—MAGNIFIC—

Maximizing Genetic Findings and Prediction

Florian Privé

Aarhus University (DK)

privefl

About me

About me

- [2013–2016] Engineer in Informatics and Applied Mathematics (ENSIMAG)

About me

- [2013–2016] Engineer in Informatics and Applied Mathematics (ENSIMAG)
- [2016–2019] PhD with Michael Blum (TIMC, Grenoble)
  and Hugues Aschard (Institut Pasteur, Paris),
  developing statistical learning methods in human genetics

About me

- [2013–2016] Engineer in Informatics and Applied Mathematics (ENSIMAG)
- [2016–2019] PhD with Michael Blum (TIMC, Grenoble)
  and Hugues Aschard (Institut Pasteur, Paris),
  developing statistical learning methods in human genetics
- [2017–2019] Founder and organizer of the R user group of Grenoble

About me

- [2013–2016] Engineer in Informatics and Applied Mathematics (ENSIMAG)
- [2016–2019] PhD with Michael Blum (TIMC, Grenoble)
  and Hugues Aschard (Institut Pasteur, Paris),
  developing statistical learning methods in human genetics
- [2017–2019] Founder and organizer of the R user group of Grenoble
- [2018–] Teaching an advanced R course for PhD students

About me

- [2013–2016] Engineer in Informatics and Applied Mathematics (ENSIMAG)
- [2016–2019] PhD with Michael Blum (TIMC, Grenoble)
  and Hugues Aschard (Institut Pasteur, Paris),
  developing statistical learning methods in human genetics
- [2017–2019] Founder and organizer of the R user group of Grenoble
- [2018–] Teaching an advanced R course for PhD students
- [2019–2022] Postdoc at Aarhus University (Denmark)
- [2022–] Senior Researcher (same place)

About me

- [2013–2016] Engineer in Informatics and Applied Mathematics (ENSIMAG)
- [2016–2019] PhD with Michael Blum (TIMC, Grenoble)
  and Hugues Aschard (Institut Pasteur, Paris),
  developing statistical learning methods in human genetics
- [2017–2019] Founder and organizer of the R user group of Grenoble
- [2018–] Teaching an advanced R course for PhD students
- [2019–2022] Postdoc at Aarhus University (Denmark)
- [2022–] Senior Researcher (same place)
- [2021–] Working remotely from France (near Lyon)

