Quantile-specific confounding: correction for subtle population stratification via quantile regression

Chen Wang[1], Marco Masala[2], Edoardo Fiorillo[2], Marcella Devoto[2], Francesco Cucca[2,3], Iuliana Ionita-Laza[1,4,#]

[1]Department of Biostatistics, Columbia University, New York, USA

[2]Institute for Genetic and Biomedical Research, National Research Council, Italy

[3]Department of Biomedical Sciences, University of Sassari, Sassari, Italy

[4]Department of Statistics, Lund University, Lund, Sweden

[#]ii2135@cumc.columbia.edu

Summary

Subtle population structure remains a significant concern in genome-wide association studies. Using human height as an example, we show how quantile regression, a natural extension of linear regression, can better correct for subtle population structure due to its inherent ability to adjust for quantile-specific effects of covariates such as principal components. We utilize data from the UK biobank and the SardiNIA/ProgeNIA project for demonstration.

Many complex traits are highly polygenic, and genome-wide association studies (GWAS) allow for the discovery of genetic variation associated with such traits. A well-known confounder in GWAS is population structure; left unaccounted, it leads to false positive associations. Adjusting for this confounding effect using principal component analyses (PCA) or mixed effect modeling has become standard practice in GWAS [1]; yet there is always the risk of residual confounding especially in large and heterogeneous GWAS analyses. Indeed, recent studies have shown that population stratification still accounts for a substantial fraction of GWAS associations for many traits [2], including height. The largest GWAS for height to date with approximately 5.4 million individuals has been conducted by the GIANT (Genetic Investigation of Anthropometric traits) consortium [3]. GIANT is by nature a large and heterogeneous dataset that includes a vast number of cohorts of mostly European-descent individuals from Europe and North America, and more than one million participants of East Asian, Hispanic, African, or South Asian ancestry, and therefore

concerns about incomplete control of population structure have been raised in the literature [4-5]. Even in the UK Biobank (UKBB), a much more homogeneous cohort, there is evidence of residual population stratification [6].

PCA performs well when the structure has a smooth and wide distribution. However, for more sharp and localized distributions, PCA may not be enough [7]. Although alternative methods for detecting finer scale population structure have been discussed in the literature [8,9], there are several drawbacks and practical challenges for implementing them in practice.

We illustrate here an application of quantile regression (QR) [10], a classical statistical technique that can better adjust for sharp and localized population stratification. In such cases, the confounding effect may vary across parts of the phenotype distribution and may even be concentrated on specific quantiles. Linear regression (LR) by design assumes that covariate effects are homogeneous across quantiles and therefore can under- or over- adjust when effects are heterogeneous. In contrast, QR allows for covariate adjustment at specific quantiles, and therefore can be more appropriate than LR when the covariate effects vary across quantiles. For example, physical activity can have stronger effects on upper BMI levels, but lower effects on upper height quantiles ([11], **Fig. S1**) and QR naturally adjusts for these heterogeneous effects.

QR is a natural extension of LR and consists of fitting a series of linear regression models at different quantile levels $\tau \in (0,1)$. To test for association between the jth genetic variant and phenotype Y at quantile level τ we fit the following conditional quantile regression model:

$$Q_Y(\tau \mid X_j, C) = X_j\beta_j(\tau) + C\alpha(\tau),$$

where $Q_Y(\tau|X, C)$ is the τth conditional quantile of phenotype Y given genotypes X and covariates C; $\beta_j(\tau)$ and $\alpha(\tau)$ are quantile-specific coefficients and can be estimated by minimizing the pinball loss function [10]. Since both the genotype effects $\beta_j(\tau)$ and the covariate effects $\alpha(\tau)$ are allowed to vary by quantile, QR can (*i*) detect heterogeneity in effects across quantiles, and (*ii*) adjust for covariates with varying effects across quantiles. For testing the null hypothesis H_0: $\beta_j(\tau) = 0$, we employ the rank score test to test and obtain a p-value for each SNP at quantile level τ [12-13]. In practice, in addition to quantile specific p-values we also compute a composite p-value by combining p-values at multiple quantile levels using Cauchy's combination method [14].

To illustrate the scenario of local and sharp population stratification, we utilized data from 325,667 white British unrelated individuals in UKBB, and 1,946 unrelated individuals from the ProgeNIA/SardiNIA project, a longitudinal study of a cohort of Sardinian subjects [15-16].

