KDPS: a tool for 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.

2Weill Institute for Neurosciences, Department of Neurology, University of California San Francisco, San Francisco, CA 94158.

**Word count**: 3,081

**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; they 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 relatedness (kinship or identity by descent) scores and allows 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 quantitative and categorical, 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 the ability of this method to process biobank-scale studies within practical timeframes and 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.

**Keywords**: Subject relatedness, kinship decoupling, sample selection, phenotype prioritization, genetic epidemiology, greedy algorithm

**Key points**:

* Kinship decoupling
* Phenotype-aware sample selection
* Applicability to biobank-scale genomic and epidemiology research

**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 [1,2], a significant number of genetic analysis methods require studies of unrelated individuals, including: selection scans [3,4], admixture mapping [5], gene by environment interaction analysis [6,7], and analyses across stratum [8].

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 [9], especially in large-scale initiatives like the UK Biobank [10] and All of US research program [11]. Cryptic relatedness occurs when two or more genetically related individuals are included in a study, and the genetic relationship is unrecognized. This unrecognized relatedness can skew results by artificially inflating genetic similarities or associations, leading to biased estimates of genetic effects. 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 [12], PLINK2 [13], Friends and family [14], SampleSeq2 [15] and FastIndep [16]. 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 a 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 [12] or identity by descent (IBS) score [17] 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 (Supplementary Figure s1). Given a higher fuzziness score, the algorithm downplays the importances of complex network topology, de-prioritize the removal of individuals related to many other subjects and prioritize the removal of individuals with lower phenotypic weights. With lower fuzziness score, the algorithm adheres more closely to the network topology and prioritize the removal of individuals that are related to more subjects to minimize the number of subjects that need to be removed. 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. After each iteration, the algorithm recalculates m to reflect the current maximum number of relationships in the updated network, ensuring that the pruning criterion m − f dynamically adapts to the evolving structure of the cohort. This iterative update continues until no subject exceeds the relatedness threshold and helps maintain optimal pruning sensitivity as the network is refined. Eventually this iterative removal completes when 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. For simulations, a complex relatedness network (~100,000 relationship pairs) was considered based on the UK Biobank kinship matrix. The matrix was anonymized by removing subject identifiers and randomly permuting individual labels, preserving the underlying topology and relationship structure while ensuring de-identification. To understand the performance of KDPS on phenotypic traits of different modalities, we simulated phenotype data included three configurations: a binary trait, a categorical trait with three levels, and a quantitative trait drawn from a mean-centered, normally distributed range. Detailed simulation parameters can be found in Table 1.

To assess the impact of heritability on subject retention, simulated phenotypic traits were assigned to subjects in a relatedness network in a stepwise manner. First, seed cases were randomly assigned at 10% prevalence across all individuals. Phenotypes were then propagated to genetically related individuals using a parameter termed the heritability indicator. For a given case individual, each of its relatives inherited the phenotype with probability equal to:

Thus, at a heritability multiplier of 0, no related individual could inherit the phenotype from a seed case. At small values (e.g., 0.25 or 0.5), inheritance probabilities were modest, while at very high values (e.g., ≥15), nearly all relatives of a case were assigned the phenotype. This heuristic design allowed us to model a spectrum of traits from non-heritable to strongly heritable. Phenotypes were simulated at heritability multipliers of 0, 0.25, 0.5, 0.75, 1, 5, 10, 15, and 20. For each setting, three replicates were generated using fixed seeds for reproducibility. The KDPS algorithm was then applied with fuzziness = 0 to remove related individuals. Following pruning, we calculated both the subject retention ratio (subjects retained / total subjects) and the case retention ratio (cases retained / total cases).

Furthermore, KDPS was applied to four real-world phenotypes from the UK Biobank: schizophrenia (UKB ID 130874), acute myocardial infarction (UKB ID 131298), multiple sclerosis (UKB ID 131042), and alcohol drinking status (never drinkers, UKB ID 20117). Schizophrenia was defined using ICD-10 F20 diagnoses from hospital records (Fields 41202/41204), death registries, primary care, and self-report (Field 20002). Acute myocardial infarction was identified from hospital and death records (ICD-10 I21–I22; Fields 41202/41204, 40001/40002) and self-report (Field 20002). Multiple sclerosis was captured via “first occurrence” fields (131042, 131043; ICD-10 G35) along with self-report and hospital data. Alcohol status was derived from questionnaire Field 20117 distinguishing never versus ever drinkers. These phenotypic traits were extracted and harmonized with respect to the pre-calculated pair-wise kinship coefficients for all UK Biobank individuals. 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

