SPC: a SPectral Component approach leveraging Identity-by-Descent graphs to address recent population structure in genomic analysis

Ruhollah Shemirani[1], Gillian M. Belbin[1*], Sinead Cullina[1,2], Christa Caggiano[1], Christopher Gignoux[3,4], Noah Zaitlen[5,6,7], Eimear E. Kenny[1,2,8]

1. *Institute for Genomic Health, Icahn School of Medicine at Mount Sinai, New York, NY, USA*
2. *Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA*
3. *Colorado Center for Personalized Medicine, University of Colorado Anschutz Medical Campus, Aurora, CO, USA*
4. *Department of Biostatistics and Informatics, University of Colorado Anschutz Medical Campus, Aurora, CO, USA*
5. *Department of Neurology, University of California Los Angeles, Los Angeles, CA, USA*
6. *Department of Human Genetics, University of California Los Angeles, Los Angeles, CA, USA*
7. *Department of Computational Medicine, University of California Los Angeles, Los Angeles, CA, USA*
8. *Department of Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA*

** Current affiliation: Gencove, New York, NY*

Abstract

Population structure is a well-known confounder in statistical genetics, particularly in genome-wide association studies (GWAS), where it can lead to inflated test statistics and spurious associations. Traditional methods, such as principal components (PCs), commonly used to adjust for population structure, are limited in capturing fine-scale, non-linear patterns that arise from recent demographic events – patterns that are crucial for understanding rare variant effects. To address this challenge, we propose a novel method called SPectral Components (SPCs), which leverages identity-by-descent (IBD) graphs to capture and transform local, non-linear fine-scale population structure into continuous representations that can be seamlessly integrated into genetic analysis pipelines. Using both simulated datasets and empirical data from the UK Biobank (N ≈ 420,000), we demonstrate that SPCs outperform PCs in adjusting for fine-scale population structure. In simulations, SPCs explained over 90% of the fine-scale population structure with fewer components, while PCs captured less than 5%. In the UK Biobank, SPCs reduced the inflation of p-values in the GWAS of an environmental-driven phenotype by 12% compared to

PCs, while maintaining a similar performance to PCs in height, a highly heritable phenotype. Additionally, SPCs improved rare variant association analyses, reducing genomic inflation (e.g., from 7.6 to 1.2 in one analysis), and provided more accurate heritability estimates. Spatial autocorrelation analysis further confirmed the ability of SPCs to account for environmental effects, reducing Moran's I for both environmental and heritable phenotypes more effectively than PCs. Overall, our findings demonstrate that SPCs provide a robust, scalable adjustment for recent population structure, offering a powerful alternative or complement to PCs in large-scale biobank studies.

Keywords: GWAS, population structure, principal components, identity-by-descent, spectral graph theory, UK Biobank, heritability, fine-scale population structure, genomic inflation

Introduction

Population structure is a critical factor in statistical genetic studies, as it can confound results if not properly addressed (Price et al. 2006; Lawson et al. 2020). This structure arises from genetic and environmental differences across populations, shaped by historical processes such as migration (Gravel et al. 2013), genetic drift (Nei and Tajima 1981), selection (Berg et al. 2019; Sohail et al. 2019; Mathieson et al. 2023), and assortative mating (Border et al. 2022). These processes create systematic differences in variant frequencies within and between populations that can violate key statistical assumptions in genetic analyses (Sohail et al. 2019; Price et al. 2010). For example, population structure can violate the assumption of independence among individuals in genetic analyses, such as Genome Wide Association Studies (GWAS), leading to false-positive associations and inflated test statistics (Cook et al. 2020). To mitigate these effects, genetic analysis pipelines often use Principal Components (PCs; Price et al. 2006) and/or (generalized) linear mixed models (GLMMs; Kang et al. 2010). PCs represent major axes of genetic variation derived from dimensionality reduction of a genotype matrix, or the Genetic Relatedness Matrix (GRM), which stores pairwise genetic similarity based on variants. GLMMs

complement this approach by simultaneously accounting for population structure and cryptic relatedness by modeling a random effect whose covariance structure is defined by the estimated GRM. Both PC- and GLMM-based strategies have been successfully used to address population structure in analysis involving common variants. However, it is less clear whether these methods will be similarly effective for rare variants and more complex fine-scale population structure (Mathieson and McVean 2012; Zaidi and Mathieson 2020; Privé et al. 2020).

Large-scale sequencing studies have revealed patterns of recently arisen rare variants and have shown that they tend to be geographically localized (Fu et al. 2013; Gravel et al. 2011). This fine-scale population structure impacting rare variants has been detected in several populations, including Icelandic (Helgason et al. 2005), British (Cook et al. 2020; Gilbert et al. 2022; Nait Saada et al. 2020), and Japanese (Sakaue et al. 2020), revealing distinct genetic patterns shaped by historical movements over the past few thousand years. However, these studies also highlight a critical challenge: traditional methods to account for population structure struggle to capture this discrete, non-linear population structure occurring within recent timescales. The non-linear relationships and sparsity in the GRM introduced by rare variants can make rare variant PCs sensitive to noise and violate the linear assumptions of PCA, complicating the interpretation of results (Baye et al. 2011; Ma and Shi 2020; Zaidi and Mathieson 2020). Likewise, incorporating rare variant information in GLMMs introduces several limitations, including higher computational costs and reduced power compared to common variant counterparts (S. Lee et al. 2014; Wainschtein et al. 2022). These findings underscore the need for novel analytical methods tailored to the unique properties of rare variants in statistical genetic analyses, enabling more accurate modeling of fine-scale population structure and the complex genetic architecture they reveal.

To address these limitations, estimates of pairwise haplotypes shared identical-by-descent, or Identity-By-Descent (IBD), offers a more precise approach for detecting recent fine-scale population structure in large genomic datasets (Shemirani et al. 2021). The length of shared IBD segments provides a direct measure of evolutionary time – longer segments indicate more recent population structure, typically within the past dozen generations (Browning and Browning 2015). At the population level, the aggregated lengths of IBD segments shared between all pairs of individuals across the genome can be utilized to construct graphs of genetic similarity, where individuals are represented as nodes and shared IBD haplotypes are represented as edges (Belbin et al. 2021). Previous studies have shown the advantages of utilizing IBD data to represent recent population structure through the extraction of clusters of individuals who share IBD haplotypes across the genome (Belbin et al. 2021; Caggiano et al. 2023; Dai et al. 2020). These IBD-based clusters reveal population structure missed by traditional methods, leading to differential associations with health outcomes, unique enrichment of casual rare variants, and distinctive performance of polygenic risk scores compared to clusters identified with common variants alone (Belbin et al. 2021; Johnson et al. 2022).

Here, we introduce SPectral Components (SPCs), a graph theory-driven approach that transform the discrete, localized signals of recent population structure, represented by an IBD graph, into a continuous representation of genetic similarity that can be used in statistical genetic analysis pipelines. Our method leverages graph Laplacian transformations (Grone et al. 1990), a fundamental concept in computational fields, such as dimensionality reduction (Kunegis et al. 2010), graph-based machine learning (Si et al. 2014), and network analysis (Shuman et al. 2013), due to their ability to encode graph topology and smooth representations of variations in node features across the graph (Shuman et al. 2013). Previous work has applied spectral approaches to common variant data and demonstrated they can be more robust than PCs in finding hidden population structure compared to GRMs (A. B. Lee et al. 2010). In this study, we use simulated

and empirical data from the UK Biobank project to explore the efficacy of SPCs in adjusting for the effects of population structure compared to traditional methods. We also examine the performance of these approaches in rare variant analysis under environmental confounding. We demonstrate how SPCs optimally extract patterns of population structure that are resilient to the presence of discrete and localized structure in both variant and environmental association, via leveraging graph spectral decomposition.

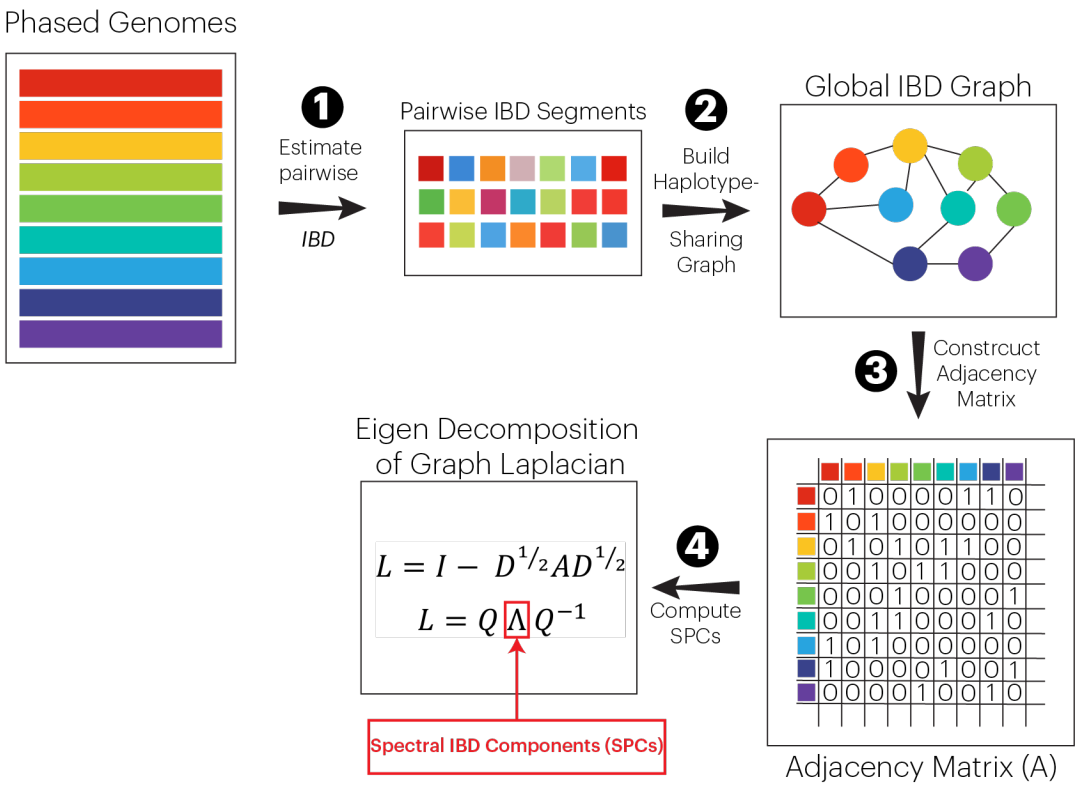


Figure 1 – General schema for calculating SPCs. First, phased genotypes are used to estimate segments of genome shared between individuals IBD. Second, the IBD segments are utilized to generate a relatedness graph with vertices representing participants and edges representing aggregated IBD sharing between pairs of participants above a threshold of 6cM genome wide. Third, the IBD relatedness graph is transformed into an adjacency matrix (A). Finally, the adjacency matrix A is transformed into its Laplacian form. The eigenvectors corresponding to the smallest eigenvalues of this matrix comprise the set of SPCs.