Following standard QC procedures, we analyzed 7,844,680 SNPs in both datasets. We used standing height measurements (at baseline) as the phenotype.

We performed LR as well as QR at several quantile levels, including $\tau = 0.1, 0.3, 0.5, 0.7, 0.9$. For both LR and QR, we adjusted for sex and 10 global PCs (**Methods**). We did not adjust for age, as age can act as a collider in the combined Sardinia and UKBB data leading to high inflation in false positives.

British individuals are on average taller than individuals from Sardinia by 8 cm (**Fig. 1(a)**), leading to more pronounced stratification at lower quantile levels (e.g. $\tau = 0.1 - 0.3$). PC1 and PC2 are describing genetic variation within UK, while PC3 and PC5 can separate the two subpopulations (**Fig. 1(b)**). The effects of PC5 on height vary as a function of quantile levels and are stronger at lower quantiles (**Fig. 1(c)**).

We first performed simulations by randomly permuting the height phenotype separately within the Sardinia and UKBB data, and report detailed results for one replicate and summary results across 20 independent replicates. We noticed a decrease in false positives induced by population stratification for QR relative to LR (**Fig. 2(a)**). Specifically, LR tends to identify more significant associations than QR across different replicates.

Similarly, for real height, we see a reduction in the number of significant associations for QR vs. LR (**Fig. 2(a)**). Although most associations are shared between LR and QR (457 loci), LR results in a much larger number of new discoveries relative to QR (189 vs. 10 loci). These new associations likely reflect both increased power for LR over QR but also increased number of false positives since by combining data from UKBB and Sardinia a certain number of false positives are expected due to effects of population stratification and assortative mating. When restricting analyses to 325,667 white British unrelated individuals, the number of discoveries decreases substantially, especially for loci identified by LR but missed by QR (139 vs. 189 above) likely reflecting an overall reduction in false positives (**Fig. 2(b)**).

By comparison, in the case of BMI the effect of population stratification appears less pronounced (**Fig. S2** and **S3**). Relative to height, BMI shows more heterogeneity in genetic effects genome-wide, with power at the upper quantiles higher than at lower quantiles, suggestive of more pronounced gene-by-environment effects at upper quantiles of BMI, as expected.

In summary, QR is a robust technique for the analysis of continuous phenotypes in large heterogeneous genomic data that provides a good balance between discovery power and control of false positives. Beyond its robustness, QR allows for a more nuanced

understanding of genotype-phenotype correlations by revealing patterns of heterogeneity that is not possible with the GWAS approach.

Acknowledgements. C.W. and I.I.L. are supported by NIH grants MH095797 and AG072272. This research has been conducted using the UK Biobank Resource under Application Number 41849.

Code Availability. Scripts for QR GWAS analysis are available at https://github.com/Iuliana-Ionita-Laza/QRGWAS. The quantile regression was performed using the R package quantreg (https://cran.r-project.org/package=quantreg) version 5.94. The rank score test was performed using the R package QRank (https://cran.r-project.org/package=QRank) version 1.0.

References

1. Price, A. L. et al. (2010). New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics, 11*(7), 459–463. https://doi.org/10.1038/nrg2813
2. Tan, T. et al. (2024). Family-GWAS reveals effects of environment and mating on genetic associations. *medRxiv*. https://doi.org/10.1101/2024.10.01.24314703v1
3. Yengo, L. et al. (2022). A saturated map of common genetic variants associated with human height from 5.4 million individuals of diverse ancestries. *Nature*. https://doi.org/10.1038/s41586-022-05275-y
4. Sohail, M. et al. (2019). Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies. *eLife, 8*, e39702. https://doi.org/10.7554/eLife.39702
5. Novembre, J., & Barton, N. H. (2018). Tread Lightly Interpreting Polygenic Tests of Selection. *Genetics*, *208*(4), 1351–1355. https://doi.org/10.1534/genetics.118.300786
6. Haworth, S. et al. (2019). Apparent latent structure within the UK Biobank sample has implications for epidemiological analysis. *Nature Communications, 10*(1), 333. https://doi.org/10.1038/s41467-018-08219-1
7. Mathieson, I., & McVean, G. (2012). Differential confounding of rare and common variants in spatially structured populations. *Nature genetics*, *44*(3), 243–246. https://doi.org/10.1038/ng.1074
8. Diaz-Papkovich, A. et al. (2019). UMAP reveals cryptic population structure and phenotype heterogeneity in large genomic cohorts. *PLoS genetics*, *15*(11), e1008432. https://doi.org/10.1371/journal.pgen.1008432
9. Nait Saada, J. 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

