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.

A diagram of a graph

AI-generated content may be incorrect.

**Supplementary Figure S1. Impact of fuzziness score on phenotype-aware pruning decisions in a representative relatedness network.**

This figure illustrates a simplified example of KDPS pruning under increasing fuzziness scores (0–3), demonstrating how subject removal decisions are influenced by both the degree of relatedness and phenotypic priority. Each node represents a subject; edges represent pairwise genetic relatedness. Node color and label reflect removal priority based on number of relatedness: with green (lower number) indicating low removal priority, and pink (higher number) indicating high removal priority. At a fuzziness score of 0, KDPS strictly prioritizes removal based on relatedness: subjects with the highest number of connections (“relatedness hubs”) are removed first, with phenotype only considered when ties in relatedness occur. As fuzziness increases to 1 and 2, KDPS begins to down-weight the importance of being a hub by grouping together subjects whose relatedness degrees are within f of the maximum (i.e., ≥ m − f). Within these groups, phenotype weight drives pruning decisions. At fuzziness = 3, all subjects are considered equally in terms of network position (as all have degree ≤ m − f), and KDPS removes subjects solely based on their phenotype ranking. This figure represents only a snapshot within the iterative pruning process: following each pruning step, m is recalculated and the threshold m − f is updated dynamically to guide subsequent removals. This design enables KDPS to flexibly balance between structural (relatedness) and phenotypic priorities during decoupling.

A graph with numbers and a bar

AI-generated content may be incorrect.

**Supplementary Figure S2. Distribution of kinship degrees among related subjects in the UK Biobank.**

This histogram summarizes the number of kinship relationships observed across different kinship degrees in the UK Biobank cohort. The kinship degree reflects the number of subjects to which a given individual is genetically related (i.e., node degree in the relatedness network). Many observed relationships are pairwise only, with 106 290 subjects exhibiting a kinship degree of 1. These correspond to individuals involved in a single related pair, forming isolated dyads. Only 30 088 subjects, less than one third of 106 290, were related to two other individuals and those related to three subjects total 8 114, representing progressively smaller fractions of the cohort. The number of subjects drops steeply with each additional connection, with only 18 individuals observed to have 10 related connections. This pattern highlights that most relatedness in large-scale cohorts like the UK Biobank occurs as isolated pairs, rather than as dense relatedness clusters or hubs. Consequently, KDPS can effectively resolve most relatedness scenarios using phenotype-aware pruning of pairwise relationships alone, and a default fuzziness score of 0 is typically sufficient to achieve optimal balance between decoupling and phenotype retention. In rare cases where subjects are embedded in larger, more complex relatedness structures, or when retaining subjects with ultra-rare phenotypes is essential, a higher fuzziness score (e.g., 3 to 5) may be warranted to down-weight topological importance and emphasize phenotype-based prioritization. Nonetheless, this empirical distribution supports the general use of fuzziness = 0 as the default setting for large-scale studies focused on unrelated subject selection.

|  |  |  |  |
| --- | --- | --- | --- |
| Categorical Data | | | |
|  | Disease 1 | Disease 2 | Healthy |
| Original Data | 1897 | 3697 | 12927 |
| Phenotypic Naïve | 979 | 1888 | 6764 |
| Phenotypic Informed | 1750 | 2944 | 4858 |
|  |  |  |  |
| Continuous Numerical Data | | | |
|  | Min | Mean | Max |
| Original Data | 61.85 | 100 | 140.9 |
| Phenotypic Naïve | 61.85 | 99.96 | 140.9 |
| Phenotypic Informed | 63.07 | 100.05 | 140.9 |

**Supplementary Table S1. Simulation results demonstrating phenotype-specific retention improvements with KDPS across categorical and continuous traits.**

This table summarizes the outcomes of simulation-based evaluations comparing phenotype-naïve versus phenotype-informed kinship decoupling using KDPS. The top panel presents results for a three-level categorical phenotype (Disease 1, Disease 2, and Healthy). In this setting, phenotype-aware pruning using KDPS retained ~79% more subjects with Disease 1 and ~56% more with Disease 2 compared to the phenotype-naïve approach, demonstrating KDPS’s capacity to prioritize and preserve individuals with traits of interest. The bottom panel presents results for a continuous quantitative phenotype. When prioritizing individuals with higher trait values, KDPS increased the minimum retained value by 2% and the mean by 0.09% relative to phenotype-naïve pruning. These results confirm KDPS’s flexibility and effectiveness in supporting both categorical and continuous phenotypic prioritization.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Original Data | | | |  |  | | | |
|  | | Condition 2 | |
| NO | YES |
| Condition 1 | NO | 1149 | 305 |
| YES | 260 | 78 |
| Prioritize based on equal weight | | | | Prioritize based on condition 2 | | | |
|  | | Condition 2 | |  | | Condition 2 | |
| NO | YES | NO | YES |
| Condition 1 | NO | 358 | 97 | Condition 1 | NO | 361 | 94 |
| YES | 91 | 19 | YES | 87 | 23 |
| Prioritize based on condition 1 | | | | Prioritize based on composite weight | | | |
|  | | Condition 2 | |  | | Condition 2 | |
| NO | YES | NO | YES |
| Condition 1 | NO | 358 | 96 | Condition 1 | NO | 358 | 92 |
| YES | 90 | 21 | YES | 87 | 27 |

**Supplementary Table S2. Simulation results demonstrating the effect of different phenotype prioritization strategies on subject retention using KDPS.**

This table illustrates a simulated example of KDPS applied to a synthetic dataset of related individuals annotated with two independent binary phenotypes: Condition 1 and Condition 2, each occurring in ~20% of subjects. The top panel presents the original distribution of subjects across the four phenotype combinations. KDPS was applied using four different prioritization strategies to evaluate their influence on post-decoupling phenotype composition. Equal Weight: All subjects were assigned the same phenotypic priority (weight = 1 or Phenotypic Naïve), treating phenotypes as irrelevant during decoupling. Prioritize Condition 1: Subjects with Condition 1 = YES were prioritized for retention over those without. Prioritize Condition 2: Subjects with Condition 2 = YES were prioritized similarly. Composite Weight: A phenotype-aware weighting scheme was constructed based on the combination of the two conditions: 1) Subjects with both conditions received a weight of 4. 2) Subjects with only Condition 1 received a weight of 3. 3) Subjects with only Condition 2 received a weight of 2. 4) Subjects with neither condition received a weight of 1. The table shows the resulting post-pruning counts of each phenotype combination under each prioritization scheme. Notably, the number of retained individuals with both conditions (Condition 1 = YES and Condition 2 = YES) increased from 19 under equal weighting to 27 under composite weighting, demonstrating the utility of composite phenotypic weighting in maximizing retention of highly valuable subjects. This example showcases how KDPS can be flexibly extended to prioritize subjects using complex, user-defined phenotypic combinations. While demonstrated here with binary phenotypes, the same principle applies to ordinal, categorical, or continuous traits, and supports the use of nonlinear functions to define customized composite scores tailored to study-specific objectives.