Phenotype-aware decoupling of related subjects

Wanjun Gu1,2, Jiachen Xi1, Steven Cao1, Rany M. Salem1

1Herbert Wertheim School of Public Health and Human Longevity Science, University of California, San Diego, La Jolla, CA, 92093.

2University of California, San Francisco, San Francisco, CA 92093.

**Word count**: XXXX

**Tables**: 2, **Figures**: 1

**Corresponding author and to whom reprint requests are to be addressed:**

Rany M. Salem, PhD, MPH.

Herbert Wertheim School of Public Health and Longevity Science,

University of California, San Diego

9500 Gilman Drive #0725, La Jolla CA, 92093

Tel: 858-246-0433

Fax: 858- 534-4642

Email: rsalem@health.ucsd.edu

**Abstract**

Relatedness within genomic cohorts is a potential source of bias for many genetic analyses. Existing tools to break relatedness are phenotype naïve, indiscriminately remove subjects to break relationship, risking the loss of valuable data, especially in studies targeting uncommon and rare phenotypes. To address this limitation, we developed the Kinship Decouple and Phenotype Selection (KDPS) tool, with a novel algorithm designed to enhance the precision of subject selection in genetic and epidemiological research by incorporating phenotype prioritization. KDPS separates related individuals by considering kinship or identity by descent scores, while simultaneously prioritizing subjects based on phenotypes of interest. This approach enables the retention of valuable subjects for analysis, even in the face of necessary exclusions due to relatedness. Furthermore, KDPS accommodates a wide range of phenotypes, including binary, ordinal, and quantitative types, and allows for customization to either prioritize specific phenotypes or maximization of sample size. In simulations based on the UK Biobank dataset and real-world datasets, KDPS demonstrated significant improvements in retention of subjects with prioritized phenotypes and computational efficiency compared to previously published software. The ability of this method to process biobank-scale studies within practical timeframes marks a considerable advancement over existing techniques. KDPS offers tailored solutions for complex analytical challenges and broad applicability in genetics and epidemiology research. To our knowledge, KDPS is the first tool to perform phenotype-aware decoupling, paving the way for more powerful genetic and epidemiological analyses.

**Introduction**

Geneticists utilize a broad suite of sophisticated methodologies to decode the complex architecture of genotype-phenotype relationship. Despite recent advances in statistical methods to accommodate relatedness (Loh et al. 2015, Zhou et al. 2018), a significant number of genetic analysis methods require studies of unrelated individuals, including: selection scans (Akbari et al. 2018, Booker, Jackson and Keightley 2017), admixture mapping (Ali-Khan and Daar 2010), gene by environment interaction analysis (Ottman 1996, Virolainen et al. 2023), and analyses across stratum (Winkler et al. 2017).

Over the past two decades, population-based genetic studies have become the central methodology for elucidating the relationships between genetic variation and complex human traits. The expansion of dataset sizes introduces potential bias of cryptic relatedness (Voight and Pritchard 2005), especially in large-scale initiatives like the UK Biobank (Bycroft et al. 2018) and All of US research program (All of Us Research Program Genomics Investigators 2024). Cryptic relatedness occurs when two or more genetically related individuals are included in a study and the genetic relationship is unrecognized. Investigators need to check for cryptic relatedness in their study population and remove subjects to break relatedness if their analytical approach cannot accommodate relatedness. However, the removal of related subjects comes at the cost of excluding individuals with relevant phenotypes, and significantly diminishing the statistical power of the study, impacting the ability to detect meaningful genetic associations, especially in gene-environment interaction studies. Sample size and resulting statistical power are critical to success of population-based genetic studies. The issue is critical in investigations of rare/prevalent phenotypes, such as cancer, psychiatric and autoimmune diseases, where each individual case is vital. Therefore, algorithms that break relatedness that optimize sample size while retaining individuals with phenotypes of interest can drastically improve statistical power to uncover the genetics of the traits and disease.

Several tools and approaches are available to break related subjects in a study, including KING (Manichaikul et al. 2010), PLINK2 (Chang et al. 2015), Friends and family (de Jager et al. 2017), SampleSeq2 (Edwards and Li 2012) and FastIndep (Abraham and Diaz 2014). However, none of which take phenotypes into account. For instance, the most widely used decoupling method is executed in PLINK2, employs a greedy algorithm that removes individuals most related to others, resorting to random selection among equally related subjects. Naïve non-selective pruning is particularly problematic in scenarios in which the phenotype of interest is rare or uncommon in the cohort. Eliminating subjects with valuable phenotypic traits results in reduced sample size and the statistical power. We introduce the Kinship Decouple and Phenotype Selection (KDPS) method to address this limitation. KDPS generates an sample with unrelated subjects by considering genetic relatedness metrics and can prioritize retention of subjects based on phenotypes of interest. This innovative approach maximizes the number of subjects with desired phenotypes and/or exposure of interest.

