each chromosome (starting with chromosome 1 in the top left corner and continuing in a row-wise fashion).

similarity matrix	window size	global	regional	global and regional
Cov	1600	1.81e-9	3.58e-9	1.44e-9
	3200	1.81e-9	2.92e-9	1.08e-9
	6400	1.81e-9	2.92e-9	1.08e-9
Jaccard	1600	1.82e-9	1.40e-9	9.90e-10
	3200	1.82e-9	1.53e-9	1.06e-9
	6400	1.82e-9	1.53e-9	1.06e-9
s-matrix	1600	1.29e-9	3.51e-9	8.61e-10
	3200	1.29e-9	3.76e-9	1.18e-9
	6400	1.29e-9	3.76e-9	1.18e-9
GRM	1600	8.61e-10	2.90e-09	5.33e-10
	3200	8.61e-10	2.82e-9	4.82e-10
	6400	8.61e-10	2.82e-9	4.82e-10

Table 3: Regression (1) on the FEV_1 for the Costa Rica population isolate. Columns show p-values for $\beta_{\text{S}} = 0$. Five principal components each for both global and regional adjustments.

We conducted the analysis in the following way. We first prepared the genetic data from the COPDGene study for chromosome 15 with PLINK2. We employed a maximal allele frequency cutoff of *--max-maf 0.01*, LD pruning with parameters *--indep-pairwise 2000 10 0.01*, and filtered out the snp of interest using the command *--snp rs16969968*. To specify regional windows around *rs16969968*, we employed *--window W*, where we chose the window size $W \in \{1600, 3200, 6400\}$. We made sure that due to its high allele frequency (maf=0.597), the snp *rs16969968* was indeed not included in any window. Since we compute our similarity matrices on rare variants having very different allele frequencies than the common ones, they are virtually uncorrelated with the common loci that are typically tested for association in single-locus analyses. After preparing the data with PLINK2, we are left with 5,765 subjects and 94,497 RVs.

Using the genetic data above and additionally the covariates *age*, *sex*, an indicator of current smoking status (*smoker*=1 for current smokers and 0 for former smokers), as well as the subject's *height* (in centimeters), we fit the regression models

$$\mathbb{E}\left(\text{FEV1}\right) = \beta_0 + \beta_{\text{S}}\text{SNP} + \beta_A\text{age} + \beta_S\text{sex} + \beta_{\text{SM}}\text{smoker} + \beta_H\text{height} + \sum_{i=1}^{5}\beta_i\text{PCA}_i, \qquad (1)$$

where *SNP* is the allele count data for *rs16969968*, and PCA_i for $i \in \{1, \ldots, 5\}$ are the first five principal components for either the global or regional similarity matrices. We test the hypothesis that $\beta_{\text{S}} = 0$ against the alternative that $\beta_{\text{S}} \neq 0$. The global PCAs are computed by applying any similarity matrix approach to the full genomic data, and computing the first eigenvectors. The regional PCAs are computed by extracting a region around *rs16969968* (of window size given in Table 3), computing the similarity matrix on that region, and then calculating the first eigenvectors of that similarity matrix.

In this way, we regress FEV_1 on the above covariates, where for global and regional adjustments we each use the first five PCAs. For the combined global and regional adjustment, we add both

the five global and the five regional eigenvectors to the model (1).

Results are given in Table 3. The table shows that for any of the four similarity matrices under investigation, and for any of the reported window sizes, the combined global and regional adjustment yields a p-value for the hypothesis $\beta_S = 0$ that is more significant than the global or regional adjustments alone. We therefore recommend to adjust for both "global" and "regional" PCs in genetic association testing. As we discussed above, in order to run the local scans, we propose here to compute the similarity matrices based on rare variants (RVs) which have very different allele frequencies than the common variants and are therefore virtually uncorrelated with the common loci that are typically tested for association in single-locus analyses. We believe therefore that the "proximal contamination" effects (Salter-Townshend and Myers, 2019; Gazal et al., 2018, 2017; Thornton and Bermejo, 2014; Baran et al., 2012; Listgarten et al., 2012) can largely be avoided here. If RVs are to be tested for association, they should be excluded from the computation of the regional similarity matrices. Given the computational efficiency of locStra, this does not create a bottleneck in terms of computation time.