Results

Overview of Spectral Components (SPC) Calculation

Previous work has demonstrated the efficacy of identity-by-descent (IBD) graphs in capturing fine-scale population structure signals absent from GRMs (Zaidi and Mathieson 2020; Caggiano et al. 2023; Browning and Browning 2023). We developed a pipeline to extract individuals frequency components underlying a graph of IBD sharing at a population level, or Spectral Components (SPCs; **Figure 1**). SPCs were calculated in 4 steps. First, we phased genotype data, and estimated IBD using iLASH (Shemirani et al. 2021), We then aggregated the total sums of IBD sharing for each pair of individuals, which were then utilized to generate an undirected, unweighted relatedness graph, with each participant represented as a node. IBD relatedness in this graph was represented as connecting edges between respective nodes, if their sum of IBD sharing surpassed a minimum threshold of 6cM. Thirdly, we generated an adjacency matrix representation of this graph. Lastly, we calculated the spectral components of this matrix as SPCs. To conduct comparative analysis, we further calculated alternative principal components of this graph, along with the principal components of the original genotype data, and rare variants.

Evaluating Performance to Detect Fine-scale Population Structure on a Simulated Dataset

To evaluate the performance of SPCs for capturing fine-scale population structure in genomic datasets, we first simulated a population, following Zaidi & Mathieson, 2020. We simulated a 5-by-5 grid of 25 demes with a low migration rate among them that emulates a homogeneous population structure (**Figure 2a**). This configuration closely resembles the recent fine-scale

population structure observed in the White British cohort in the UK Biobank (see Methods). By simulating this nuanced structure, we aimed to evaluate how well SPCs can discern subtle genetic differences that are often overlooked by traditional methods, thereby providing a more detailed understanding of population stratification.

The simulated dataset included 50,000 diploid samples. Only the first chromosome was simulated for computational tractability. We calculated PCs of common and rare variants. We then down sampled the dataset to retain only the common variants, phased the data to replicate phasing artifacts, and estimated aggregated IBD sharing using iLASH.

Compared to PCs, SPCs demonstrate significant improvements in capturing population structure (**Figure 2b-c**). We used correlation with deme of origin as a basic measurement of the ability of SPCs to dissect fine-scale population structure. We indexed demes based on their order on horizontal and vertical axes to generate coordinates of origin for each simulated sample, akin to place of birth for an individual. The first 100 PCs explain 4% and 5% of the variation in the vertical and horizontal coordinates of the deme of origin for simulated samples, respectively. However, the first 10 SPCs were sufficient to explain 90% of the variation in the same categories of data. This is reflected in better visual separation between samples based on their deme of origin.

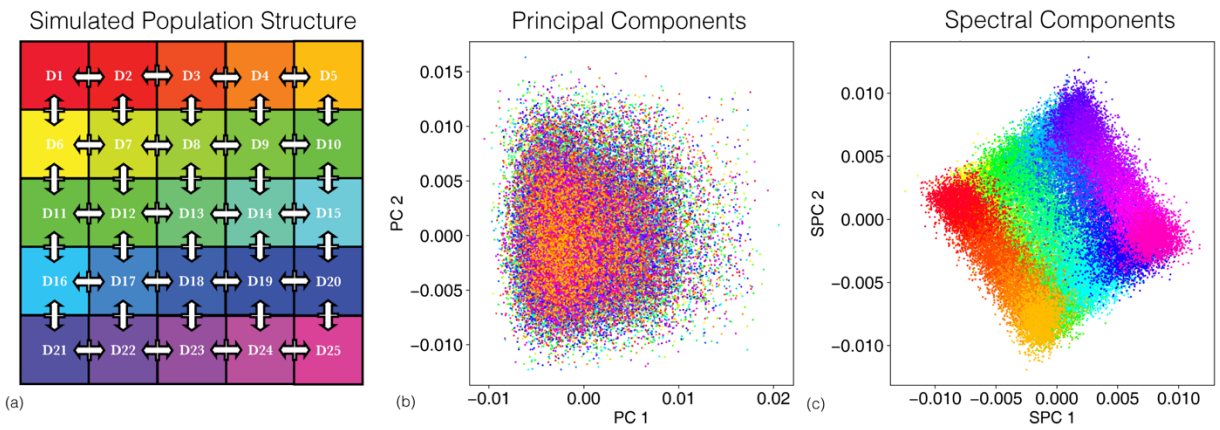


Figure 2- Reconstruction of recent population structure by covariates. **a**) A schematic of our simulation of a cohort with homogeneous population structure with 25 demes and constant migration (5% per generation) between neighboring demes. All demes coalesce to single ancestral deme 200 generations ago. Consequently, simulated individuals have similar genetic backgrounds in terms of common variants, while simultaneously harboring distinct structure in terms of rare variation and environment of origin. **b, c**) PC and SPC representation of the shared genetic ancestry among simulated samples. Dots in either panel represent samples colored based on the deme of origin.

Evaluating Performance for Disentangling Environmental and Genetic effects

To evaluate the performance of SPCs as covariates in genomic analysis pipelines, we simulated 7 continuous phenotypes (**Figure 3a**), two of which were non-heritable and 5 with different heritability schemes. The first, referred to as 'environmental smooth', was a continuous phenotype in which variation solely originated from indirect environmental effects. The mean value of the phenotype decreased from North to South (**Figure 3a**), and the effects of linear dependencies on deme of origin among the mean values of outcomes were observed. The second, which Zaidi and Mathieson called 'environmental sharp'(Zaidi and Mathieson 2020), aimed to investigate the

202 effects resulting from the invalidity of common linear assumptions (**Figure 3a**). For this phenotype,
203 a single target deme had a phenotype distribution of $y \sim \mathcal{N}(\mu, 2\sigma)$ with a non-zero μ, while other
204 demes had a phenotype distribution of $y \sim \mathcal{N}(0, \sigma)$; mimicking non-linear indirect environmental
205 effects. Due to variation stemming only from environmental effects, the expected narrow-sense
206 heritability of both environmental phenotypes simulated is zero.
207
208 The next 3 phenotypes, on the other hand, were heritable (h^2=0.1, 0.3, 0.8), and were affected
209 only by direct genetic effects and random noise. These phenotypes, which we refer to as
210 "polygenic phenotypes", are modeled through the randomized selection of causal variants with
211 effect sizes assigned based on their allele frequency and the total heritability of the trait. This
212 approach of simulating effect sizes results in negligible correlation with the deme of birth, as it
213 does not allocate a high impact to common variants that undergo significant drift, subsequently
214 impacting mean phenotype values per deme (Schoech et al. 2019). Absence of such confounding
215 effects eliminate both the biases described earlier and the requirement to adjust for them. Thus,
216 fixed-effect covariates, such as PCs, are not effective in addressing possible confounding effects
217 for this phenotype. Instead, Linear Mixed Models can be used in such scenarios to account for
218 variance components induced by genetic effects (Price et al. 2010).
219
220 The last two phenotypes, henceforth referred to as "hybrid" phenotypes, combine environmental
221 and polygenic phenotypes via replacing the random noise components of the polygenic
222 phenotype with environmental phenotypes. The hybrid smooth phenotype is generated via adding
223 the polygenic phenotype with heritability 0.3 and the environmental smooth phenotype
224 (reweighted to comprise 0.7 of the variance of the phenotype). The hybrid sharp phenotype is
225 generated similarly using the environmental sharp with the same weighting scheme. The
226 environmental category was designed to evaluate Type-1 error rate, while the polygenic

227 phenotypes assess power. Finally, the hybrid phenotypes evaluate how well-calibrated each
228 model is in the presence of both background polygenic effects and confounding environmental
229 effects.
230
231 To measure the efficacy of SPCs in adjusting for environmental effects on phenotypes, we used
232 proportion of variation in the phenotype explained by covariates (PVE; **Figure 3b, Supplemental**
233 **Table 1**). For the smooth phenotype, the first 25 SPCs accounted for 83% of the PVE, while the
234 same number of PCs explained only 0.7%. For the sharp phenotype, the PVE for both models
235 decreased, although the SPCs still explained 40-fold more environmental effects than PCs (2.5%
236 of PVE vs 0.06%). SPCs achieved their lowest PVE in the analysis of the heritable phenotypes
237 with no environmental effects from recent population structure, where both models explained less
238 than 0.5% in the variation of the phenotypes across all heritability models (**Supplemental Figure**
239 **1**). PVE values from the PC-based model were slightly higher than the SPC-based models (0.1%
240 vs. 0.4%, respectively). Analyzing the hybrid smooth phenotypes as outcomes resulted in a PVEs
241 of 58% and 0.7% for the SPC-based, and PC-based models respectively. PVEs for the hybrid
242 sharp phenotype were 1.7% and 0.2% for the two models, respectively (**Supplemental Figure**
243 **1**). Overall, these simulations demonstrate that SPCs more effectively capture environmental
244 effects compared to PCs, while not significantly impacting estimates of true heritable effects.
245
246 **Evaluating Performance for Genome Wide Association Studies**
247
248 To evaluate how SPCs perform in the context of genome-wide association studies, we conducted
249 GWAS using all 7 phenotypes, focusing on the distribution of z-scores and the inflation of p-values
250 as a measure of the unadjusted effects of population structure (Devlin and Roeder 1999; Yang et
251 al. 2011) (**Figure 3c**). In the GWAS of common variants for the smooth phenotype, the genomic
252 inflation factor (denoted by λ_{gc}) was 1.101 [confidence intervals (CI): 1.084-1.118] when including

SPCs in the model, compared to 5.410 [CI: 5.305-5.517] when including PCs in the model, indicating that SPCs were better controlling for population structure. The difference in performance in the sharp model was less pronounced, although statistically significant, 1.102 [CI: 1.066-1.138] for the PC-based models and 1.037 [CI: 1.019-1.054] for the SPC-based ones. This pattern was not observed in the inflation of p–values in the models in the GWAS of the polygenic phenotype, where the credible interval for the λ_{gc} overlapped each other for all heritability rates (**Supplemental Figure 2c**). In the GWAS of hybrid phenotypes, where the environmental noise has non-random structure, λ_{gc} was significantly higher in the PC-adjusted model (λ_{gc}= 4.498 [CI: 4.355-4.648]) compared to the SPC-adjusted model (λ_{gc}=1.581 [CI: 1.524-1.635]) for the hybrid smooth phenotypes (**Supplemental Figure 3c**). λ_{gc} for both models increased compared to the scenario with same narrow-sense heritability (h=0.3) and random environmental noise. While SPCs had a lower λ_{gc} in the GWAS of hybrid sharp phenotype (λ_{gc}=1.321 [CI: 1.287-1.355]) compared to PCs (λ_{gc} =1.365 [CI: 1.334-1.396]), their credible intervals overlapped.

To better understand the impact of population structure in the analysis of the polygenic phenotype, we stratified variants into causal and non-causal, referring to variants neither truly associated with the phenotype nor in Linkage Disequilibrium (LD) with causal variants. We assumed that any systematic differences in calculated p-values or effect sizes of non-causal variants in between the two models are likely due to population structure rather than true genetic association. Among the non-causal variants, p-values calculated using PCs were consistently lower (Wilcoxon P-value: 1.54×10^{-53}), which suggests they are more affected by confounding. However, the mean effect size estimates for these variants were higher for the SPC models; suggesting the lower p-values in the PC model were not due to higher power but may rather reflect differences in how population structure is accounted for. For causal variants and variants in linkage disequilibrium with causal variants, there were no differences in estimated effect sizes using SPCs and PCs.