10. Koenker, R., & Bassett, G. (1978). Regression quantiles. *Econometrica, 46*(1), 33–50. https://doi.org/10.2307/1914337
11. Bann, D., Fitzsimons, E., & Johnson, W. (2020). Determinants of the population health distribution: an illustration examining body mass index. *International journal of epidemiology*, *49*(3), 731–737. https://doi.org/10.1093/ije/dyz245
12. Koenker, R. (2005). *Quantile regression* (Vol. 38). Cambridge University Press.
13. Song, X., et al. (2017). QRank: A novel quantile regression tool for eQTL discovery. *Bioinformatics, 33*(14), 2123–2130. https://doi.org/10.1093/bioinformatics/btx141
14. Wang, C., et al. (2024). Genome-wide discovery for biomarkers using quantile regression at biobank scale. *Nature Communications, 15*, 6460. https://doi.org/10.1038/s41467-024-25963-4
15. Pilia, G., et al. (2006). Heritability of cardiovascular and personality traits in 6,148 Sardinians. *PLoS Genetics, 2*(6), e132. https://doi.org/10.1371/journal.pgen.0020132
16. Sidore, C., et al. (2015). Genome sequencing elucidates Sardinian genetic architecture and augments association analyses for lipid and blood inflammatory markers. *Nature Genetics, 47*(12), 1272–1281. https://doi.org/10.1038/ng.3402