In both simulations and real-world applications, KDPS successfully pruned the complex relatedness networks, resulting in a final dataset of entirely unrelated individuals after kinship decoupling. Because KDPS only requires kinship matrix files and phenotype files as inputs, its memory footprint remained modest, never exceeding 4 GB of RAM. This makes the method practical and accessible on most consumer-grade computers and standard workstations. 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). In biobank scale cohorts such as the UK Biobank, most relatedness is expected to be pair-wise relationships instead of complex relatedness networks (Supplementary Figure S2). Accordingly, KDPS can resolve most relatedness scenarios with a fuzziness score of 0, which offers a practical default for balancing decoupling and phenotype retention. In cases involving complex relatedness networks or ultra-rare phenotypes, a higher fuzziness score (e.g., 3 - 5) may be warranted to prioritize phenotype over topology.

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 binary 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%.

In the evaluation of the performance of KDPS on multi-class categorical phenotypes, compared to phenotype-naïve pruning, phenotype-aware KDPS increased the retention of disease-relevant individuals by ~79% for disease 1 and ~56% for disease 2, demonstrating substantial gains in preserving prioritized classes (Supplementary Table s1). When KDPS was applied on a continuous phenotype using a simulated normally distributed quantitative trait, compared to phenotype-naïve pruning, phenotype-aware KDPS resulted in modest upward shifts across the distribution of retained subjects. The minimum phenotype value increased by 2%, and the mean by 0.09% (Supplementary Table s1). In more complex scenarios where multiple phenotypes of interest are involved, KDPS also demonstrated the capability to maximize targeted subject retainment based on a composite weight. A simulation involving two independent binary phenotypes (~20% prevalence each) showed that applying composite weights, prioritizing subjects with both traits, resulted in a 42% (19 to 27) increase in the number of retained individuals with both conditions compared to equal-weight pruning (Supplementary Table S2).

Additionally, we examined the impact of trait heritability on subject and case retention. Simulation results demonstrated that subject retention ratio remained stable across all levels of heritability (Supplementary Figure S3A), reflecting the fact that overall pruning is determined primarily by network structure rather than phenotype assignment. Whereas, as expected, case retention ratio decreased as heritability increased. At low heritability (multiplier = 0), seed cases were scattered randomly, and many were retained after pruning. However, at higher heritability multipliers, cases increasingly clustered within related families. To fully decouple related individuals, KDPS necessarily removed larger proportions of cases, leading to reduced case retention (Supplementary Figure S3B).

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. In the real-world examples, KDPS was applied to phenotypes with diverse genetic architectures and heritability values, including schizophrenia (heritability (H2)~80%)[18], multiple sclerosis (H2~30%)[19], acute myocardial infarction (H2 ~40–50%)[20,21], and alcohol drinking status (~20–30%)[22,23], all of which showed strong phenotype retention performance. Moreover, the use of composite weights enables highly flexible prioritization strategies, allowing users to specify phenotype combinations such as categorical values of a particular type (e.g., case status) in addition to a numeric variable within a defined range (e.g., BMI between 18–25). This capacity broadens the applicability of KDPS to complex study designs, enabling tailored subject retention across diverse phenotype-driven analytical objectives.

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*.)[24–26]. 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 language and/or novel algorithms, may improve performance. While we considered a diverse set of real and simulated phenotypes, this set is not exhaustive, we expect the generalizable framework of KDPS should apply broadly to phenotypes with varying prevalence and genetic contribution. Additionally, it is also important to consider population structure in the generation of the relatedness matrix. In ancestrally diverse or admixed populations, standard IBD or kinship estimation methods may be inaccurate or biased due to the confounding effect of genetic admixture [27]. Users are advised to select appropriate methods that account for ancestry when generating the subject relatedness matrix [28,29]. Finally, users are cautioned that phenotype-based subject selection may to introduction of collider bias [30]. 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://ucsd-salem-lab.github.io/kdps/. The KDPS R package can be installed via the Comprehensive R Archive Network (CRAN) and 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.