We used the hybrid phenotypes to evaluate the calibration of each model's effect size estimate against the ground truth simulated effects. Looking exclusively at estimated effect sizes for variants in high LD (r^2 > 0.1) with the causal variants in the GWAS of medium heritability phenotype, both models showed similar accuracy in estimating the true simulated effect sizes. (**Supplemental Figure 4a**). Using the same effect sizes along with non-random environmental noise, in the form of the hybrid smooth phenotype, resulted in drastically lower accuracy in the PC-adjusted model, while not significantly affecting the accuracy SPC adjusted model (**Supplemental Figure 4b**). Conversely, when looking at variants in low LD (r^2 < 0.1) with the causal variants, under random environmental noise, PC-adjusted model found two false-positive variants with a p-value lower than the multiple-testing threshold (p-value $< 5 \times 10^{-8}$), whereas the SPC-adjusted model found one. After adding structured noise in the hybrid smooth models, SPC-adjusted model highlighted the same variant, while the PC-adjusted model highlighted 1874 false-positive variants that passed the multiple testing threshold.

To explore the efficacy of SPCs in adjusting for confounding in rare variant analysis, we conducted GWAS of rare variants with all simulated phenotypes. The GWAS of rare variants showed similar λ_{gc} trends to the GWAS of common variants (**Figure 3d**) across all phenotypes, with performance gap decreasing due to lower power, with the exceptions of smooth and hybrid smooth phenotypes. In GWAS of the smooth phenotype, λ_{gc} of the PC model increased significantly to 7.576 [CI: 7.514-7.639], while λ_{gc} of the SPC model exhibited a modest increase to 1.119 [CI: 1.114-1.124]. Similarly, in the analysis of hybrid smooth phenotype, we estimated λ_{gc} values of 5.480 [CI: 5.336-5.630] and 1.162 [CI: 1.157-1.167] for the PC-adjusted and the SPC adjusted models respectively.

As an additional analysis, we used the environmental phenotypes with no underlying causal genetic variation to calculate false-positive rates in the GWAS of common and rare variants. In

the GWAS of the smooth phenotype, adjustment using PCs resulted in 1918.9 [std:22.38] false-positives variants passing the Bonferroni multiple testing threshold (5×10^{-8}) in the analysis of common variants and 12042.35 [std:71.32] false-positives in the analysis of rare variants. Adjustment by SPCs resulted in 3.6 [std:4.86] and 0.3 [std:0.56] false positives in the analysis of common and rare variants, respectively, suggesting that the SPC method has significantly lower Type I error rates in the presence of population structure confounding. These results demonstrate the robustness of SPC-adjustment to both rare and common allele frequency variants.

Comparison against alternative strategies of covariate extraction

We next compared SPCs against alternative strategies to calculate fixed-effect covariates to adjust for recent population structure. Previous work by Zaidi and Mathieson demonstrated that PCs derived from rare variants can partially alleviate the limitations of common PCs (Zaidi and Mathieson 2020). In our simulations, SPCs outperformed PCs of rare variants both in terms of predictive performance (R^2 score) and decreasing the inflation of p-values. While PCs of rare variants had a higher R^2 score with the environmental phenotypes compared to PCs of common variants, their performance was consistently lower than that of SPCs (**Supplemental Figure 1, Supplemental Table 1**).

Additionally, principal components derived from the IBD relatedness matrix were previously used to improve representation of recent population structure compared to PCs in simulations (Zaidi and Mathieson 2020). In line with those findings, we explored alternative strategies for the calculation of continuous covariates using the IBD relatedness matrix for comparison with SPCs. We calculated 11 additional sets of covariates from IBD relatedness data, comparing principal components to spectral components and exploring a range of possible minimum IBD sharing thresholds in the generation of the relatedness matrix (6cM, 10cM, and 15cM). We found SPCs,

330 calculated using a threshold of 6cM, to be among the best performers in terms of PVE of simulated
331 phenotype (**Supplemental Figure 1**). We also repeated the SPC-adjusted GWAS using two
332 higher IBD sharing thresholds (10cM and 15cM) and found that 6cM threshold yielded λ_{gc} values
333 that were equal or lower than those of higher thresholds (**Supplemental Figure 5**). Overall, our
334 results demonstrate that SPCs provide a more accurate adjustment for recent population structure
335 than alternative covariate strategies (further details in Supplemental Material).
336

337 **Evaluating performance in narrow-sense heritability estimation**

338 Finally, we evaluated the impact of SPCs on estimating narrow-sense heritability, the proportion
339 of phenotypic variance that is due to a genetic component (Yang et al. 2017), where adjustment
340 of phenotype values using PCs is a common measure to lower the inflation of estimates due to
341 population structure (Abdellaoui et al. 2019; Cook et al. 2020). We calculated the narrow-sense
342 heritability of the smooth phenotype using the GREML method implemented in the GCTA software
343 package along with a GRM calculated based on common variants. This resulted in a heritability
344 estimate of 1.00, indicating the phenotype is fully heritable genetically (**Supplemental Figure 6**).
345 Adjustment using PCs of rare variants resulted in a heritability estimate of 0.45. However,
346 adjustment using SPCs resulted in a 12-fold decrease of the estimate to 0.08, suggesting a
347 significantly lower genetic heritability component, which is in line with the non-heritability of the
348 phenotype. This illustrates the optimal effectiveness of SPCs against PCs when analyzing
349 phenotypes with strong environmental components.
350

351 To test for overcorrection of genetic factors that leads to the underestimation of narrow-sense
352 heritability (Evans et al. 2018), we calculated the heritability of the polygenic phenotypes. Across
353 all heritability values, the estimated credible intervals of narrow-sense heritability for both
354 approaches overlapped with each other (**Supplemental Table 2**). Both approaches yielded
355 heritability estimates similar to the unadjusted heritability estimate, demonstrating that the mixed

effects model deployed by GCTA adequately addresses population structure in these scenarios. We simulated another heritable phenotype with high heritability (0.8) and lower polygenicity to explore the effects of lower polygenicity on heritability estimates. There, uncorrected estimated heritability for this phenotype was 0.43 (std:0.04). The estimated heritability after correcting for SPCs did not significantly change (0.42-std:0.04). However, correcting for PCs decreased the heritability estimate to 0.36 (std:0.04) (**Supplemental Figure 7**). Together, this suggests that SPCs have advantages over PCs in accurately estimating heritability in the context of population stratification.

Measuring the effects of upstream noise

Artifacts of noise and errors in IBD estimation pipelines, including phasing switch errors, false-positive IBD segments, missing IBD segments, and incorrect IBD segment boundaries negatively affect the quality of IBD estimation and, consequently, global sharing estimates. For example, we previously showed that the boundaries of IBD segments might be underestimated (Shemirani et al. 2021). We measured the effects of these artifacts on the performance of SPCs in terms of PVE by comparing SPCs derived from inferred IBD (estimated SPCs) against SPCs calculated using ground-truth IBD data extracted directly from the simulated ancestral recombination graphs (ARGs; c.f. Methods). We calculated the PVE in the hybrid smooth phenotype by the first 25 PCs, SPCs, and ground-truth SPCs (**Supplemental Figure 8**). While ground-truth SPCs outperformed the estimated SPCs, their performance was qualitatively similar, with PVEs of 0.59 vs. 0.58 for ground-truth and estimated SPCs, respectively, suggesting that upstream artifacts do not cause serious hindrance to the performance of SPCs.

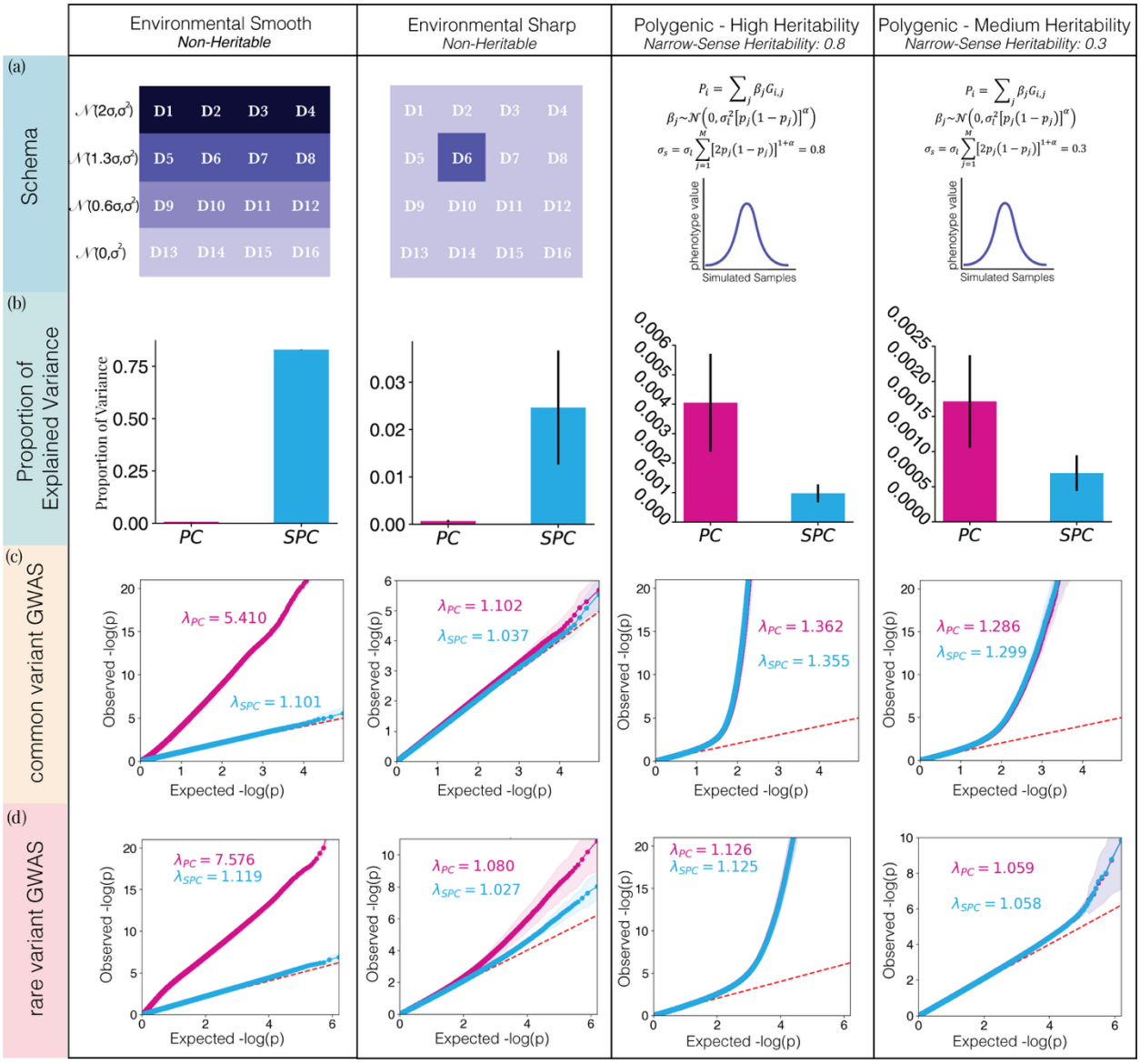


Figure 3 – Comparison of PCs and SPCs as covariates to correct for population structure on the outcomes in simulations. **a)** The schema of simulated phenotypes. First, the environmental smooth phenotype, where the values are determined based on the vertical coordinate of origin for each individual. The phenotype values are drawn from a normal distribution with shared standard deviation. The mean of the distribution in each row is lower than the row above, resulting in a difference of twice the standard deviation between the top and the bottom row. Second, the environmental sharp phenotype is determined by the deme of origin. The phenotype values for all demes other than D6 is drawn from a zero-centered normal distribution, whereas in D6,