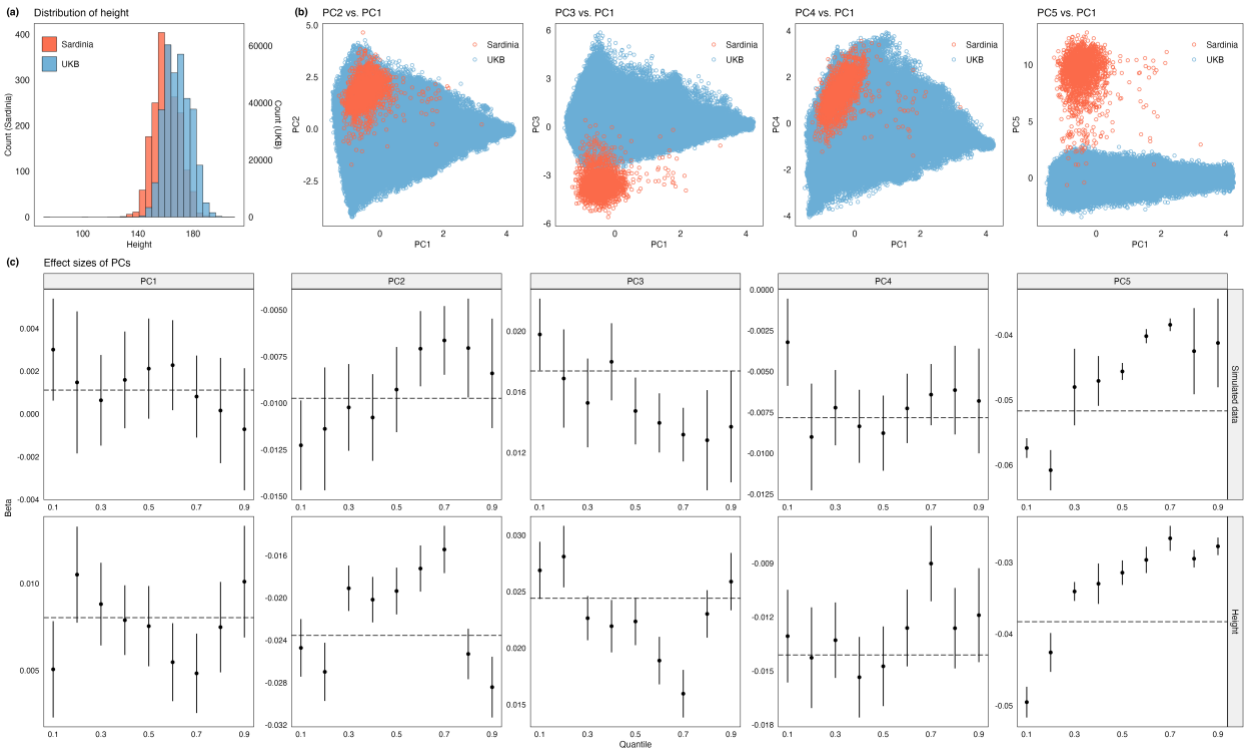

Figure 1: **Height distribution and PCs of genetic variation in two cohorts, Sardinia and UKBB. (a)** Height distributions; **(b)** PC plots for top PCs; **(c)** PCs' estimated quantile-specific effects (and standard errors) on height at quantile levels $\tau = 0.1 - 0.9$. Upper panel: simulated data; lower panel: height. The horizontal line corresponds to the mean-based effect in LR.