Subsequent research here is also needed to evaluate the effects of other confounding variables, e.g. sequencing depth, batch effects, etc., on the globally and regionally computed similarity matrices. Approaches similar to the one developed for association testing (Sankararaman et al., 2008) could be utilized here.

References

Baran, Y., Pasaniuc, B., Sankararaman, S., Torgerson, D. G., Gignoux, C., Eng, C., Rodriguez-Cintron, W., Chapela, R., Ford, J. G., Avila, P. C., Rodriguez-Santana, J., Burchard, E. G., and Halperin, E. (2012). Fast and accurate inference of local ancestry in latino populations. *Bioinformatics*, 28(10):1359–1367.

Bates, D. and Eddelbuettel, D. (2013). Fast and Elegant Numerical Linear Algebra Using the RcppEigen Package. *J Stat Softw*, 52(5):1–24.

Bodmer, W. and Bonilla, C. (2008). Common and rare variants in multifactorial susceptibility to common diseases. *Nat Genet*, 40:695–701.

Chang, C., Chow, C., Tellier, L., Vattikuti, S., Purcell, S., and Lee, J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*, 4.

Devlin, B. and Roeder, K. (1999). Genomic control for association studies. *Biometrics*, 55(4):997–1004.

Gazal, S., Finucane, H., Furlotte, N., Loh, P., Palamara, P., Liu, X., Schoech, A., Bulik-Sullivan, B., Neale, B., Gusev, A., and Price, A. (2017). Linkage disequilibrium-dependent architecture of human complex traits shows action of negative selection. *Nat Genet*, 49(10):1421–1427.

Gazal, S., Loh, P., Finucane, H., Ganna, A., Schoech, A., Sunyaev, S., and Price, A. (2018). Functional architecture of low-frequency variants highlights strength of negative selection across coding and non-coding annotations. *Nat Genet*, 50(11):1600–1607.

Hunninghake, G. M., Soto-Quiros, M. E., Avila, L., Ly, N. P., Liang, C., Sylvia, J. S., Klanderman, B. J., Silverman, E. K., and Celedón, J. C. (2007). Sensitization to Ascaris lumbricoides and severity of childhood asthma in Costa Rica. *Journal of Allergy and Clinical Immunology*, 119(3):654–661.

Keinan, A. and Clark, A. (2012). Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science*, 11(336(6082)):740–3.

Kryukov, G. V., Shpunt, A., Stamatoyannopoulos, J. A., and Sunyaev, S. R. (2009). Power of deep, all-exon resequencing for discovery of human trait genes. *PNAS*, 106(10):3871–3876.

Laird, N. M. and Lange, C. (2010). *The fundamentals of modern statistical genetics*. Springer Science & Business Media.

Lee, S., Epstein, M., Duncan, R., and Lin, X. (2012). Sparse principal component analysis for identifying ancestry-informative markers in genome-wide association studies. *Genet Epidemiol*, 36(4):293–302.

Listgarten, J., C, L., CM, K., RI, D., E, E., and D, H. (2012). Improved linear mixed models for genome-wide association studies. *Nat Methods*, 9(6):525–6.

Lutz, S.M, C.-M. e. a. (2015). A genome-wide association study identifies risk loci for spirometric measures among smokers of European and African ancestry. *BMC Genet*, 16(138):1–11.

Martin, E. R., Tunc, I., Liu, Z., Slifer, S. H., Beecham, A. H., and Beecham, G. W. (2018). Properties of Global and Local Ancestry Adjustments in Genetic Association Tests in Admixed Populations. *Genet Epidemiol*, 42(2):214–229.

Morrison, A., Voorman, A., Johnson, A., Liu, X., Yu, J., Li, A., and J., B. (2013). Whole genome sequence-based analysis of a model complex trait, high density lipoprotein cholesterol. *Nat Genet*, 45(8):899–901.