390 phenotype values are drawn from a normal distribution with the mean set to twice the standard
391 deviation. Finally, the two polygenic phenotypes were determined by assigning effect sizes to
392 causal variants, randomly selected from windows of 10,000 base pairs, and calculating the
393 polygenic score based on the effect sizes. Effect sizes were drawn from a random normal
394 distribution with a mean of zero and a variance derived from minor allele frequency, heritability,
395 and selective pressure. σ_s determines the expected heritability of the phenotype, while α
396 determines the selective pressure, with negative values resulting in higher effect sizes assigned
397 to variant with low minor allele frequency. Tested values for σ_s were 0.8 for the high heritability
398 phenotype and 0.3 for the medium heritability phenotype. α was set to -0.5 across both
399 experiments **b)** proportion of variation in the phenotypes explained by the first 25 PCs and SPCs
400 in 4 categories of simulated phenotypes. Adjustment methods are listed on the x-axes, while the
401 y-axes represent proportion of variance explained by each model. The error bars represent the
402 standard deviation. **c)** Inflation of p-values in a GWAS analysis of simulated phenotypes using
403 common variants only. **d)** Inflation of p-values in a GWAS analysis of simulated phenotypes using
404 rare variants (10<MAC<100).

405 **Analysis of the UK Biobank**

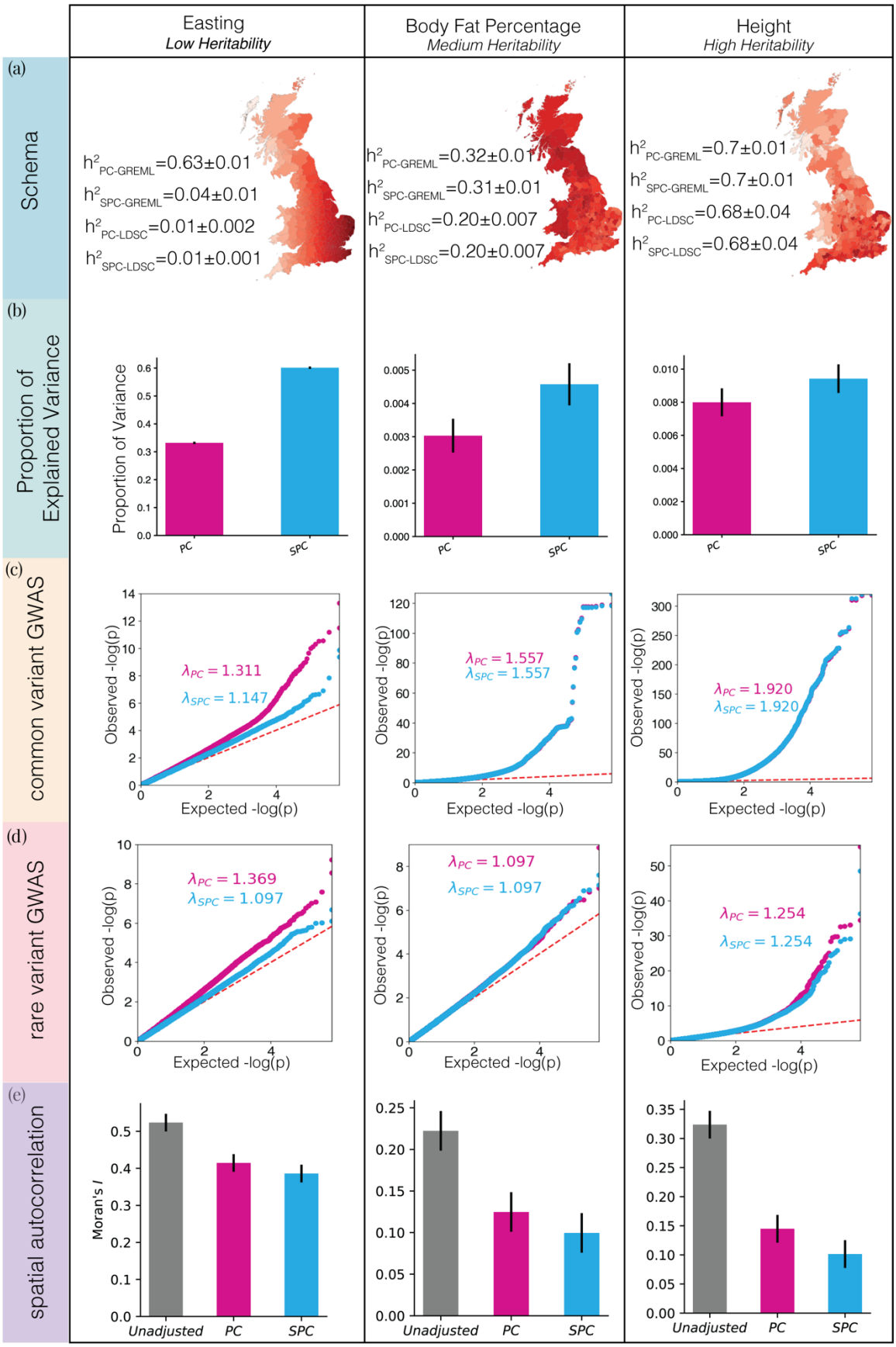
406

407 We then sought to evaluate the performance of SPCs in real data using genetic data from
408 individuals with self-reported White British ancestry in the UK Biobanks (UKBB; N~420,000). We
409 focused on three continuous phenotypes (**Figure 3a**): Easting coordinates of birthplace
410 (eastings), which display a linear geographical cline and have negligible true heritability, Body Fat
411 Percentage (BFP), which is established to have medium heritability (h^2~0.3), and height, which is
412 established to have high heritability (h^2~0.7; Karczewski et al. 2024). In terms of proportion of
413 variance explained, we saw similar results as the simulated data, with SPCs outperforming PCs
414 for all phenotypes. The difference was particularly significant for eastings. The GREML narrow-

415 sense heritability estimates revealed similar patterns, though the gap in estimates was smaller
416 compared to the simulated analysis. The heritability analysis for height and BFP resulted in the
417 same estimates in both models (0.7±0.01 for height, ~0.31±0.01 for BFP), while the estimates for
418 eastings were 0.63 (std: 0.01) and 0.04 (std:0.01), using PCs and SPCs, respectively. This
419 distinction between easting and the other two phenotypes is similar to that of environmental
420 smooth and polygenic phenotypes in the simulations, respectively.
421
422 Next, we conducted GWAS for all 3 phenotypes to measure the inflation of the test statistics due
423 to population structure, using the first 25 PCs and SPCs, age, age squared, sex, and an age/sex
424 interaction variable, as covariates. To account for indirect genetic effects, and cryptic relatedness,
425 which were not present in the simulated phenotypes, we used the BOLT-LMM software, which
426 applies a Bayesian Shrinkage Prior to adjust for those effects (Loh et al. 2015). In the GWAS of
427 eastings, SPCs resulted in a lower inflation of p-values, as evidenced by a lower genomic inflation
428 factor compared to PCs: 1.200 and 1.369, respectively. In contrast, the GWAS of height and BFP
429 showed no significant differences in the inflation of p-values as both approaches yielded the same
430 inflation factor (**Figure 3c**). We then conducted a GWAS of rare variants using imputed genotype
431 data, using BOLT-LMM. Looking exclusively at the imputed rare variants (MAF < 0.01), we
432 observed lower genomic inflation factor values for both models in all three phenotypes along with
433 similar patterns as the non-imputed GWAS, with SPCs having a lower inflation factor (1.147)
434 compared to PCs (1.311) in easting and both models having the same inflation factor for height
435 and BFP.
436
437 We then sought to explore the differences between the effect sizes estimated by each model. LD
438 score regression found identical heritability estimates from both models in all phenotypes (**Figure**
439 **4**), with a higher intercept for PCs in eastings GWAS (1.362 vs. 1.172). This suggests the
440 difference in performance in eastings was mostly driven by population structure confounding in

the PC model (Bulik-Sullivan et al. 2015). For height and BFP, aggregate analysis revealed negligible differences, with SPC-adjusted slope being lower but not significantly. We explored this further for height by extracting the list of loci for which the estimated single variant effect sizes differed significantly between the two approaches. We found that the minor alleles in these loci tended to be more uncommon, compared to all tested variants (Kolmogorov-Smirnov test p-value:5×10^{-83}), highlighting that the impact of SPCs is most pronounced at lower allele frequencies. The same pattern was also observed in the results for the eastings and BFP GWAS.

We calculated spatial autocorrelation of the phenotypes, in the form of Moran's I, as a measure of indirect environmental effects (Moran 1948). A previous study showed the spatial autocorrelation of phenotype distributions is reduced after adjustment using PCs, illustrating the efficacy of PC adjustment in addressing confounding due to such effects (Abdellaoui et al. 2019). Here we replicated that experiment and added SPC adjustment (**Figure 4e**) as another comparison approach. The Moran's I for all three phenotypes decreased after adjustment using PCs, 0.52 to 0.41 for easting, 0.22 to 0.12 for BFP, and 0.32 to 0.15 for height. Adjustment using SPCs further reduced Moran's I estimates to 0.38 for easting, and 0.10 for BFP and height. In summary, our analysis demonstrates that SPCs consistently outperform PCs in reducing population structure confounding, particularly for phenotypes with strong environmental components, and show comparable performance for highly and moderately heritable traits like height and BFP, especially when considering rare variant effect sizes and spatial autocorrelation.