**Materials and methods**

Loading relatedness and phenotype data

KDPS requires an kinship relatedness (Manichaikul et al. 2010) or identity by descent (IBS) score (Su et al. 2012) and phenotypic data files. These are used to ensure that only subjects present in both the kinship and phenotype datasets are included (Figure 1A). KDPS accepts two categories of phenotypes: categorical, with two or more categories, or numerical. KDPS can be run with phenotype prioritization or phenotype naïve. For phenotype prioritization, users are required to specify the primary phenotypes of interest. For categorical phenotypes, an ordered list indicating priority is required. For numerical phenotypes, users must designate whether higher or lower values are prioritized. More complex scenarios with multiple phenotypes and exposures of interest can be accommodated via composite scores, which facilitate prioritization based on a combination of traits (e.g., sex and body height), thereby allowing for nuanced selection within the study sample. Two additional user parameters are required: relatedness cutoff values (Kinship or IBS) and a fuzziness score. The relatedness cutoff value is used to sets the degree relatedness threshold and thereby tolerance towards the variability in the number of connections each subject has within the network. A fuzziness score (f) can be used to tune prioritization by setting the criteria for selection by assigning subjects who are related with m individuals and subjects who are related to m – f (f < m) individuals with the same kinship weight and prioritize the pruning of their relatedness network based on their corresponding phenotype weights. The score allows user granularity in determining which subjects are systematically excluded based on their phenotype prioritizations, e.g., prioritizing subjects with phenotype of interest versus maximizing set of unrelated subjects.

Pruning complex relatedness network

The first step of the relatedness pruning process is to identify sets of subjects that are related to each other. First, subjects that are not related with any other are removed from the kinship matrix. Next, pairs of subjects that are each related only to each other (related pairs) are identified and split according to the phenotype prioritization criteria (Figure 1A). The algorithm next proceeds stepwise increasing the relatedness group size to tackle more complex relatedness networks. To this end, KDPS employs two different strategies based on the user-specified fuzziness score. With a fuzziness score of zero, the program adopts a simplified approach by prioritizing the removal of super-subjects. Super-subjects act as hubs, linked to multiple subjects within the network that are otherwise not related to each other or any other subjects in the study. This targeted pruning is particularly effective in cohort studies, where such super-subjects can constitute a considerable portion of the network. Conversely, with a fuzziness score greater than zero, a greedy algorithm is employed, sequentially eliminating subjects with the lowest phenotypic weight who are related to more than m - f subjects, where m is the number of related pairs of the subject in the cohort who is related to the most people, and f represents the fuzziness score. This iterative removal continues until only unrelated pairs remain, which are then subjected to a final work-up step. The culmination of this process is a curated list of subjects to be excluded, that is used to construct the cohort of unrelated individuals.

Benchmarking the method performance using simulations

We evaluated the performance of the KDPS method via simulation and real-world scenarios. A complex related network (n~100 000 pairs) was simulated based on UK Biobank kinship structure and a simulated set of categorical and numerical phenotypes, detailed in Table 1. Furthermore, KDPS was applied to real-world phenotypes in the UK Biobank dataset, using four binary phenotypes: schizophrenia, acute myocardial infarction, multiple sclerosis and alcohol drinking (ever consumed alcohol). A fuzziness score of zero was used and results were compared between the approach where phenotypic prioritization was considered (phenotype-aware) and in which phenotypic information ignored (phenotype-naïve). All testing was performed using a single thread job (Intel(R) Xeon(R) CPU E5-4650 v3 @ 2.10GHz) and 4 Gb of memory.

**Results**

Simulation test results

The computational time required by KDPS exhibits a logarithmic dependency on the aggregate level of relatedness among subjects, as demonstrated in simulations with increasing number of relatedness, where fuzziness score was set at zero (Figure 1B). Similarly, the computational time increases logarithmically in relation to increasing fuzziness score, adhering to an approximate O(log(n)) complexity. Specifically, with a dataset comprising 50 000 subjects with over 10 000 relationship pairs the execution time of KDPS spans approximately 1.5 minutes at a fuzziness score of zero, increasing to over 10 minutes when the fuzziness score is set to 10 (Figure 1C). Notably, in simulations with ‘UK Biobank’ scale complex related network (~100 000 related pairs) KDPS computational time was under 15 minutes. Real world computation times may diverge, influenced by the intricacy of the relatedness network and additional factors such as the relatedness cutoff and fuzziness score.