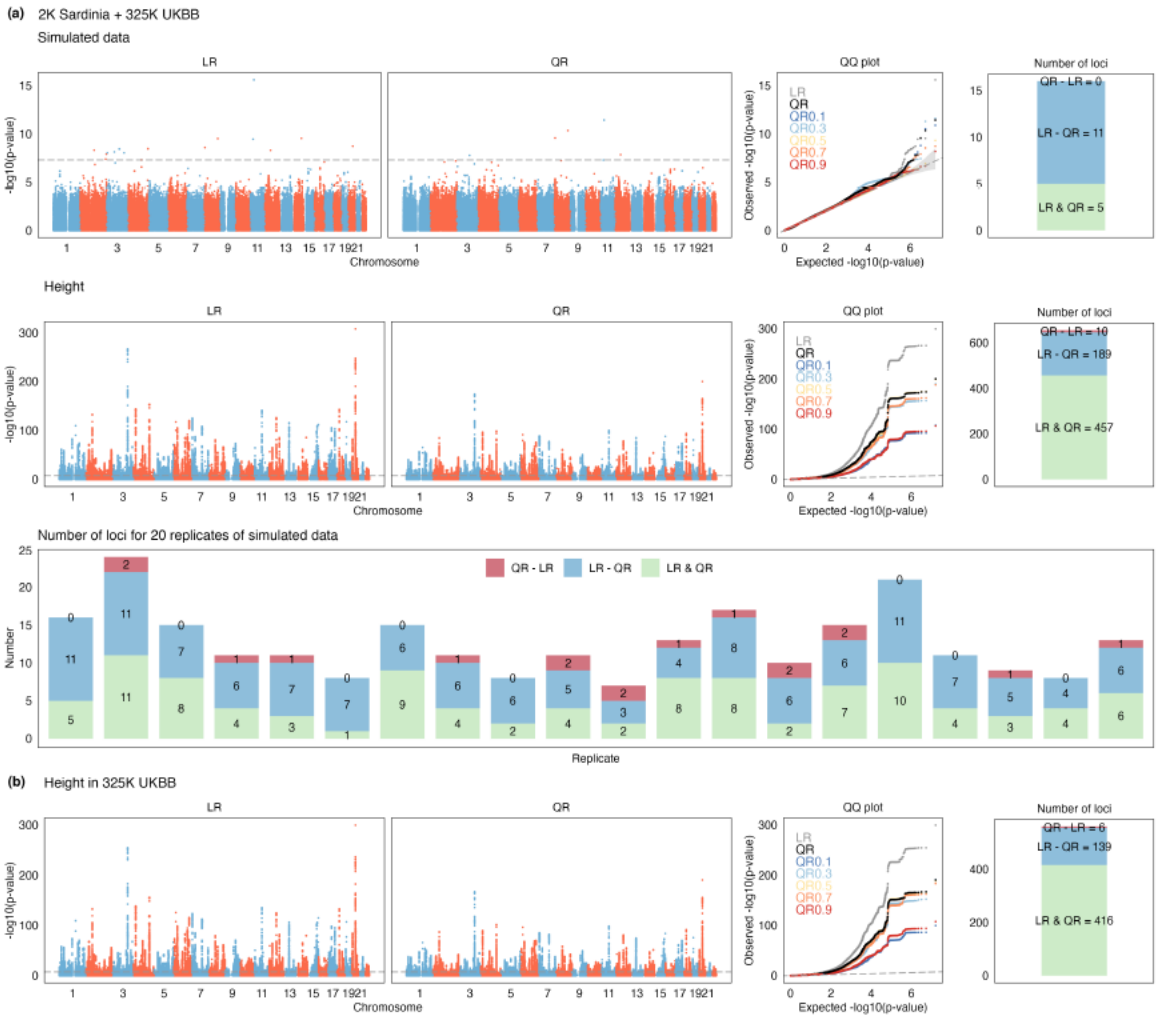

Figure 2: **LR and QR GWAS on simulated and height data with stratification**. Manhattan and QQ plots are shown for **(a)** simulated (no real signal) and height data when combining UKBB and Sardinia; **(b)** only UKBB. We report the number of independent loci (unique and shared) identified by LR and QR.

Methods

Overview of quantile regression. We assume that we have n independent samples from a population. We denote by $Y = (Y_1, \cdots, Y_n)'$ the $n \times 1$ sample phenotype vector, by X the $n \times p$ genotype matrix, and by C the $n \times q$ matrix for covariates including age, gender and principal components of genetic variation. In GWAS, the typical approach is to perform marginal (unconditional) testing for one variant at a time, assuming the following model:

$$E(Y \mid X_j, C) = X_j\beta_j + C\alpha.$$

In QR, we denote the τth conditional quantile function of Y as $Q_Y(\tau \mid X, C)$. For linear QR, we can write the conditional quantile regression model for the jth variant and specific quantile level $\tau \in (0,1)$ as:

$$Q_Y(\tau \mid X_j, C) = X_j\beta_j(\tau) + C\alpha(\tau),$$

where $\beta_j(\tau)$ and $\alpha(\tau)$ are quantile-specific coefficients that can be estimated by minimizing the pinball loss function. This optimization problem can be formulated equivalently as a linear programming problem that can be solved efficiently using the simplex algorithm or interior point methods. For QR, a commonly used hypothesis testing tool is the rank score test that allows us to obtain well-calibrated p-values at individual quantile levels. The rank score test statistic for a fixed quantile level τ is defined as:

$$S_{QRank,j,\tau} = n^{-1/2}\sum_{i=1}^{n} X_{ij}^{*}\phi_\tau\left(Y_i - C_i\hat{\alpha}(\tau)\right) \text{ and } V_{QRank,j} = n^{-1}\tau(1-\tau)X_j^{*'}X_j^{*},$$

where $X^{*} = P_C X$, $P_C = I - C(C'C)^{-1}C'$ and $\phi_\tau(u) = \tau - I(u < 0)$ is the τ-pinball loss. Under the null hypothesis $H_0: \beta_j(\tau) = 0$, $V_{QRank,j}^{-1/2} S_{QRank,j,\tau} \sim N(0,1)$. Note that the asymptotic distribution of the test statistic is independent of the distribution of the phenotype. Hence it can be applied to any phenotype without requiring a pre-transformation to achieve normality, which is another important advantage of QR over LR. Cauchy combination can be used to integrate results over different quantile levels.

PCA. We perform PCA analysis as follows. We combined genotype data of unrelated samples from the UKBB White British population (n=325,667) and the ProgeNIA/SardiNIA cohort (n=1,946). To identify unrelated individuals, we used KING [1] to estimate the genetic relatedness between samples and only retained the unrelated samples that were more distant than 2nd-degree relatives (KING kinship coefficient <= 0.0884). We selected SNPs with a minor allele frequency (MAF) > 0.01 and genotype missingness < 5%. We performed linkage disequilibrium (LD) pruning using PLINK2 [2] with a sliding window of 1000 SNPs, a step size of 100 SNPs, and an LD threshold r2 < 0.1 to retain independent variants. We also excluded three long-range LD regions (chr6: 25.5M-33.5M, chr8: 8M-12M, and chr11: 46M-

57M) [3] from the LD pruning. We computed the first 10 PCs on the LD-pruned variants using randomized partial PCA implemented in FlashPCA [4].

Locus identification. We performed LD clumping using PLINK2 [2] with a 1 Mb distance threshold and an LD threshold of r2 > 0.1 to identify independent loci. Adjacent loci within 1 Mb of each other were subsequently merged into a single locus.