NHLBI TOPMed (2018). Boston Early-Onset COPD Study in the National Heart, Lung, and Blood Institute (NHLBI) Trans-Omics for Precision Medicine (TOPMed) Program. `https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000946.v3.p1`.

NHLBI TOPMed (2019). The Genetic Epidemiology of Asthma in Costa Rica. `https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000988.v3.p1`.

Patterson, N., Price, A., and Reich, D. (2006). Population Structure and Eigenanalysis. *PLoS Genet*, 2(12):e190.

Pillai, S. G., Ge, D., Zhu, G., Kong, X., Shianna, K. V., Need, A. C., Feng, S., Hersh, C. P., Bakke, P., Gulsvik, A., Ruppert, A., Lødrup-Carlsen, K. C., Roses, A., Anderson, W., ICGN-Investigators, Rennard, S. I., Lomas, D. A., Silverman, E. K., and Goldstein, D. B. (2009). A Genome-Wide Association Study in Chronic Obstructive Pulmonary Disease (COPD): Identification ofTwo Major Susceptibility Loci. *PLoS Genetic*, 5(3):e1000421.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*, 38:904–909.

Price, A. L., Tandon, A., Patterson, N., Barnes, K. C., Rafaels, N., Ruczinski, I., Beaty, T. H., Mathias, R., Reich, D., and Myers, S. (2009). Sensitive Detection of Chromosomal Segments of Distinct Ancestry in Admixed Populations. *PLoS Genet*, 5(6):e1000519.

Pritchard, J., Stephens, M., Rosenberg, N., and Donnelly, P. (2000). Association mapping in structured populations. *Am J Hum Genet*, 67(1):170–181.

Prokopenko, D., Hecker, J., Silverman, E. K., Pagano, M., Nöthen, M. M., Dina, C., Lange, C., and Fier, H. L. (2016). Utilizing the Jaccard index to reveal population stratification in sequencing data: a simulation study and an application to the 1000 Genomes Project. *Bioinformatics*, 32(9):1366–1372.

Purcell, S. and Chang, C. (2019). PLINK2.

Regan E.A., e. a. (2010). Genetic epidemiology of COPD (COPDGene) study design. *COPD*, 7(7):32–43.

Salter-Townshend, M. and Myers, S. (2019). Fine-Scale Inference of Ancestry Segments Without Prior Knowledge of Admixing Groups. *Genetics*, 212(3):869–889.

Sankararaman, S., Sridhar, S., Kimmel, G., and Halperin, E. (2008). Estimating Local Ancestry in Admixed Populations. *Am J Hum Genet*, 82(2):290–303.

Schlauch, D. (2016). Implementation of the stego algorithm - Similarity Test for Estimating Genetic Outliers.

Schlauch, D., Fier, H., and Lange, C. (2017). Identification of genetic outliers due to sub-structure and cryptic relationships. *Bioinformatics*, 33(13):1972–1979.

The 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. *Nature*, 526(68–74).

Thornton, T. A. and Bermejo, J. L. (2014). Local and Global Ancestry Inference, and Applications to Genetic Association Analysis for Admixed Populations. *Genet Epidemiol*, 38(0 1):S5S12.

von Mises, R. and Pollaczek-Geiringer, H. (1929). Praktische Verfahren der Gleichungsaufloesung. *ZAMM Zeitschrift fr Angewandte Mathematik und Mechanik*, 9:152–164.

Wang, B., Sverdlov, S., and Thompson, E. (2017). Efficient Estimation of Realized Kinship from Single Nucleotide Polymorphism Genotypes. *Genetics*, 205(3):1063–1078.

Yang, J., Lee, S., Goddard, M., and Visscher, P. (2011). GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet*, 88(1):76–82.

Yazdani, A., Yazdani, A., and Boerwinkle, E. (2015). Rare variants analysis using penalization methods for whole genome sequence data. *BMC Bioinformatics*, 16(1):405.

Zhong, Y., Perera, M. A., and Gamazon, E. R. (2019). On Using Local Ancestry to Characterize the Genetic Architecture of Human Traits: Genetic Regulation of Gene Expression in Multiethnic or Admixed Populations. *Am J Hum Genet*, 104(6):1097–1115.