In addition to increased computational demand, the choice of fuzziness score also subtly influences the retention ratio, defined as the percentage of subjects remaining after kinship decoupling compared to the original total. For instance, in a simulation involving 50 000 subjects, fuzziness scores range from zero to ten marginally reduces the retention ratio from 0.51 to 0.49 (Figure 1D). Moreover, we compared performance of phenotype selection and retention rates using KDPS and standard phenotype-agnostic approaches (e.g., PLINK2). Using the simulated datasets, KDPS demonstrated an advantage in increasing the prevalence of subjects possessing the phenotype of interest after kinship decoupling (Figure 1E). In simulations where the baseline prevalence of the phenotype of interest was set at 20%, KDPS significantly enhances this retention of subjects with phenotype of interest to approximately 30%, while the phenotype-naïve approach 20%.

Results for the real-world datasets

We next evaluated the efficacy of the KDPS in real-world scenarios, using UK Biobank cohort and varied set of outcomes and exposures, detailed in Table 2. KDPS phenotype aware selection preserved a significantly higher proportion of subjects possessing phenotypes of interest when juxtaposed against conventional phenotype-naïve methodologies. When prioritizing case subjects, KDPS resulted in 11.8% increase in the number of case subjects with schizophrenia, 11.1% increase in subjects with acute myocardial infarction, 12.1% increase in subjects with multiple sclerosis and 8.7% increase in subjects who have self-reported to have never consumed alcohol. KDPS successfully completed the decoupling and phenotype selection processes for all tested phenotypes in the UK Biobank within 35 minutes.

**Discussion**

In this report, we introduce KDPS, a novel tool and algorithm to address the lack of phenotype-aware kinship decoupling tools in genetic and epidemiological investigations. KDPS substantially improves over existing phenotype naïve selection methods by allowing incorporation of phenotypic information in subject selection. KDPS can be extended to allow complex and tailored sample selections via the use of composite trait scores (combination of traits and/or exposures of interest). KDPS has maximum utility in scenarios when the analytical method cannot accommodate relatedness and maximizing trait sample count is crucial to achieving the necessary statistical power.

Strengths of KDPS include efficiently of algorithm to process biobank-scale studies within a practical timeframe. KDPS accommodates a broad set of phenotypes for prioritization, including numeric (binary, ordinal and quantitative measures) and categorical phenotype definitions. Simulation and real-world applied analyses illustrate KDPS's computational efficiency and its capability to substantially conserve subjects with traits.

There are important considerations and limitations of KDPS. One potential challenge arises when dealing with datasets substantially more extensive and/or complex relatedness than UK Biobank (>100K related pairs), such as national biobanks and studies leveraging medical systems (e.g., AllofUS, UCLA ATLAS, BioVU, *etc*.)(All of Us Research Program Investigators et al. 2019, Johnson et al. 2023, Pulley et al. 2010). Computation times may significantly increase with sample size, and the amount and complexity of relatedness. However, this may not represent a major burden as sample selection is typically performed only once per study. Moreover, future improvements such as reimplementing KDPS in a lower-level programming languages or novel algorithms, may improve performance. Finally, users are cautioned that subject selection may to introduction of collider bias (Munafò et al. 2018). In brief, collider bias occurs when the selection of subjects is based on criteria that is associated with both the exposure and outcome of interest, potentially leading to spurious or artificial association between the exposure and outcome variables. Mitigation of collider bias should optimally be done at the design stage. Researchers are also advised to consider strategies such as conducting sensitivity analyses to check for collider bias, *e.g.*, repeating analyses in randomly selected subjects and compare to phenotype-selected results.

In conclusion, KDPS is a fast, computationally efficient, and powerful tool for phenotype-aware kinship decoupling, offering substantial improvements in both the inclusion of relevant subjects and computational efficiency. The integration of KDPS paves the way for phenotype-informed selection of unrelated subjects, offering broad applicability in genetics and epidemiology research.

**Data and code availability**

The latest release of KDPS and documentation can be found at https://github.com/UCSD-Salem-Lab/kdps. The KDPS R package can be installed via GitHub release. Scripts for the benchmarking are available at https://github.com/UCSD-Salem-Lab/kdps\_dev.