1. Manichaikul, A et al. (2010). Robust relationship inference in genome-wide association studies. *Bioinformatics (Oxford, England)*, *26*(22), 2867–2873. https://doi.org/10.1093/bioinformatics/btq559
2. Chang, C. C. et al. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*, *4*, 7. https://doi.org/10.1186/s13742-015-0047-8
3. Weissbrod, O. et al. (2020). Functionally informed fine-mapping and polygenic localization of complex trait heritability. *Nature genetics*, *52*(12), 1355–1363. https://doi.org/10.1038/s41588-020-00735-5
4. Abraham, G. et al. (2017). FlashPCA2: principal component analysis of Biobank-scale genotype datasets. *Bioinformatics (Oxford, England)*, *33*(17), 2776–2778. https://doi.org/10.1093/bioinformatics/btx299

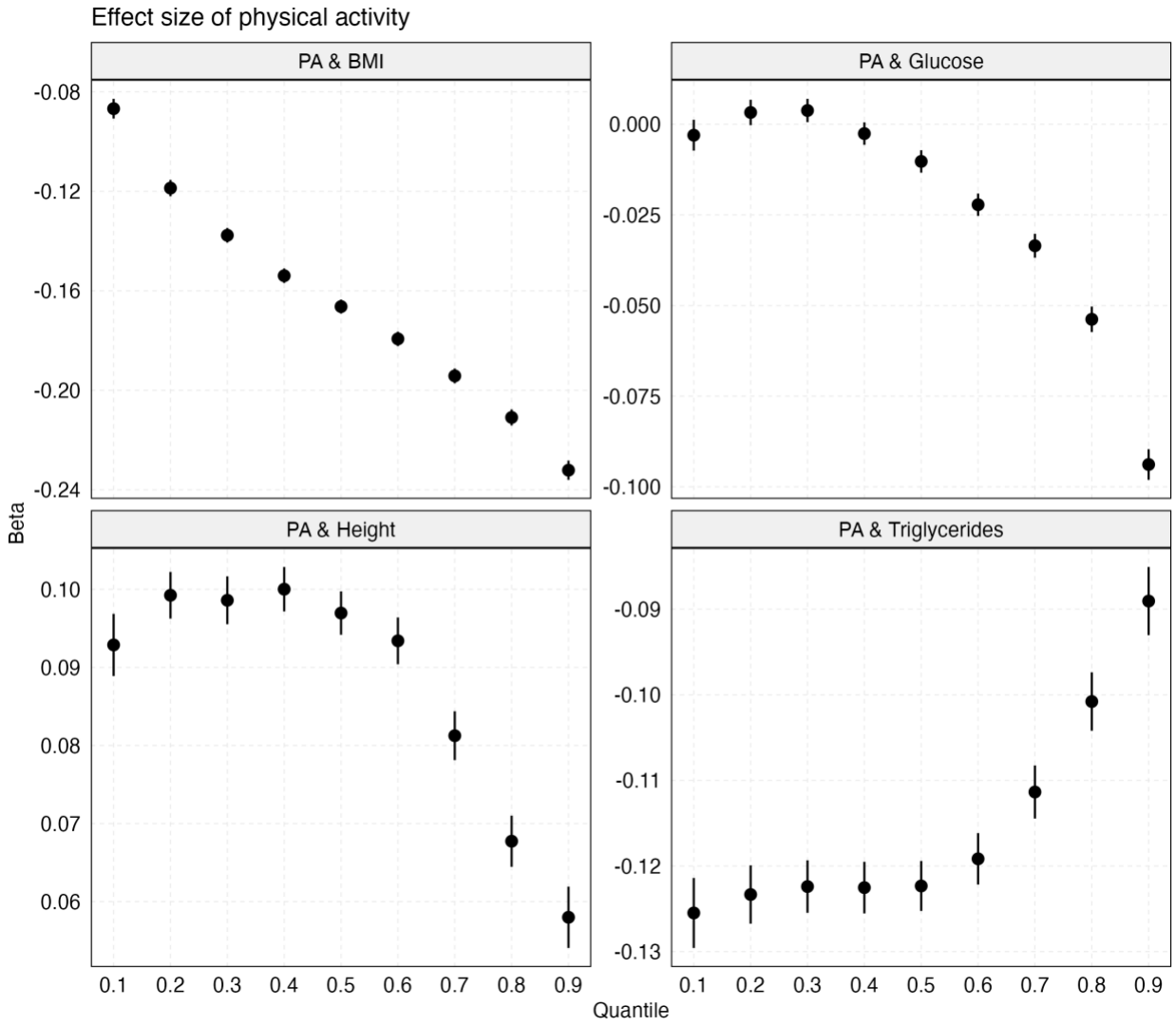