**References**

1. Loh P-R, Tucker G, Bulik-Sullivan BK, et al. Efficient Bayesian mixed-model analysis increases association power in large cohorts. Nat. Genet. 2015; 47:284–290

2. Zhou W, Nielsen JB, Fritsche LG, et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. Nat. Genet. 2018; 50:1335–1341

3. Akbari A, Vitti JJ, Iranmehr A, et al. Identifying the favored mutation in a positive selective sweep. Nat. Methods 2018; 15:279–282

4. Booker TR, Jackson BC, Keightley PD. Detecting positive selection in the genome. BMC Biol. 2017; 15:98

5. Ali-Khan SE, Daar AS. Admixture mapping: from paradigms of race and ethnicity to population history. Hugo J. 2010; 4:23–34

6. Ottman R. Gene-environment interaction: definitions and study designs. Prev. Med. 1996; 25:764–770

7. Virolainen SJ, VonHandorf A, Viel KCMF, et al. Gene-environment interactions and their impact on human health. Genes Immun. 2023; 24:1–11

8. Winkler TW, Justice AE, Cupples LA, 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 2017; 12:e0181038

9. Voight BF, Pritchard JK. Confounding from cryptic relatedness in case-control association studies. PLoS Genet. 2005; 1:e32

10. Bycroft C, Freeman C, Petkova D, et al. The UK Biobank resource with deep phenotyping and genomic data. Nature 2018; 562:203–209

11. All of Us Research Program Genomics Investigators. Genomic data in the All of Us Research Program. Nature 2024;

12. Manichaikul A, Mychaleckyj JC, Rich SS, et al. Robust relationship inference in genome-wide association studies. Bioinformatics 2010; 26:2867–2873

13. Chang CC, Chow CC, Tellier LC, et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience 2015; 4:7

14. de Jager D, Swarts P, Harper C, et al. Friends and family: A software program for identification of unrelated individuals from molecular marker data. Mol. Ecol. Resour. 2017; 17:

15. Edwards TL, Li C. Optimized selection of unrelated subjects for whole-genome sequencing studies of rare high-penetrance alleles. Genet. Epidemiol. 2012; 36:472–479

16. Abraham KJ, Diaz C. Identifying large sets of unrelated individuals and unrelated markers. Source Code Biol. Med. 2014; 9:6

17. Su S-Y, Kasberger J, Baranzini S, et al. Detection of identity by descent using next-generation whole genome sequencing data. BMC Bioinformatics 2012; 13:121

18. Sullivan PF, Daly MJ, O’Donovan M. Genetic architectures of psychiatric disorders: the emerging picture and its implications. Nat. Rev. Genet. 2012; 13:537–551

19. International Multiple Sclerosis Genetics Consortium. Multiple sclerosis genomic map implicates peripheral immune cells and microglia in susceptibility. Science 2019; 365:eaav7188

20. Marenberg ME, Risch N, Berkman LF, et al. Genetic susceptibility to death from coronary heart disease in a study of twins. N. Engl. J. Med. 1994; 330:1041–1046

21. Inouye M, Abraham G, Nelson CP, et al. Genomic risk prediction of coronary artery disease in 480,000 adults: Implications for primary prevention. J. Am. Coll. Cardiol. 2018; 72:1883–1893

22. Verhulst B, Neale MC, Kendler KS. The heritability of alcohol use disorders: a meta-analysis of twin and adoption studies. Psychol. Med. 2015; 45:1061–1072

23. Clarke T-K, Adams MJ, Davies G, et al. Genome-wide association study of alcohol consumption and genetic overlap with other health-related traits in UK Biobank (N=112 117). Mol. Psychiatry 2017; 22:1376–1384

24. All of Us Research Program Investigators, Denny JC, Rutter JL, et al. The “All of Us” Research Program. N. Engl. J. Med. 2019; 381:668–676

25. Johnson R, Ding Y, Bhattacharya A, et al. The UCLA ATLAS Community Health Initiative: Promoting precision health research in a diverse biobank. Cell Genom 2023; 3:100243

26. Pulley J, Clayton E, Bernard GR, et al. Principles of human subjects protections applied in an opt-out, de-identified biobank. Clin. Transl. Sci. 2010; 3:42–48

27. Dou J, Sun B, Sim X, et al. Estimation of kinship coefficient in structured and admixed populations using sparse sequencing data. PLoS Genet. 2017; 13:e1007021

28. Thornton T, Tang H, Hoffmann TJ, et al. Estimating kinship in admixed populations. Am. J. Hum. Genet. 2012; 91:122–138

29. Conomos MP, Reiner AP, Weir BS, et al. Model-free estimation of recent genetic relatedness. Am. J. Hum. Genet. 2016; 98:127–148

30. Munafò MR, Tilling K, Taylor AE, et al. Collider scope: when selection bias can substantially influence observed associations. Int. J. Epidemiol. 2018; 47:226–235

**List of figures:**

**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.