**Acknowledgements**

The authors extend their sincere gratitude to Dr. Erin Richard and Mijia Ma at the University of California, San Diego, for their invaluable support and insightful comments regarding both the methodology and the manuscript. This work was supported in part by awards from the Foundation of the National Institute of Health and National Institute of Health grants R00HL122515 and R01 DK135868 to R.M.S.

**Disclosures:** RMS reports a service contract with Travere. The authors have declared that no other conflict of interest exists.

**Conflict of interest**

The authors of the manuscript declare no conflict of interest.

**Funding**

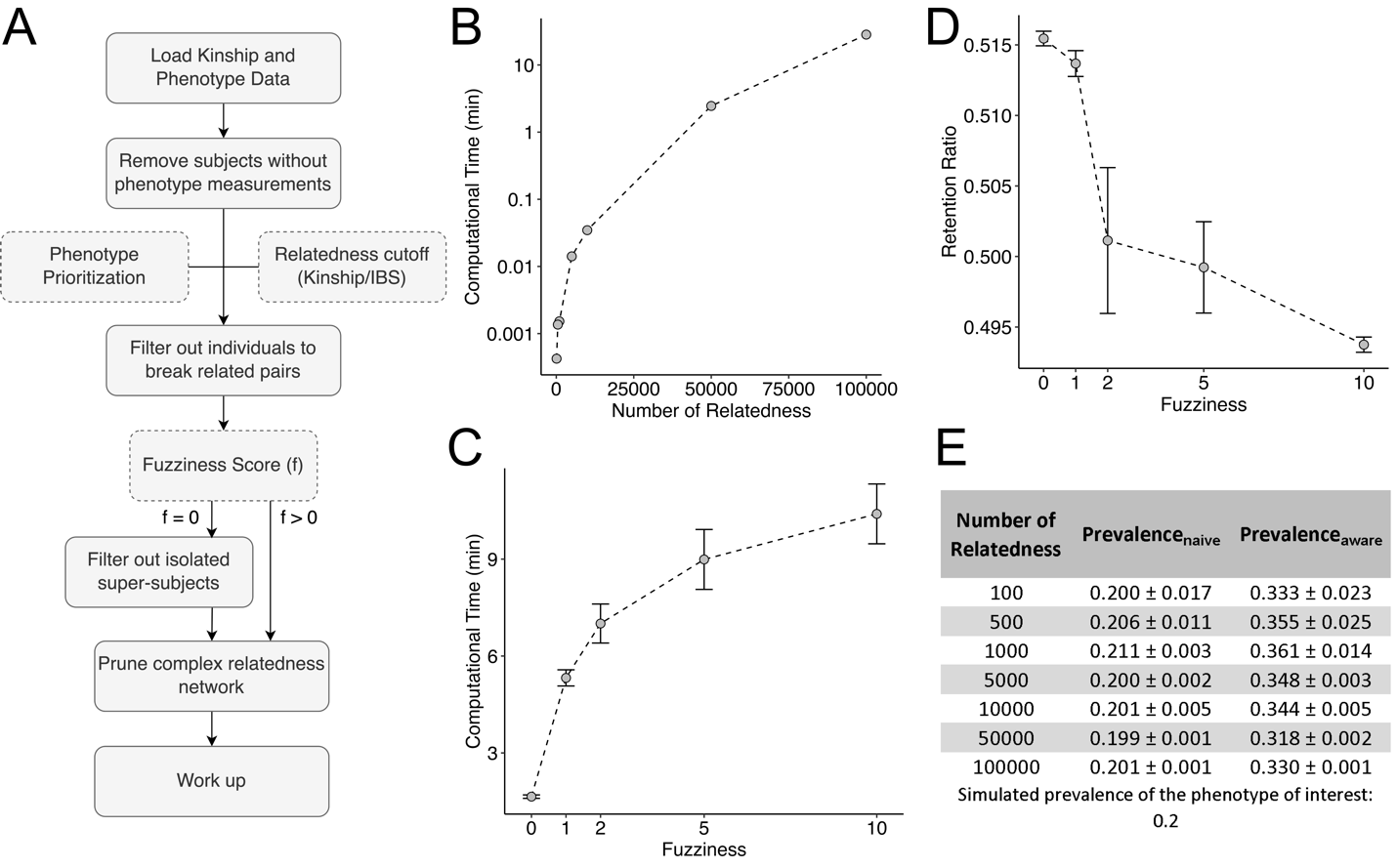
This study is partially funded by NIH R00HL122515 and R01 DK135868.