Figure S1: **Heterogeneous effects of physical activity (PA) on BMI, height, glucose and triglycerides in UKBB.** Estimated effects and standard errors across different quantile levels are shown. These analyses are based on unrelated white British individuals (between 264,854 and 325,825 depending on the trait). PA is a categorical variable (high, moderate, low) defined based on questionnaire data.

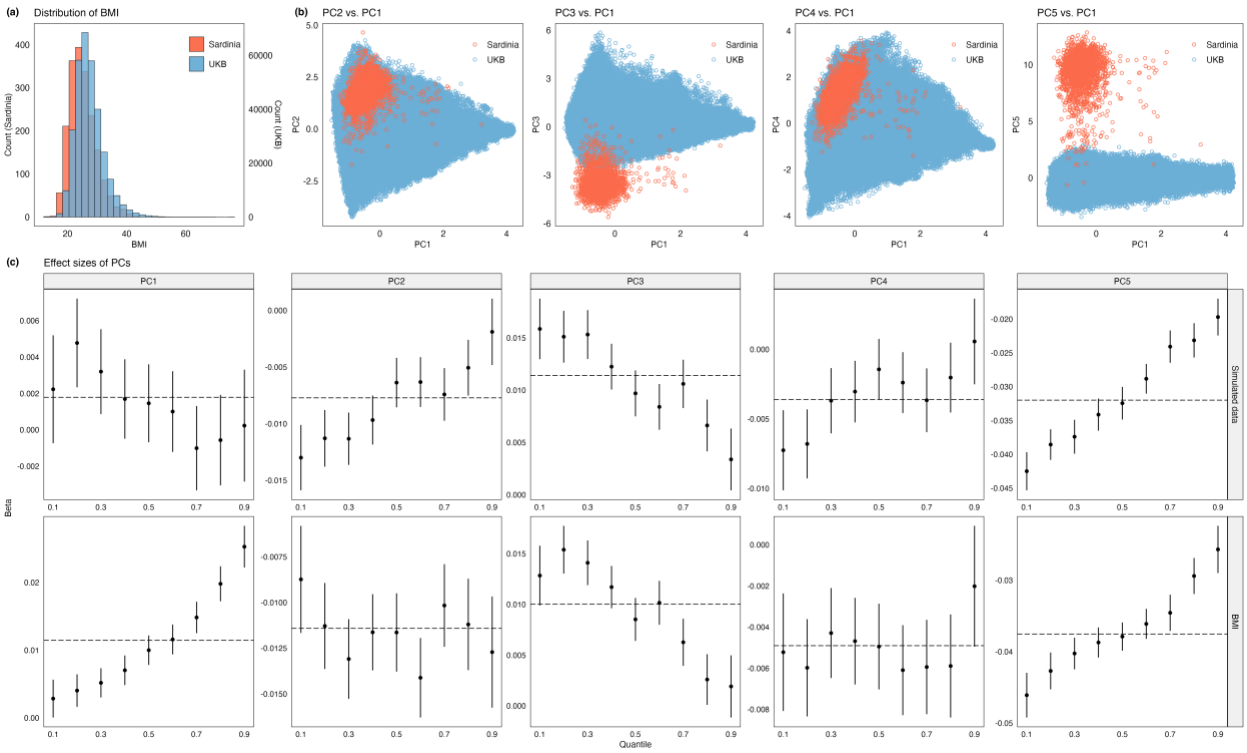

Figure S2: **BMI distribution and PCs of genetic variation in two cohorts, Sardinia and UKBB. (a)** BMI distributions; **(b)** PC plots for top PCs; **(c)** PCs' estimated quantile-specific effects (and standard errors) on BMI at quantile levels $\tau = 0.1 - 0.9$. Upper panel: simulated data; lower panel: BMI. The horizontal line corresponds to the mean-based effect in LR.