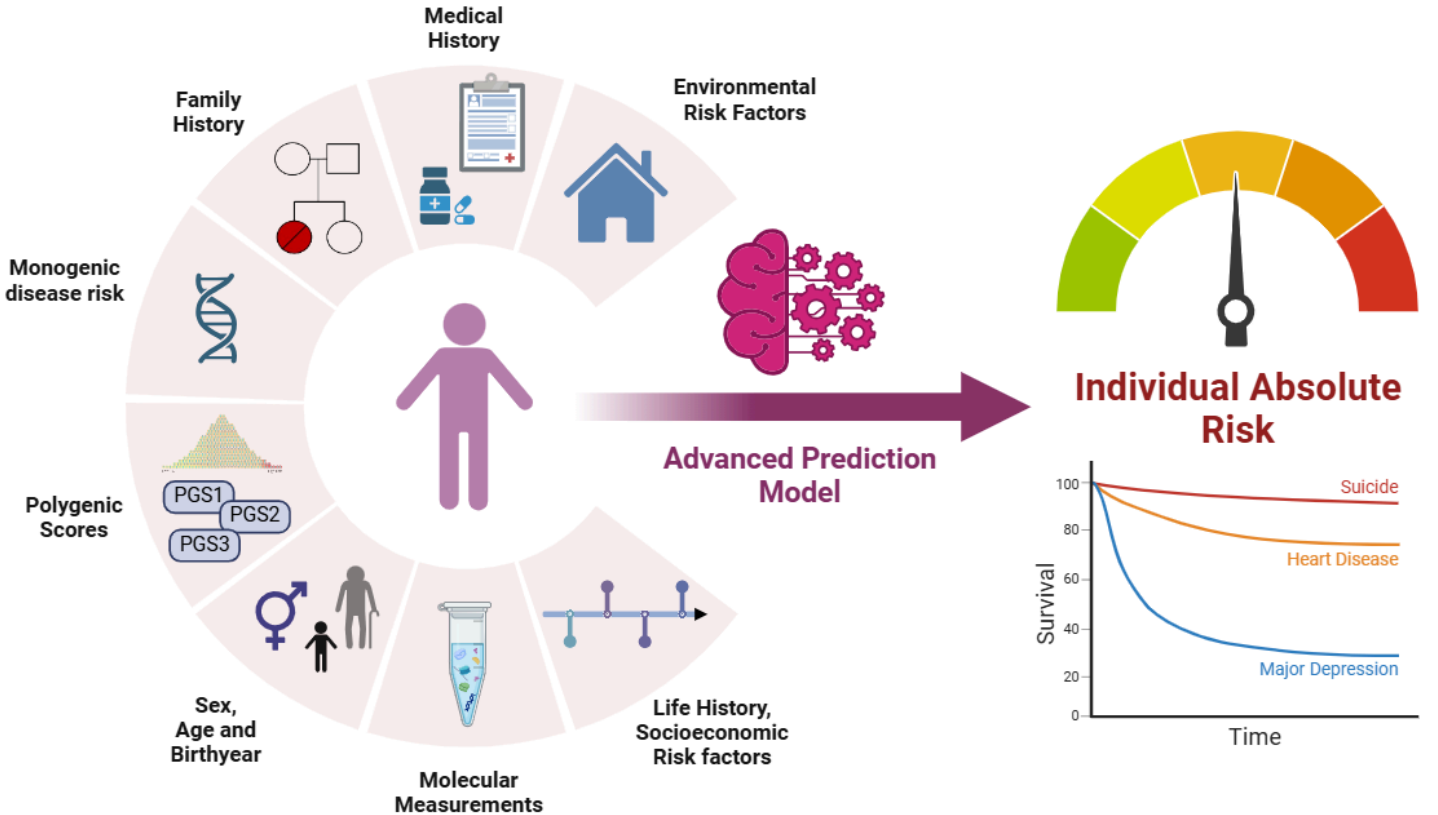
About me

- [2013–2016] Engineer in Informatics and Applied Mathematics (ENSIMAG)
- [2016–2019] PhD with Michael Blum (TIMC, Grenoble)
  and Hugues Aschard (Institut Pasteur, Paris),
  developing statistical learning methods in human genetics
- [2017–2019] Founder and organizer of the R user group of Grenoble
- [2018–] Teaching an advanced R course for PhD students
- [2019–2022] Postdoc at Aarhus University (Denmark)
- [2022–] Senior Researcher (same place)
- [2021–] Working remotely from France (near Lyon)
- Plan to apply to be a CRCN