Easting
Low Heritability
Body Fat Percentage
Medium Heritability
Height
High Heritability
(a)
Schema
h2PC-GREML=0.63±0.01
h2SPC-GREML=0.04±0.01
h2PC-LDSC=0.01±0.002
h2SPC-LDSC=0.01±0.001
h2PC-GREML=0.32±0.01
h2SPC-GREML=0.31±0.01
h2PC-LDSC=0.20±0.007
h2SPC-LDSC=0.20±0.007
h2PC-GREML=0.7±0.01
h2SPC-GREML=0.7±0.01
h2PC-LDSC=0.68±0.04
h2SPC-LDSC=0.68±0.04
(b)
Proportion of Explained Variance
Proportion of Variance
PC
SPC
(c)
common variant GWAS
Observed -log(p)
Expected -log(p)
λPC = 1.311
λSPC = 1.147
λPC = 1.557
λSPC = 1.557
λPC = 1.920
λSPC = 1.920
(d)
rare variant GWAS
λPC = 1.369
λSPC = 1.097
λPC = 1.097
λSPC = 1.097
λPC = 1.254
λSPC = 1.254
(e)
spatial autocorrelation
Moran's I
Unadjusted
PC
SPC

462

463 ***Figure 4*** - Comparison of PCs and SPCs as covariates to correct for population structure on the
464 outcomes in the UK Biobank. **a)** The structure of each phenotype. We analyzed 3 phenotypes
465 with varying degrees of true heritability, increasing from left to right. Heritability estimation using
466 GREML and LDSC methods, after correcting for recent population structure using both
467 approaches is shown, along with display of average phenotype values in each administrative
468 county in the UK. **b)** Proportion of explained variance in each phenotype by a model that only
469 includes PCs or SPCs as covariates. Error bars represent standard deviation. **c)** QQ plot of the
470 distribution of p-values in the GWAS of common variants for each phenotype that measure the
471 ability of covariates to deal with the effects of recent population structure on the outcomes. **d)** QQ
472 plot of the distribution of p-values in the GWAS of rare variants (MAF < 1%) in the imputed dataset.
473 **e)** Spatial autocorrelation of each phenotype measured in terms of Moran's I as a comparison of
474 the ability of each set of covariates to correct against the artifacts of environmental effects.

475

476 As a sensitivity analysis, we conducted a comparison of SPCs against other covariate extraction
477 techniques by evaluating their PVE on a random subset of the UK Biobank participants
478 (N=50,000; **Supplemental Figure 9**). SPCs outperformed both PCs of rare and common variants.
479 We note that in this dataset (unlike in our simulations), PCs of rare variants could not outperform
480 PCs of common variants. We then calculated additional component sets based on minimum IBD
481 sharing thresholds (6cM, 10cM, and 15cM), graph weighting schemes (weighted vs. unweighted
482 graphs), and component calculation methods (spectral vs. principal components of IBD sharing),
483 resulting in 12 sets of covariates (including SPCs), replicating our analysis using simulated data.
484 For each sharing threshold, SPCs outperformed their alternatives, highlighting the ability of
485 spectral components in extracting localized, fine-scale signals of population structure, along with
486 the advantages of utilizing unweighted matrices, especially in the first 5, and the first 10
487 dimensions. We further showed how a higher sharing threshold can negatively impact PVE by

excluding a significant proportion of IBD signals (**Figure 5**). Together with our simulated results, these analyses demonstrate that SPCs, calculated at the 6cM threshold, provide a more accurate adjustment for recent population structure than alternative adjustment strategies (further details in Supplemental Material).

To measure the robustness of SPCs in the presence of ascertainment bias, we explored the effects of network sparsity on the SPCs. In PC calculation, a lack of balance in representation may result in the PC's inability to capture population structure signals from subpopulations with lower representation (A. B. Lee et al. 2010). The IBD analysis of the UKBB cohort, due to its demographic history, leads to a densely connected graph, where most individuals have hundreds of IBD connections to others (**Supplemental Figure 10**). However, imbalances in representation result in a limited number of participants with lower connectivity. Here, connectivity (or node degree) is the number of edges per node. We searched for individuals with the lowest connectivity levels in a random subset (N=50,000). The lowest 20 node degrees ranged from 4 to 131 (mean = 84.65; median = 103). We looked at the distribution of SPC components to see if the nodes with low connectivity cluster near zero due to lower variability in their latent representation, suggesting failure to capture their underlying structure. We calculated the variance of each SPC dimension and compared them to the variances of a random sample of the same size (**Supplemental Figure 11**) and found that the variances calculated for lower connectivity nodes were not significantly lower than those of the random sample. When applying this to principal components, low-connectivity nodes exhibited significantly reduced variance. The distributions across SPC dimensions further illustrate SPCs' robustness to ascertainment bias (**Supplemental Figure 12**).

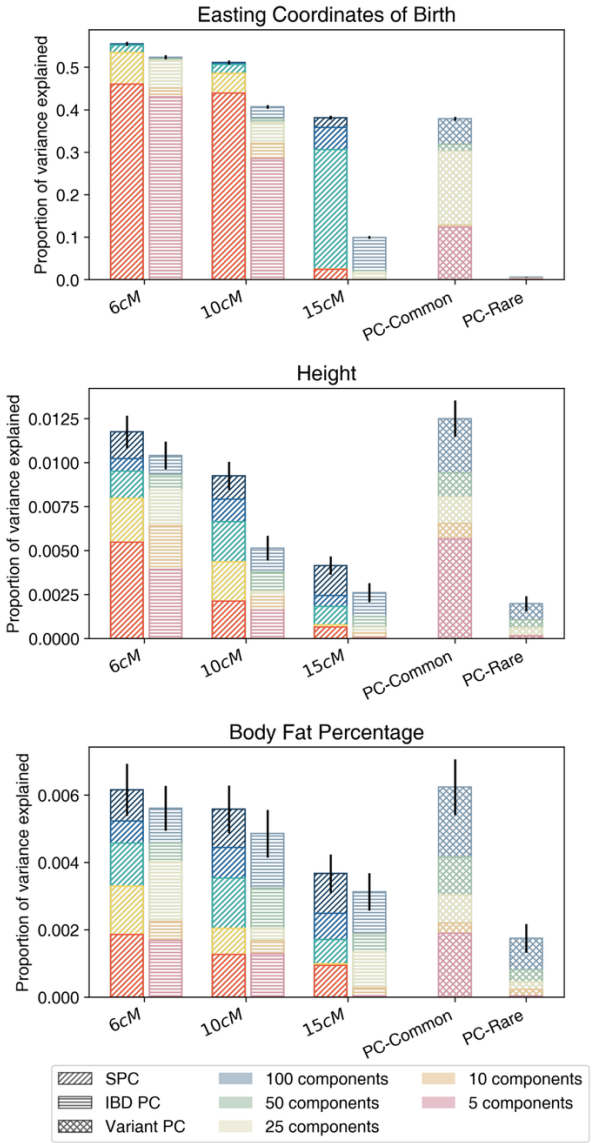


Figure 5 – Bar plots of the PVE in eastings, height, and BFP explained by SPCs (derived from binary IBD relatedness), IBD PCs (derived from binary IBD relatedness), and PCs (derived from genetic variants) in the UK Biobank, Exploring the effects of minimum IBD sharing parameter, and decomposition approach on the performance of covariates. Principal components of common and rare variants are displayed on the rightmost side of the figure. For other components, labels on the x-axis indicate the minimum IBD sharing threshold used to generate the components. Spectral (SPC) and principal (IBD PC) components were calculated per relatedness graph, indicated by

diagonal and horizontal shading lines, respectively. SPCs, are highlighted in brighter colors. PCs of common and rare variants are subscripted as PC-Common and PC-Rare, respectively.

Discussion

We introduced SPectral Components (SPCs), a graph-based approach that leverages IBD graphs to capture localized, non-linear fine-scale population structure, and transforms them into a continuous representation for statistical genetics analyses. This method overcomes some limitations of Principal Components (PCs), which are optimized for modeling continuous linear ancestral relationships, but struggle to fully account for recent demographic events that shape patterns of rare variation. Our results demonstrate that SPCs outperform PCs, explaining over 90% of fine-scale population structure in simulations while reducing genomic inflation in genome-wide association studies (GWAS) for both environmental and heritable phenotypes. Notably, SPCs improved rare variant association analyses by reducing genomic inflation, while providing more accurate heritability estimates and avoiding overcorrection for the highly heritable trait of height. Spatial autocorrelation analysis further confirmed that SPCs better control confounding from environmental effects, reducing Moran's I more effectively than PCs. These findings establish SPCs as a powerful and scalable tool to improve population structure adjustment and confounding control in large-scale, diverse biobank studies.

This work builds on previous research on IBD- or rare variant-based GRMs which have shown that the latter models can often overcorrect for population structure, leading to loss of true signal in genetic analysis. For example, several studies (Browning and Browning 2015; Evans et al.

2018) demonstrate that overcorrection can occur because IBD GRMs disproportionately emphasize recent, localized relatedness, which can obscure genuine associations. Additionally, it extends prior applications of spectral approaches to common variant data, which demonstrated they can be more robust than PCs in finding hidden population structure (A. B. Lee et al. 2010). However, those applications use variants that are shared identical-by-state, rather than haplotypes shared identical-by-descent, which lack genealogical context and are expected to have poorer resolution of confounding for recent population structure and rare variant association. The distinction between fixed effects and random effects on these models further complicates the adjustment process. Fixed effects, like PCs, explicitly control for population structure by removing specific sources of variation, whereas random effects models, like those incorporating GLMMs, aim to partition variance across all genetic relationships, but may inadvertently attenuate rare variant signals by treating them as background noise. SPCs provide a middle ground by capturing fine-scale population structure without overcorrection, using a fixed-effect approach tailored to discrete, localized signals, while maintaining the broader interpretability of linear mixed models. Moreover, the use of SPCs showed a clear reduction in genomic inflation for rare variants, therefore providing a significantly more robust adjustment. Other work on alternative approaches, such as using ARGs to infer genome-wide genealogies that can be combined with GLLMs to improve power for rare variant association has recently been shown to scale to large biobank datasets (Nait Saada et al. 2020; Zhang et al. 2023; Gunnarsson et al. 2024; Christ et al. 2024). While these approaches show promise, they currently have some limitations, such as implicating large genomic regions in variant association and requiring methodological improvements to LMM scalability. Future work could explore the use of PCs, ARGs, and SPCs in tandem as a powerful tool for addressing the challenges posed by population structure in modern genomics (Zaidi and Mathieson 2020; Ma and Shi 2020).

Another focus of this study was to examine how SPCs capture environmental effects, which often overlap with recent population structure (Zaidi and Mathieson 2020). The increased accuracy of SPCs compared to PCs is especially important given the increasing recognition of environmental confounding in GWAS, where geography and patterns of recent demography can lead to spurious associations. In this context, SPCs explain a significantly greater proportion of variance for environmental phenotypes, such as easting coordinates. This result aligns with earlier work (Abdellaoui et al. 2019; Cook et al. 2020; Leslie et al. 2015; Zaidi and Mathieson 2020), highlighting the difficulty of decoupling environmental and genetic factors in geographically clustered populations. Our findings support this previous work, demonstrating that population structure confounding remains a major challenge even in large-scale homogeneous cohorts like the White Britons from the UK Biobank. We showed this for the highly polygenic trait of height, as demonstrated by measuring the spatial autocorrelation in the form of Moran's I. SPCs were able to further decrease the spatial autocorrelation of height compared to PCs. Our results also support the use of SPCs in heritability analysis, where adjustment using SPCs reduced genetic heritability estimates by accounting for environmental confounding without overcorrection. By improving adjustment for the environmental confounding effects, SPCs provide a critical advance in the accuracy of genomic analysis for phenotypes that are heavily influenced by recent demography or environmental variables.