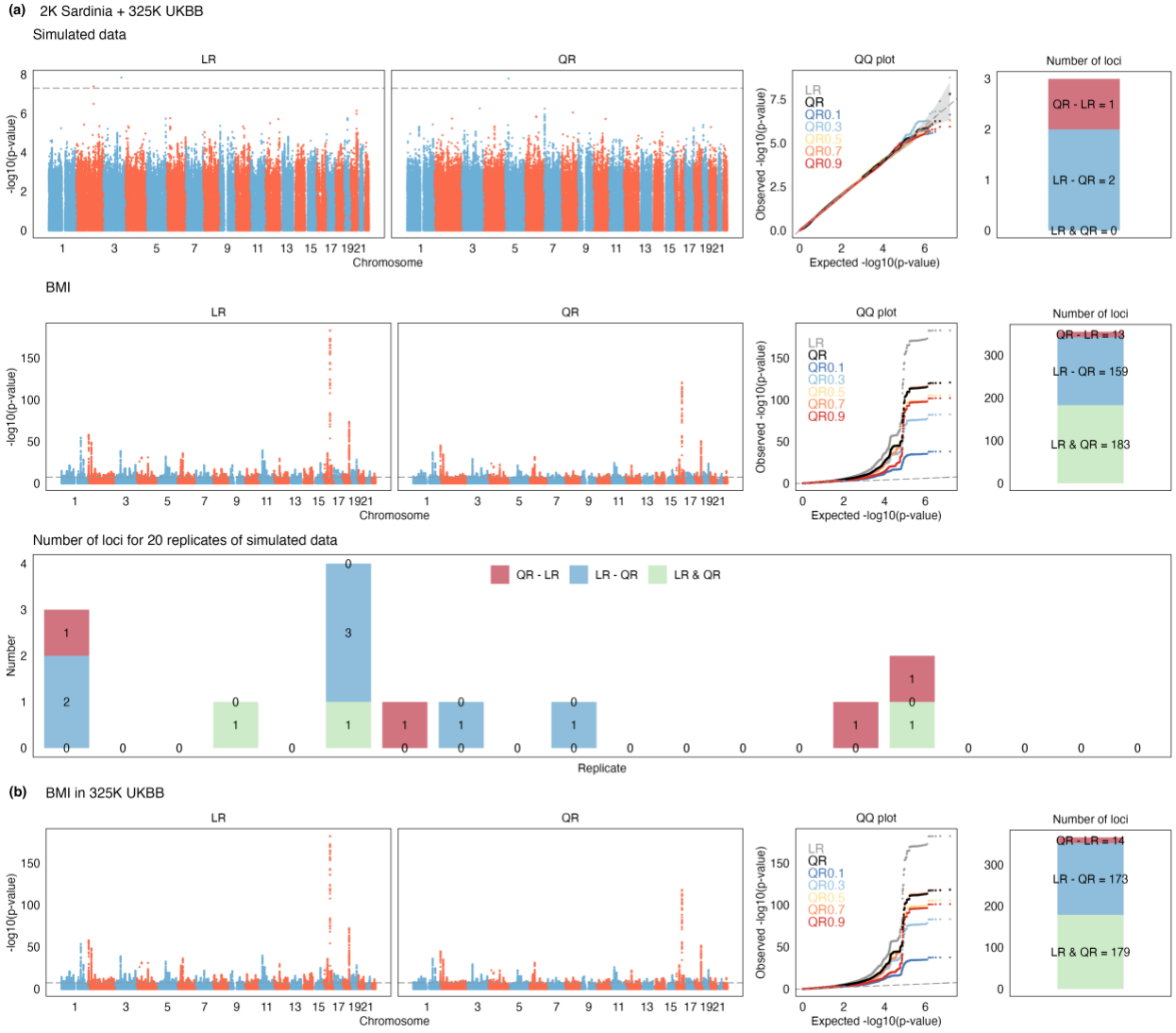

Figure S3: **LR and QR GWAS on simulated and BMI data with stratification**. Manhattan and QQ plots are shown for **(a)** simulated (no real signal) and BMI data when combining UKBB and Sardinia; **(b)** only UKBB. We report the number of independent loci (unique and shared) identified by LR and QR.