COLUMBIA
COLUMBIA UNIVERSITY
DEPARTMENT OF
BIOMEDICAL INFORMATICS

kirill.nikitin@columbia.edu
@ni_kirill

PRIVATE INFORMATION LEAKAGE FROM POLYGENIC RISK SCORES

Kirill Nikitin *Gamze Gürsoy*

NEW YORK GENOME CENTER
G²LAB

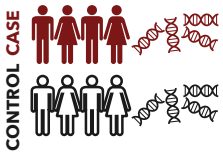
SUMMARY

Polygenic Risk Scores (PRSs) are a popular clinical tool for the estimation of an individual's genetic susceptibility to diseases. We show that a profit-seeking actor, e.g., an insurance company, can recover the associated genotypes from publicly shared PRSs and use them to either de-anonymize individuals or infer sensitive traits of known patients.

BACKGROUND

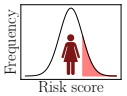
Genotypes as the difference count from the reference

GENOME-WIDE ASSOCIATION STUDIES



	SNP₁	SNP₂	SNP₃
Effect weight β	0.759	0.123	0.611

$$PRS = \frac{\sum_{i=1}^{N} \beta_i \times G_i}{Ploidy \times N}$$

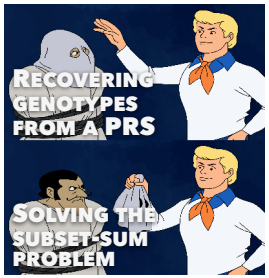


PRS example

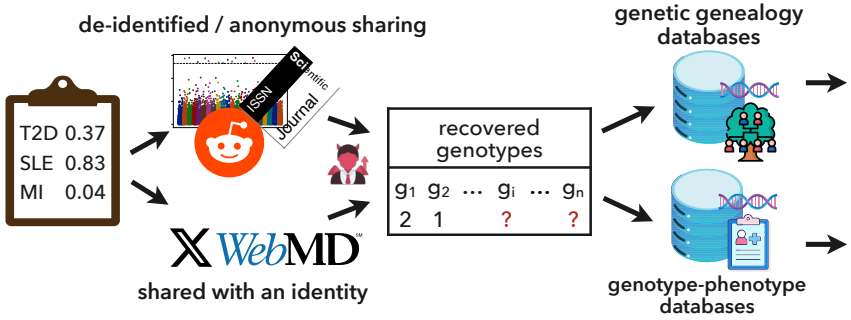
```
##POLYGENIC SCORE (PGS) INFORMATION
#pgs_id=PGS000073
#trait_reported=Cervical cancer
#variants_number=10
#weight_type=log(OR)
rsID      chr_name chr_position effect  other.  effect_weight
rs3130196.    6      33063219     C       T     0.3576744442718159

rs3132461.    6      31480668     G       A     0.3074846997479607

rs2523557     6      31331257     G       A     -0.35065687161316933
```



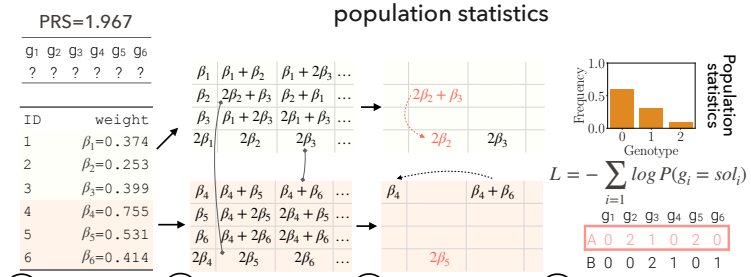
RISK SCENARIOS



GENOTYPE RECOVERY

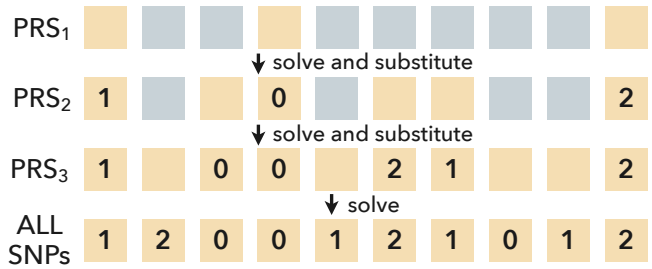
Given a PRS, we need to find how many times it includes each effect weight β

➡ Use dynamic programming and population statistics



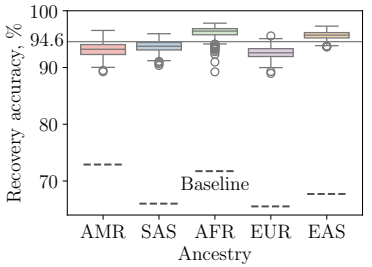
1. Split in half and calculate all the subset sums ≤ PRS
2. Find pairs of sums that add up to PRS
3. Backtrack the sums to find solution weights
4. Rank solutions based on their likelihood

SOLVING A PRS PANEL VIA PRS CHAINING



EVALUATION

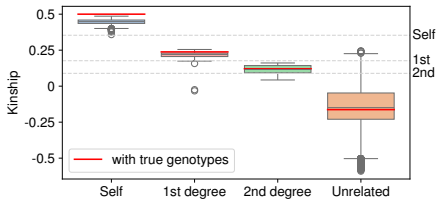
GENOTYPE RECOVERY ACCURACY



- Panel of 298 PRSs (2 to 49 SNPs) 4,821 SNPs total, 2,654 unique
- 2,535 genome samples from the 1000 Genomes Project
- We calculate PRSs and then attempt to recover genotypes
- 2,600 guessed SNPs on average

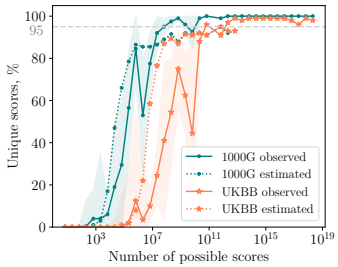
At least 454 out of 4,723 PRS models in the PGS Catalog are vulnerable to genotype recovery

GENEALOGY DATABASE SEARCH



- 68 samples with 1st or 2nd kins
- 100% accuracy identifying individuals and >80% accuracy identifying relatives

DATASET LINKING



In the UK Biobank dataset with 500K samples, a PRS model with 27 SNPs leads to 95% unique scores