We highlight several limitations to this study and directions for future development. First, our study was restricted to unrelated samples of White British descent in the UK Biobank. SPCs should be further evaluated in more diverse datasets where relatedness, cryptic family structure, admixture, and population differentiation may introduce additional challenges. Second, though SPCs consistently outperform PCs in this study, additional work is needed to further validate the power of SPCs across settings such as smaller cohorts with sparser graphs, phenotypes with lower heritability, and dichotomous outcomes where collinearity may be more pronounced. Third, further

work could extend SPCs to additional analytic frameworks, like burden testing, generalized additive models, and polygenic risk score calculation, to assess their ability to improve prediction accuracy and calibration. Fourth, while our analysis focused on SPCs from global IBD graphs, alternative strategies merit exploration, including clustering-based partitioning of IBD graphs and multi-segment IBD detection approaches (Temple and Thompson 2025) which may capture additional fine-scale relatedness. Finally, SPCs could also be evaluated as covariates in IBD-based mapping of quantitative trait loci, and their ability to resolve fine-scale population structure differences suggest they may help improve mapping accuracy (Cai and Browning 2025; Chen et al. 2023).

In summary, SPCs offer a robust, continuous alternative to PCs for adjusting for recent population structure in genomic studies. By capturing fine-scale, non-linear demographic patterns, SPCs provide more accurate adjustment for confounding due to environmental effects and recent ancestry, particularly in large-scale cohorts. As large-scale biobanks continue to grow in size and diversity, the impact of rare variants in genetic analysis will become increasingly pronounced. In this context, tools like SPCs will be essential for ensuring accurate results, particularly as researchers delve deeper into the genetic architecture of rare variants on common disease and traits influenced by population-structured dynamics. By addressing the combined challenges of population structure, environmental confounding, and rare variant associations, SPC-based approaches pave the way to uncover novel insights into complex disease architecture.

Methods

Calculating Spectral and Principal Components

The calculation of SPCs can be described in 4 main steps (**Figure 1**).

Step 1: We conducted QC and phasing on array data. The phased data was used by the IBD estimation algorithm, iLASH, to detect pairwise for IBD for each dataset. An in-depth explanation of genotype QC and phasing, along with details about running iLASH and estimating IBD is described in (Shemirani et al. 2021). Throughout this paper we use the term IBD to refer to long haplotype segments (≥6cM) that are shared identical-by-state (IBS) and, given their length, are extremely likely to be inherited from a recent common ancestor. This operational definition is consistent with prior work (Browning and Browning 2015; Shemirani et al. 2021; Gusev et al. 2009). We acknowledge that some definitions of IBD differ, and we have added this clarification to reduce ambiguity. The results of the first step can be represented as S the set of all IBD segments, long enough to be detected. Each segment has a starting point and an end point s_1, s_2, located on autosomal chromosome k.

$$S = \left\{(a, b, s_1, s_2) \,\middle|\, a, b \in N \ \wedge \ s_1, s_2 \in L_k \wedge k = 1, \dots, 22\right\}$$

Step 2: the aggregated sums of pairwise IBD sharing (i.e. the total or combined length of IBD haplotypes that are shared per pair of individuals) were calculated and used to generate an IBD relatedness graph. On this graph, nodes represented participants and edge weights represented the aggregated IBD sharing. A minimum threshold of aggregated IBD sharing is enforced both as a QC measure (Belbin et al. 2021) and to reflect the population structure of the dataset.

$$G_{IBD} = (V, E)$$

$$V = N$$

$$E = \left\{(a, b) \middle| a, b \in N \ \wedge \ \sum_{(s_1, s_2) \in S_{a,b}} (s_2 - s_1) > l\right\}$$

To calculate SPCs we used a threshold of 6cM. However, we explored two other thresholds of 10cM and 15cM for comparison (c.f. Supplementary Material).

Step 3: The $n \times n$ adjacency matrix $\boldsymbol{A}$ was generated, representing the relatedness graph, where rows and columns both represent the n participants in the dataset. Each element a_{ij} in this matrix is equal to 1 if the aggregated length of IBD segments shared between the two participants i and j on the graph is above the threshold of $l = 6cM$; otherwise, the value of the elements is set to zero.

$$a_{ij} = \begin{cases} 1 & if\ (a,b) \in E \\ 0 & otherwise \end{cases}$$

We refer to this matrix as the binary IBD relatedness matrix. We also explored the use of non-binary relatedness matrices for comparison due to their similarity with common Genetic Relatedness Matrices (GRM; c.f. Supplementary Material).

Step 4: Any graph adjacency matrix can be transformed into a Laplacian form (Shuman et al. 2013). Here we defined a symmetrically normalized Laplacian:

$$L^{Sym} := I - (D^{+})^{\frac{1}{2}} A (D^{+})^{\frac{1}{2}}$$

Where I is the identity matrix, D is defined as the degree matrix, a diagonal matrix where each element d_{ii} is equal to the total number of neighbors for participant i, and $D^{+}$is the (Moore-Penrose) inverse of D. The Eigenvectors of the Laplacian matrix are commonly referred to as the spectral components of the graph. Briefly, SPCs are the spectral components of the binary IBD relatedness graph with a cut-off threshold of 6cM. Thus, akin to PCs, SPCs attribute a set of numerical values to each vertex based on its projections on the set of principal axes of variation in the graph. The level of detail represented by each axis depends on the corresponding eigenvalue association with it. Eigenvector associated with the smaller eigenvalues will assign similar number to neighbors on the graph. As the associated eigenvalue increases, similar values

are assigned to participants that are incrementally further from each other on the graph (further details in Supplemental Material).

We have implemented this pipeline as a Python library that transforms flat text IBD graph files to SciPy sparse matrices and uses CPU or GPU acceleration to calculate SPCs using the same library.

Generating simulated data

We used the MSprime python library to implement the simulation pipeline to ensure the distribution of IBD connections and their length is concordant with real datasets (Baumdicker et al. 2022). Following guidelines on simulation of homogeneous population with recent structure akin to that of UK Biobanks (Zaidi and Mathieson 2020), our pipeline simulated N-by-N grids of demes with constant migration rate of 0.05. We set N=5 (25 demes) for our main simulation of chromosome 1 and 50,000 diploid individuals. We also simulated a replication dataset with N=4 with 22 chromosomes and 8,000 diploid individuals. The length of the chromosomes in the 8,000-sample simulation were adopted from the HapMap project (International HapMap Consortium et al. 2007). Chromosome specific recombination rates were adopted from the stdpopsim library metadata (Adrion et al. 2020). In the 50,000-sample simulation, we used the same library to obtain chromosome 1 specifications, with 248,956,422 base pairs and recombination rate of 1.149×10^{-8}. The migration rate of 0.05 has been shown to establish demes with average divergence, measured by F_{st}, equal to that of neighboring administrative counties in the UK Biobank dataset (Leslie et al. 2015). All demes originate from a single ancestral population 300 generations ago. The first 150 generations going back from the current day were simulated using Discrete Time Wright-Fisher (DTWF) model, implemented in the MSprime library to ensure IBD length distribution is similar to that observed in human genotype datasets. The simulation pipeline is available publicly on GitHub.

While calculating true IBD data is readily possible with MSprime, we sought to replicate common noise present in real-world applications, such as switch errors and false-positive IBD segments. The simulated tree sequences were converted to whole-genome sequence data in PLINK 1.9 format in order to delete the ground-truth phase information. The data was then filtered to retain only common variants (MAF>1%) and then phased using Eagle v4.2 with default parameters. Genetic map files for Eagle were generated using the recombination rates applied in the simulations. We have provided the code to generate these files in our simulation GitHub repository. IBD segments were estimated using the iLASH software v1.0.1 with default parameters (min_length 2.9, perm_count: 20, shingle_size: 15, shingle_overlap: 0, auto_slice: True, cm_overlap: 1, bucket_count: 5, match_threshold: 0.99, interest_threshold: 0.70). For the comparative analysis of downstream applications, 7 types of phenotypes were simulated (Figure 3a). The first category, referred to as environmental smooth phenotypes, are non-heritable phenotypes with a distribution that depends on the deme of origin. All values were generated using a normal distribution with constant variance across all demes. Moving vertically from top to bottom, the mean of the normal distribution decreases monotonically per row. It decreases from a value of $2\boldsymbol{\sigma}$, for the top row, to zero, for the bottom row. Briefly, in a 5 by 5 grid simulation, this means the phenotype values on the top row are drawn from the distribution $y \sim \mathcal{N}(2\sigma, \sigma)$, the values on the second row are drawn from the distribution $y \sim \mathcal{N}(1.5\sigma, \sigma)$, each subsequent row has a mean that is 0.5σ smaller than the previous, with last row having a mean of zero. Second phenotype, called the environmental sharp phenotype is also non-heritable with a normal distribution with a non-zero mean ($y \sim \mathcal{N}(\mu, \sigma)$) at the target deme and a zero-centered normal distribution ($y \sim \mathcal{N}(0, \sigma)$) in all other demes. We used a $\sigma = 1$ in our analysis for this manuscript. We generated 3 heritable polygenic phenotypes, to which we refer, based on their heritability, polygenic – low heritability (h^2=0.1), polygenic – medium heritability (h^2=0.3), polygenic high

717 heritability (h^2=0.08). For these phenotypes, we selected causal SNPs randomly from each
718 window of 10 KB along the genome (24,813 causal variants per phenotype) with effect sizes
719 drawn from $\beta_j \sim \mathcal{N}\left(0, \sigma_l^2\left[p_j\left(1-p_j\right)\right]^{\alpha}\right)$ where σ_l^2 is calculated based on selective pressure
720 parameter $\boldsymbol{\alpha}$=-0.5, and the total heritability of the phenotype ($\sigma_s = 0.1$, 0.3, and 0.8) via the
721 formula $\sigma_s = \sigma_l \sum_{j=1}^{M}\left[2p_j\left(1-p_j\right)\right]^{1+\alpha}$. We then used the weights along with the simulated whole
722 genome sequence data, and the score calculation functionality in the PLINK software (--score) to
723 generate the mean phenotype values for each individual. The remaining variance of the
724 phenotypes were added using a random normal distribution with variances of 0.9, 0.7, and 0.2
725 respectively.
726 Finally, we simulated two hybrid phenotypes by combining the environmental smooth and sharp
727 phenotypes and the polygenic phenotype with medium heritability (h^2 = 0.3), by reweighting the
728 environmental phenotype to have a variance of 0.7 and adding the values of the two phenotypes
729 to generate hybrid smooth and hybrid sharp phenotypes, respectively.
730
731 LD data to detect variants tagging causal ones was generated using PLINK with parameters: "--
732 list-all, --show-tags, --tag-kb 2000, --tag-r2 0.1.
733
734 Ground-truth IBD data were extracted using the ibd_segment function in the tskit library with a
735 minimum span of 3cM and maximum TMRCA of 200.
736
737
738
739 **UK Biobank Data**
740