|  |  |  |
| --- | --- | --- |
| **Parameter** | **Description** | **Values** |
| Number of Relationships | Defined as the number of kinship relationships to consider in the analysis. | 100, 500, 1000, 5 000, 10 000, 50 000, 100 000 |
| Fuzziness | The degree of fuzziness allowed in the model. | 0, 1, 2, 5, 10 |
| Phenotypic Naïve | A Boolean parameter representing whether subjects are prioritized based on their phenotypic information. | FALSE, TRUE |
| Phenotype | Tested with three phenotype configurations (pheno1: binary; pheno2: categorical with three categories, pheno3: numerical). | pheno1, pheno2, pheno3 |

**Table 1. Simulation parameters used to run KDPS.**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Phenotype** | **Schizophrenia** | **Acute Myocardial Infarction** | **Multiple**  **Sclerosis** | **Alcohol Drinking (Never)** |
|  | UK Biobank ID | 130 874 | 131 298 | 131 042 | 20 117 |
| **Full Sample** | Sample Size | 502 420 | 502 420 | 502 420 | 502 420 |
| Case/Exposure | 1 356 | 22 635 | 2 500 | 22 382 |
| Control/Ref | 501 064 | 479 785 | 499 920 | 480 038 |
| Prevalence | 2.70E-03 | 4.51E-02 | 4.98E-03 | 4.45E-02 |
| **Phenotypic Naïve Selection** | Sample Size | 427 803 | 427 797 | 427 795 | 427 793 |
| Case/Exposure | 1 184 | 19 662 | 2 175 | 19815 |
| Control/Ref | 426 619 | 408 135 | 425 620 | 407 978 |
| Prevalence | 2.77E-03 | 4.60E-02 | 5.08E-03 | 4.63E-02 |
| **Phenotypic Aware Selection** | Sample Size | 427 817 | 427 819 | 427 817 | 427 822 |
| Case/Exposure | 1 324 | 21 836 | 2 439 | 21 547 |
| Control/Ref | 426 493 | 405 983 | 425 378 | 406 275 |
| Prevalence | 3.09E-03 | 5.10E-02 | 5.70E-03 | 5.04E-02 |
| % Additional Case/Exposure Subjects Included | 11.8% | 11.1% | 12.1% | 8.7% |

**Table 2. Results and performance of KDPS in real life scenarios using UK Biobank for 4 phenotypes of interest.**



**Figure 1. Algorithm flowchart and performance benchmark of KDPS.** (A) Flowchart illustration of KDPS algorithm steps (solid boxes), including user inputs (dashed boxes). (B) Benchmark results of the computational time of KDPS with respect to various numbers of relatedness present in the simulated cohort. (C) Benchmark results of the computational time of KDPS with respect to different user-specified fuzziness scores using simulated dataset based on UK Biobank. (D) Benchmark results of the retention ratio upon removing related subjects with respect to the fuzziness score. (E) Simulation results comparing the prevalence of case subjects between the phenotype-naïve (PLINK2) and the phenotype-aware (KDPS) approaches.

**References**

Abraham, Kuruvilla Joseph, and Diaz, Clara, “Identifying Large Sets of Unrelated Individuals and Unrelated Markers,” *Source Code for Biology and Medicine*, 9/1 (2014), 6

Akbari, Ali, Vitti, Joseph J., Iranmehr, Arya, Bakhtiari, Mehrdad, Sabeti, Pardis C., Mirarab, Siavash, et al., “Identifying the Favored Mutation in a Positive Selective Sweep,” *Nature Methods*, 15/4 (2018), 279–82

Ali-Khan, Sarah E., and Daar, Abdallah S., “Admixture Mapping: From Paradigms of Race and Ethnicity to Population History,” *The HUGO Journal*, 4/1–4 (2010), 23–34

All of Us Research Program Genomics Investigators, “Genomic Data in the All of Us Research Program,” *Nature*, 2024 <http://dx.doi.org/10.1038/s41586-023-06957-x>

All of Us Research Program Investigators, Denny, Joshua C., Rutter, Joni L., Goldstein, David B., Philippakis, Anthony, Smoller, Jordan W., et al., “The ‘All of Us’ Research Program,” *The New England Journal of Medicine*, 381/7 (2019), 668–76