741 Phased genotype data along with the imputed data for the UK Biobank project was downloaded
742 through the accession code 84541. The UK Biobank participants were genotyped using the
743 Applied Biosystems UK BiLEVE Axiom Array by Affymetrix 807,411. The genotype data was
744 phased using the SHAPEIT3 software and the 1000 Genomes Project as the reference panel
745 (Bycroft et al. 2018; Auton et al. 2015; Delaneau et al. 2019). After removing indels from the 22
746 autosomal chromosomes, a total of 655,532 SNPs were retained. Data field 21000, based on self-
747 reported race/ethnicity was used to extract (N=471,931) individuals with White British ancestry.
748 Information on allele frequency and missingness in both phasing and QC process for GWAS data
749 is available in (Loh et al. 2018). We estimated fine-scale IBD sharing using the iLASH software
750 (Shemirani et al. 2021) using the same parameters mentioned in the previous section. Second
751 degree relatives were identified using KING (parameters: --related --degree 2; N=24,799). From
752 each pair of these relatives, at least one individual was excluded from the analysis (N=405,508).
753 Height information was accessed using data field 20 (N=352,254). BFP was accessed using data
754 field 23099 (N=346,655). Easting coordinates of birth was accessed using data field 129 (N=
755 343,845). PCs were calculated using PLINK2 software.
756

757 **Genome-wide association study**

758

759 The GWAS of simulated data was conducted using the GCTA software with default parameters
760 and FastGWA option to run linear regression (Jiang et al. 2019). For the common variant analysis,
761 we excluded variants with MAF<1% using PLINK2 software. For the rare/uncommon variant
762 analysis we excluded variant with MAF>1% using the same software. The GWAS in UK Biobank
763 was conducted using the BOLT-LMM software version 1 to account for cryptic relatedness and
764 direct genetic effects, absent in the simulated phenotype analysis. Age, Age squared, age, sex,
765 age by sex, and age squared by sex were used as covariates (Loh et al. 2015). The LD reference

panel was calculated using genotype data from individuals with reported European ancestry in the 1000 Genomes Project (https://broad-alkesgroup-ukbb-ld.s3.amazonaws.com; access date January 2023). The default genetic map file available in the SHAPEIT software package (Delaneau et al. 2019) calculated for Human Genome version 19 (hg19) was used for genetic distances. Minor allele frequency of 0.001 and a minimum INFO score of 0.3 filters were used for the imputed data.

Computing Environment

All analyses in this manuscript were performed on the High-Performance Computing Cluster at Mount Sinai. The compute node used featured an Intel Emerald Rapids 8568Y+ CPU with 96 cores at 2.3 GHz (20 cores were used for our applications), and 1.5 TB of RAM (approximately 100 GB were used during our runs). The operating system was CentOS v7.0, with GCC v8.3.0 and Python v3.5.

Software availability

A python implementation of the SPC algorithm is available at https://github.com/roohy/spc

Our simulation pipeline is available at https://github.com/roohy/msprime_ancestry_simulation

The Eagle phasing software is available at: https://alkesgroup.broadinstitute.org/Eagle/

The PLINK2 software is available at: https://zzz.bwh.harvard.edu/plink/plink2.shtml

The iLASH software is available at: https://github.com/roohy/iLASH

GCTA software is available at: https://yanglab.westlake.edu.cn/software/gcta/

BOLT-LMM software is available at: https://alkesgroup.broadinstitute.org/BOLT-LMM/

Data availability

Simulated genotype data, along with phenotypes is available on Mendeley repository with doi:10.17632/9dcpdbzv4m.1

Competing interest statement

The authors declare to competing interests

Acknowledgements

Conceptualization by R.S., E.E.K., C.R.G., and N.Z. Methodology developed by R.S., and E.E.K., and N.Z., Conceptualization of the simulation of population structure by R.S., and G.M.B., Formal analysis by R.S., E.E.K., and N.Z., reviewing and editing by all authors, and supervision by E.E.K.

This work was supported by the National Human Genome Research Institute of the National Institutes of Health under award numbers R01HG011345 and U01HG010971.

This work was supported in part through the computational and data resources and staff expertise provided by Scientific Computing and Data at the Icahn School of Medicine at Mount Sinai and supported by the Clinical and Translational Science Awards (CTSA) grant UL1TR004419 from the National Center for Advancing Translational Sciences. Research reported in this publication was also supported by the Office of Research Infrastructure of the National Institutes of Health under award number S10OD026880 and S10OD030463. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

Abdellaoui, Abdel, David Hugh-Jones, Loic Yengo, et al. 2019. "Genetic Correlates of Social Stratification in Great Britain." *Nature Human Behaviour* 3 (12): 1332–42. https://doi.org/10.1038/s41562-019-0757-5.

Adrion, Jeffrey R, Christopher B Cole, Noah Dukler, et al. 2020. “A Community-Maintained Standard Library of Population Genetic Models.” *eLife* 9 (June): e54967. https://doi.org/10.7554/eLife.54967.

Auton, Adam, Gonçalo R. Abecasis, David M. Altshuler, et al. 2015. “A Global Reference for Human Genetic Variation.” *Nature* 526 (7571): 68–74. https://doi.org/10.1038/nature15393.

Baumdicker, Franz, Gertjan Bisschop, Daniel Goldstein, et al. 2022. “Efficient Ancestry and Mutation Simulation with Msprime 1.0.” *Genetics* 220 (3): iyab229. https://doi.org/10.1093/genetics/iyab229.

Baye, Tesfaye M., Hua He, Lili Ding, Brad G. Kurowski, Xue Zhang, and Lisa J. Martin. 2011. “Population Structure Analysis Using Rare and Common Functional Variants.” *BMC Proceedings* 5 (9): S8. https://doi.org/10.1186/1753-6561-5-S9-S8.

Belbin, Gillian M., Sinead Cullina, Stephane Wenric, et al. 2021. “Toward a Fine-Scale Population Health Monitoring System.” *Cell* 184 (8): 2068-2083.e11. https://doi.org/10.1016/j.cell.2021.03.034.

Berg, Jeremy J, Arbel Harpak, Nasa Sinnott-Armstrong, et al. 2019. “Reduced Signal for Polygenic Adaptation of Height in UK Biobank.” *eLife* 8 (March): e39725. https://doi.org/10.7554/eLife.39725.

Border, Richard, Georgios Athanasiadis, Alfonso Buil, et al. 2022. “Cross-Trait Assortative Mating Is Widespread and Inflates Genetic Correlation Estimates.” *Science* 378 (6621): 754–61. https://doi.org/10.1126/science.abo2059.

Browning, Sharon R., and Brian L. Browning. 2015. “Accurate Non-Parametric Estimation of Recent Effective Population Size from Segments of Identity by Descent.” *The American Journal of Human Genetics* 97 (3): 404–18. https://doi.org/10.1016/j.ajhg.2015.07.012.

Browning, Sharon R., and Brian L. Browning. 2023. “Biobank-Scale Inference of Multi-Individual Identity by Descent and Gene Conversion.” Preprint, bioRxiv, November 5. https://doi.org/10.1101/2023.11.03.565574.

Bulik-Sullivan, Brendan K., Po-Ru Loh, Hilary K. Finucane, et al. 2015. “LD Score Regression Distinguishes Confounding from Polygenicity in Genome-Wide Association Studies.” *Nature Genetics* 47 (3): 291–95. https://doi.org/10.1038/ng.3211.

Bycroft, Clare, Colin Freeman, Desislava Petkova, et al. 2018. “The UK Biobank Resource with Deep Phenotyping and Genomic Data.” *Nature* 562 (7726): 203–9. https://doi.org/10.1038/s41586-018-0579-z.

Caggiano, Christa, Arya Boudaie, Ruhollah Shemirani, et al. 2023. "Disease Risk and Healthcare Utilization among Ancestrally Diverse Groups in the Los Angeles Region." *Nature Medicine* 29 (7): 1845–56. https://doi.org/10.1038/s41591-023-02425-1.

Cai, Ruoyi, and Sharon R. Browning. 2025. "Identity-By-Descent Mapping Using Multi-Individual IBD With Genome-Wide Multiple Testing Adjustment." *Genetic Epidemiology* 49 (6): e70015. https://doi.org/10.1002/gepi.70015.

Chen, Han, Ardalan Naseri, and Degui Zhi. 2023. "FiMAP: A Fast Identity-by-Descent Mapping Test for Biobank-Scale Cohorts." *PLOS Genetics* 19 (12): e1011057. https://doi.org/10.1371/journal.pgen.1011057.

Christ, Ryan, Xinxin Wang, Louis J. M. Aslett, David Steinsaltz, and Ira Hall. 2024. "Clade Distillation for Genome-Wide Association Studies." Preprint, bioRxiv, October 1. https://doi.org/10.1101/2024.09.30.615852.

Cook, James P, Anubha Mahajan, and Andrew P Morris. 2020. "Fine-Scale Population Structure in the UK Biobank: Implications for Genome-Wide Association Studies." *Human Molecular Genetics* 29 (16): 2803–11. https://doi.org/10.1093/hmg/ddaa157.

Dai, Chengzhen L., Mohammad M. Vazifeh, Chen-Hsiang Yeang, et al. 2020. "Population Histories of the United States Revealed through Fine-Scale Migration and Haplotype Analysis." *The American Journal of Human Genetics* 106 (3): 371–88. https://doi.org/10.1016/j.ajhg.2020.02.002.

Delaneau, Olivier, Jean-François Zagury, Matthew R. Robinson, Jonathan L. Marchini, and Emmanouil T. Dermitzakis. 2019. "Accurate, Scalable and Integrative Haplotype Estimation." *Nature Communications* 10 (1): 5436. https://doi.org/10.1038/s41467-019-13225-y.

Devlin, B., and Kathryn Roeder. 1999. "Genomic Control for Association Studies." *Biometrics* 55 (4): 997–1004. https://doi.org/10.1111/j.0006-341X.1999.00997.x.

Evans, Luke M., Rasool Tahmasbi, Matt Jones, et al. 2018. "Narrow-Sense Heritability Estimation of Complex Traits Using Identity-by-Descent Information." *Heredity* 121 (6): 616–30. https://doi.org/10.1038/s41437-018-0067-0.

Freyman, William A, Kimberly F McManus, Suyash S Shringarpure, et al. 2021. "Fast and Robust Identity-by-Descent Inference with the Templated Positional Burrows–Wheeler Transform." *Molecular Biology and Evolution* 38 (5): 2131–51. https://doi.org/10.1093/molbev/msaa328.

Fu, Wenqing, Timothy D. O'Connor, Goo Jun, et al. 2013. "Analysis of 6,515 Exomes Reveals the Recent Origin of Most Human Protein-Coding Variants." *Nature* 493 (7431): 216–20. https://doi.org/10.1038/nature11690.

Gilbert, Edmund, Ashwini Shanmugam, and Gianpiero L. Cavalleri. 2022. "Revealing the Recent Demographic History of Europe via Haplotype Sharing in the UK Biobank." *Proceedings of the National Academy of Sciences* 119 (25): e2119281119. https://doi.org/10.1073/pnas.2119281119.

Gravel, Simon, Brenna M. Henn, Ryan N. Gutenkunst, et al. 2011. "Demographic History and Rare Allele Sharing among Human Populations." *Proceedings of the National Academy of Sciences of the United States of America* 108 (29): 11983–88. https://doi.org/10.1073/pnas.1019276108.

Gravel, Simon, Fouad Zakharia, Andres Moreno-Estrada, et al. 2013. "Reconstructing Native American Migrations from Whole-Genome and Whole-Exome Data." *PLOS Genetics* 9 (12): e1004023. https://doi.org/10.1371/journal.pgen.1004023.

Grone, Robert, Russell Merris, and V. S. Sunder. 1990. "The Laplacian Spectrum of a Graph." *SIAM Journal on Matrix Analysis and Applications* 11 (2): 218–38. https://doi.org/10.1137/0611016.

Gunnarsson, Árni Freyr, Jiazheng Zhu, Brian C. Zhang, Zoi Tsangalidou, Alex Allmont, and Pier Francesco Palamara. 2024. "A Scalable Approach for Genome-Wide Inference of Ancestral Recombination Graphs." Preprint, bioRxiv, September 2. https://doi.org/10.1101/2024.08.31.610248.

Gusev, Alexander, Jennifer K. Lowe, Markus Stoffel, et al. 2009. "Whole Population, Genome-Wide Mapping of Hidden Relatedness." *Genome Research* 19 (2): 318–26. https://doi.org/10.1101/gr.081398.108.

Helgason, Agnar, Bryndís Yngvadóttir, Birgir Hrafnkelsson, Jeffrey Gulcher, and Kári Stefánsson. 2005. "An Icelandic Example of the Impact of Population Structure on Association Studies." *Nature Genetics* 37 (1): 90–95. https://doi.org/10.1038/ng1492.

International HapMap Consortium, Kelly A. Frazer, Dennis G. Ballinger, et al. 2007. "A Second Generation Human Haplotype Map of over 3.1 Million SNPs." *Nature* 449 (7164): 851–61. https://doi.org/10.1038/nature06258.

Jiang, Longda, Zhili Zheng, Ting Qi, et al. 2019. "A Resource-Efficient Tool for Mixed Model Association Analysis of Large-Scale Data." *Nature Genetics* 51 (12): 1749–55. https://doi.org/10.1038/s41588-019-0530-8.

Johnson, Ruth, Yi Ding, Vidhya Venkateswaran, et al. 2022. "Leveraging Genomic Diversity for Discovery in an Electronic Health Record Linked Biobank: The UCLA ATLAS

Community Health Initiative." *Genome Medicine* 14 (1): 104. https://doi.org/10.1186/s13073-022-01106-x.

Kang, Hyun Min, Jae Hoon Sul, Susan K. Service, et al. 2010. "Variance Component Model to Account for Sample Structure in Genome-Wide Association Studies." *Nature Genetics* 42 (4): 348–54. https://doi.org/10.1038/ng.548.

Karczewski, Konrad J., Rahul Gupta, Masahiro Kanai, et al. 2024. "Pan-UK Biobank GWAS Improves Discovery, Analysis of Genetic Architecture, and Resolution into Ancestry-Enriched Effects." Preprint, medRxiv, October 1. https://doi.org/10.1101/2024.03.13.24303864.

Kunegis, Jérôme, Stephan Schmidt, Andreas Lommatzsch, Jürgen Lerner, Ernesto W. De Luca, and Sahin Albayrak. 2010. "Spectral Analysis of Signed Graphs for Clustering, Prediction and Visualization." In *Proceedings of the 2010 SIAM International Conference on Data Mining (SDM)*. Proceedings. Society for Industrial and Applied Mathematics. https://doi.org/10.1137/1.9781611972801.49.

Lawson, Daniel John, Neil Martin Davies, Simon Haworth, et al. 2020. "Is Population Structure in the Genetic Biobank Era Irrelevant, a Challenge, or an Opportunity?" *Human Genetics* 139 (1): 23–41. https://doi.org/10.1007/s00439-019-02014-8.

Lee, Ann B., Diana Luca, and Kathryn Roeder. 2010. "A SPECTRAL GRAPH APPROACH TO DISCOVERING GENETIC ANCESTRY." *The Annals of Applied Statistics* 4 (1): 179–202. https://doi.org/10.1214/09-AOAS281.

Lee, Seunggeung, Gonçalo R. Abecasis, Michael Boehnke, and Xihong Lin. 2014. "Rare-Variant Association Analysis: Study Designs and Statistical Tests." *American Journal of Human Genetics* 95 (1): 5–23. https://doi.org/10.1016/j.ajhg.2014.06.009.

Leslie, Stephen, Bruce Winney, Garrett Hellenthal, et al. 2015. "The Fine-Scale Genetic Structure of the British Population." *Nature* 519 (7543): 309–14. https://doi.org/10.1038/nature14230.

Loh, Po-Ru, Gleb Kichaev, Steven Gazal, Armin P. Schoech, and Alkes L. Price. 2018. "Mixed-Model Association for Biobank-Scale Datasets." *Nature Genetics* 50 (7): 906–8. https://doi.org/10.1038/s41588-018-0144-6.

Loh, Po-Ru, George Tucker, Brendan K. Bulik-Sullivan, et al. 2015. "Efficient Bayesian Mixed-Model Analysis Increases Association Power in Large Cohorts." *Nature Genetics* 47 (3): 284–90. https://doi.org/10.1038/ng.3190.

Ma, Shengqing, and Gang Shi. 2020. "On Rare Variants in Principal Component Analysis of Population Stratification." *BMC Genetics* 21 (1): 34. https://doi.org/10.1186/s12863-020-0833-x.

Mathieson, Iain, Felix R. Day, Nicola Barban, et al. 2023. “Genome-Wide Analysis Identifies Genetic Effects on Reproductive Success and Ongoing Natural Selection at the FADS Locus.” *Nature Human Behaviour* 7 (5): 790–801. https://doi.org/10.1038/s41562-023-01528-6.

Mathieson, Iain, and Gil McVean. 2012. “Differential Confounding of Rare and Common Variants in Spatially Structured Populations.” *Nature Genetics* 44 (3): 243–46. https://doi.org/10.1038/ng.1074.

Moran, P. A. P. 1948. “The Interpretation of Statistical Maps.” *Journal of the Royal Statistical Society. Series B (Methodological)* 10 (2): 243–51.

Nait Saada, Juba, Georgios Kalantzis, Derek Shyr, et al. 2020. “Identity-by-Descent Detection across 487,409 British Samples Reveals Fine Scale Population Structure and Ultra-Rare Variant Associations.” *Nature Communications* 11 (1): 6130. https://doi.org/10.1038/s41467-020-19588-x.

Nei, Masatoshi, and Fumio Tajima. 1981. “GENETIC DRIFT AND ESTIMATION OF EFFECTIVE POPULATION SIZE.” *Genetics* 98 (3): 625–40. https://doi.org/10.1093/genetics/98.3.625.

Price, Alkes L., Nick J. Patterson, Robert M. Plenge, Michael E. Weinblatt, Nancy A. Shadick, and David Reich. 2006. “Principal Components Analysis Corrects for Stratification in Genome-Wide Association Studies.” *Nature Genetics* 38 (8): 904–9. https://doi.org/10.1038/ng1847.

Price, Alkes L., Noah A. Zaitlen, David Reich, and Nick Patterson. 2010. “New Approaches to Population Stratification in Genome-Wide Association Studies.” *Nature Reviews Genetics* 11 (7): 7. https://doi.org/10.1038/nrg2813.

Privé, Florian, Keurcien Luu, Michael G B Blum, John J McGrath, and Bjarni J Vilhjálmsson. 2020. “Efficient Toolkit Implementing Best Practices for Principal Component Analysis of Population Genetic Data.” *Bioinformatics* 36 (16): 4449–57. https://doi.org/10.1093/bioinformatics/btaa520.

Sakaue, Saori, Jun Hirata, Masahiro Kanai, et al. 2020. “Dimensionality Reduction Reveals Fine-Scale Structure in the Japanese Population with Consequences for Polygenic Risk Prediction.” *Nature Communications* 11 (1): 1569. https://doi.org/10.1038/s41467-020-15194-z.

Schoech, Armin P., Daniel M. Jordan, Po-Ru Loh, et al. 2019. “Quantification of Frequency-Dependent Genetic Architectures in 25 UK Biobank Traits Reveals Action of Negative Selection.” *Nature Communications* 10 (1): 1. https://doi.org/10.1038/s41467-019-08424-6.

Shemirani, Ruhollah, Gillian M. Belbin, Christy L. Avery, Eimear E. Kenny, Christopher R. Gignoux, and José Luis Ambite. 2021. “Rapid Detection of Identity-by-Descent Tracts for Mega-Scale Datasets.” *Nature Communications* 12 (1): 3546. https://doi.org/10.1038/s41467-021-22910-w.

Shuman, David I., Sunil K. Narang, Pascal Frossard, Antonio Ortega, and Pierre Vandergheynst. 2013. “The Emerging Field of Signal Processing on Graphs: Extending High-Dimensional Data Analysis to Networks and Other Irregular Domains.” *IEEE Signal Processing Magazine* 30 (3): 83–98. https://doi.org/10.1109/MSP.2012.2235192.

Si, Si, Donghyuk Shin, Inderjit S. Dhillon, and Beresford N. Parlett. 2014. “Multi-Scale Spectral Decomposition of Massive Graphs.” *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2* (Cambridge, MA, USA), NIPS’14, vol. 2 (December): 2798–806.

Sohail, Mashaal, Robert M Maier, Andrea Ganna, et al. 2019. “Polygenic Adaptation on Height Is Overestimated Due to Uncorrected Stratification in Genome-Wide Association Studies.” *eLife* 8 (March): e39702. https://doi.org/10.7554/eLife.39702.

Temple, Seth D., and Elizabeth A. Thompson. 2025. “Identity-by-Descent Segments in Large Samples.” *Theoretical Population Biology* 165 (October): 10–21. https://doi.org/10.1016/j.tpb.2025.06.003.

Wainschtein, Pierrick, Deepti Jain, Zhili Zheng, et al. 2022. “Assessing the Contribution of Rare Variants to Complex Trait Heritability from Whole-Genome Sequence Data.” *Nature Genetics* 54 (3): 263–73. https://doi.org/10.1038/s41588-021-00997-7.

Yang, Jian, Michael N. Weedon, Shaun Purcell, et al. 2011. “Genomic Inflation Factors under Polygenic Inheritance.” *European Journal of Human Genetics* 19 (7): 807–12. https://doi.org/10.1038/ejhg.2011.39.

Yang, Jian, Jian Zeng, Michael E. Goddard, Naomi R. Wray, and Peter M. Visscher. 2017. “Concepts, Estimation and Interpretation of SNP-Based Heritability.” *Nature Genetics* 49 (9): 1304–10. https://doi.org/10.1038/ng.3941.

Zaidi, Arslan A, and Iain Mathieson. 2020. “Demographic History Mediates the Effect of Stratification on Polygenic Scores.” *eLife* 9 (November): e61548. https://doi.org/10.7554/eLife.61548.

Zhang, Brian C., Arjun Biddanda, Árni Freyr Gunnarsson, Fergus Cooper, and Pier Francesco Palamara. 2023. “Biobank-Scale Inference of Ancestral Recombination Graphs Enables Genealogical Analysis of Complex Traits.” *Nature Genetics* 55 (5): 768–76. https://doi.org/10.1038/s41588-023-01379-x.