Booker, Tom R., Jackson, Benjamin C., and Keightley, Peter D., “Detecting Positive Selection in the Genome,” *BMC Biology*, 15/1 (2017), 98

Bycroft, Clare, Freeman, Colin, Petkova, Desislava, Band, Gavin, Elliott, Lloyd T., Sharp, Kevin, et al., “The UK Biobank Resource with Deep Phenotyping and Genomic Data,” *Nature*, 562/7726 (2018), 203–9

Chang, Christopher C., Chow, Carson C., Tellier, Laurent Cam, Vattikuti, Shashaank, Purcell, Shaun M., and Lee, James J., “Second-Generation PLINK: Rising to the Challenge of Larger and Richer Datasets,” *GigaScience*, 4 (2015), 7

Edwards, Todd L., and Li, Chun, “Optimized Selection of Unrelated Subjects for Whole-Genome Sequencing Studies of Rare High-Penetrance Alleles,” *Genetic Epidemiology*, 36/5 (2012), 472–79

de Jager, Deon, Swarts, Petrus, Harper, Cindy, and Bloomer, Paulette, “Friends and Family: A Software Program for Identification of Unrelated Individuals from Molecular Marker Data,” *Molecular Ecology Resources*, 17/6 (2017) <https://pubmed.ncbi.nlm.nih.gov/28503747/> [accessed 28 February 2024]

Johnson, Ruth, Ding, Yi, Bhattacharya, Arjun, Knyazev, Sergey, Chiu, Alec, Lajonchere, Clara, et al., “The UCLA ATLAS Community Health Initiative: Promoting Precision Health Research in a Diverse Biobank,” *Cell Genomics*, 3/1 (2023), 100243

Loh, Po-Ru, Tucker, George, Bulik-Sullivan, Brendan K., Vilhjálmsson, Bjarni J., Finucane, Hilary K., Salem, Rany M., et al., “Efficient Bayesian Mixed-Model Analysis Increases Association Power in Large Cohorts,” *Nature Genetics*, 47/3 (2015), 284–90

Manichaikul, Ani, Mychaleckyj, Josyf C., Rich, Stephen S., Daly, Kathy, Sale, Michèle, and Chen, Wei-Min, “Robust Relationship Inference in Genome-Wide Association Studies,” *Bioinformatics* , 26/22 (2010), 2867–73

Munafò, Marcus R., Tilling, Kate, Taylor, Amy E., Evans, David M., and Davey Smith, George, “Collider Scope: When Selection Bias Can Substantially Influence Observed Associations,” *International Journal of Epidemiology*, 47/1 (2018), 226–35

Ottman, R., “Gene-Environment Interaction: Definitions and Study Designs,” *Preventive Medicine*, 25/6 (1996), 764–70

Pulley, Jill, Clayton, Ellen, Bernard, Gordon R., Roden, Dan M., and Masys, Daniel R., “Principles of Human Subjects Protections Applied in an Opt-out, de-Identified Biobank,” *Clinical and Translational Science*, 3/1 (2010), 42–48

Su, Shu-Yi, Kasberger, Jay, Baranzini, Sergio, Byerley, William, Liao, Wilson, Oksenberg, Jorge, et al., “Detection of Identity by Descent Using Next-Generation Whole Genome Sequencing Data,” *BMC Bioinformatics*, 13 (2012), 121

Virolainen, Samuel J., VonHandorf, Andrew, Viel, Kenyatta C. M. F., Weirauch, Matthew T., and Kottyan, Leah C., “Gene-Environment Interactions and Their Impact on Human Health,” *Genes and Immunity*, 24/1 (2023), 1–11

Voight, Benjamin F., and Pritchard, Jonathan K., “Confounding from Cryptic Relatedness in Case-Control Association Studies,” *PLoS Genetics*, 1/3 (2005), e32

Winkler, Thomas W., Justice, Anne E., Cupples, L. Adrienne, Kronenberg, Florian, Kutalik, Zoltán, Heid, Iris M., et al., “Approaches to Detect Genetic Effects That Differ between Two Strata in Genome-Wide Meta-Analyses: Recommendations Based on a Systematic Evaluation,” *PloS One*, 12/7 (2017), e0181038

Zhou, Wei, Nielsen, Jonas B., Fritsche, Lars G., Dey, Rounak, Gabrielsen, Maiken E., Wolford, Brooke N., et al., “Efficiently Controlling for Case-Control Imbalance and Sample Relatedness in Large-Scale Genetic Association Studies,” *Nature Genetics*, 50/9 (2018), 1335–41