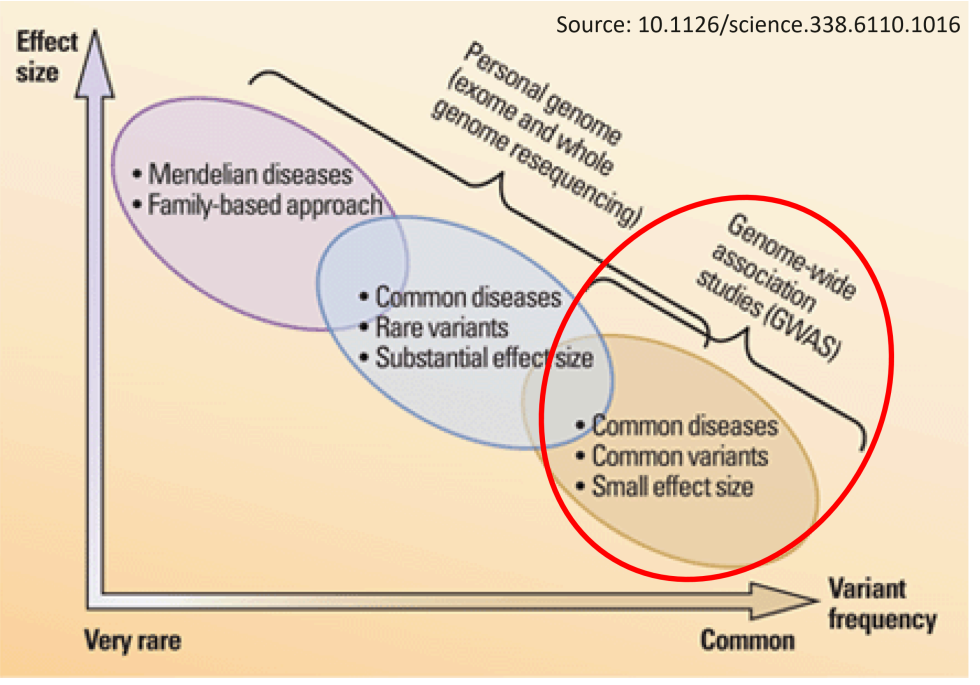
Introduction & Motivation

Personalized medicine: best predict disease risk

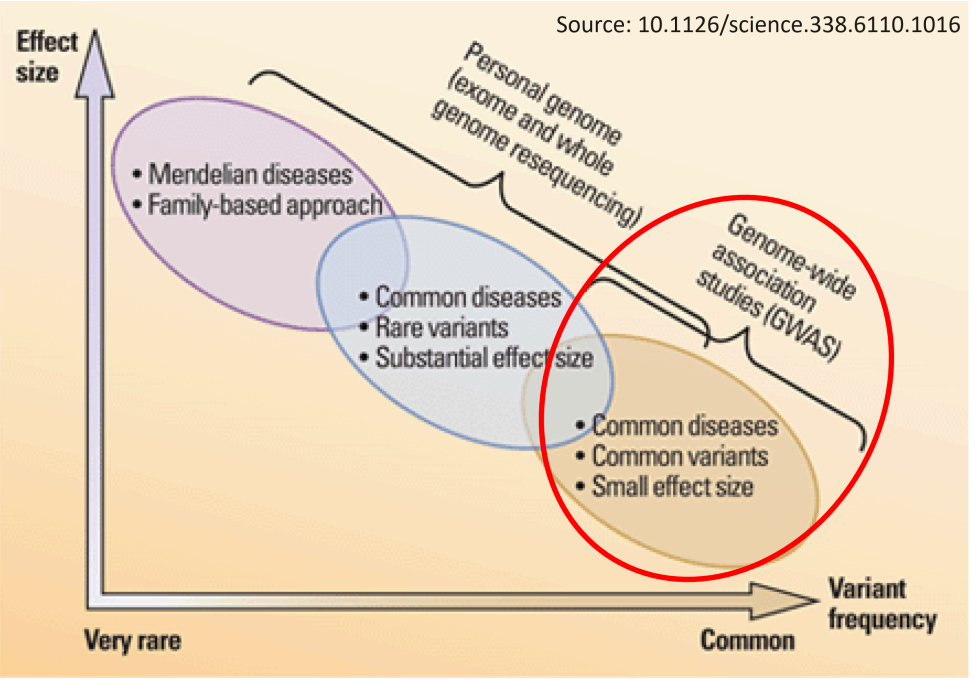
A polygenic score is a new risk factor (a very interesting one!)



Disease genetic architecture

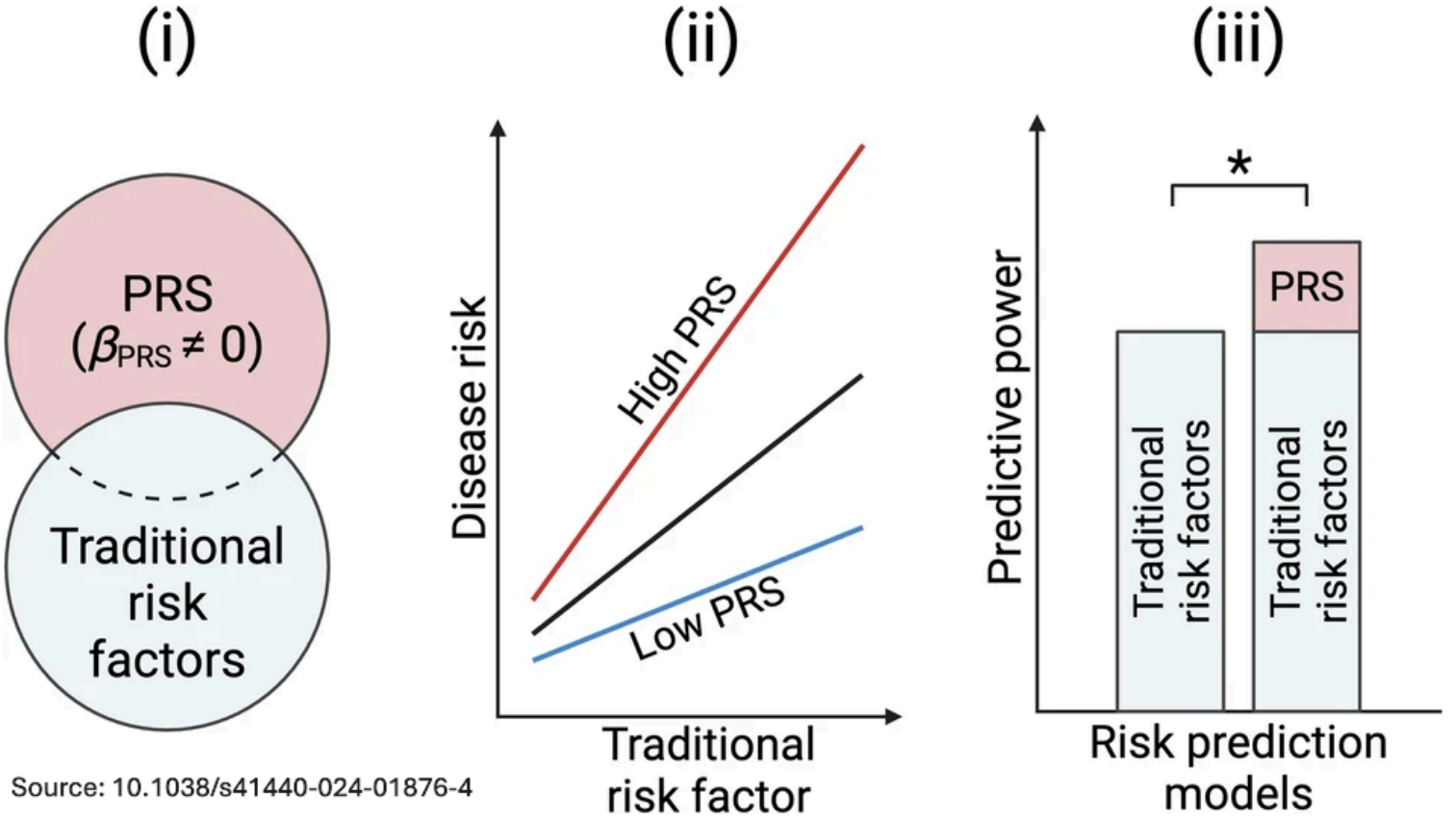


Disease genetic architecture



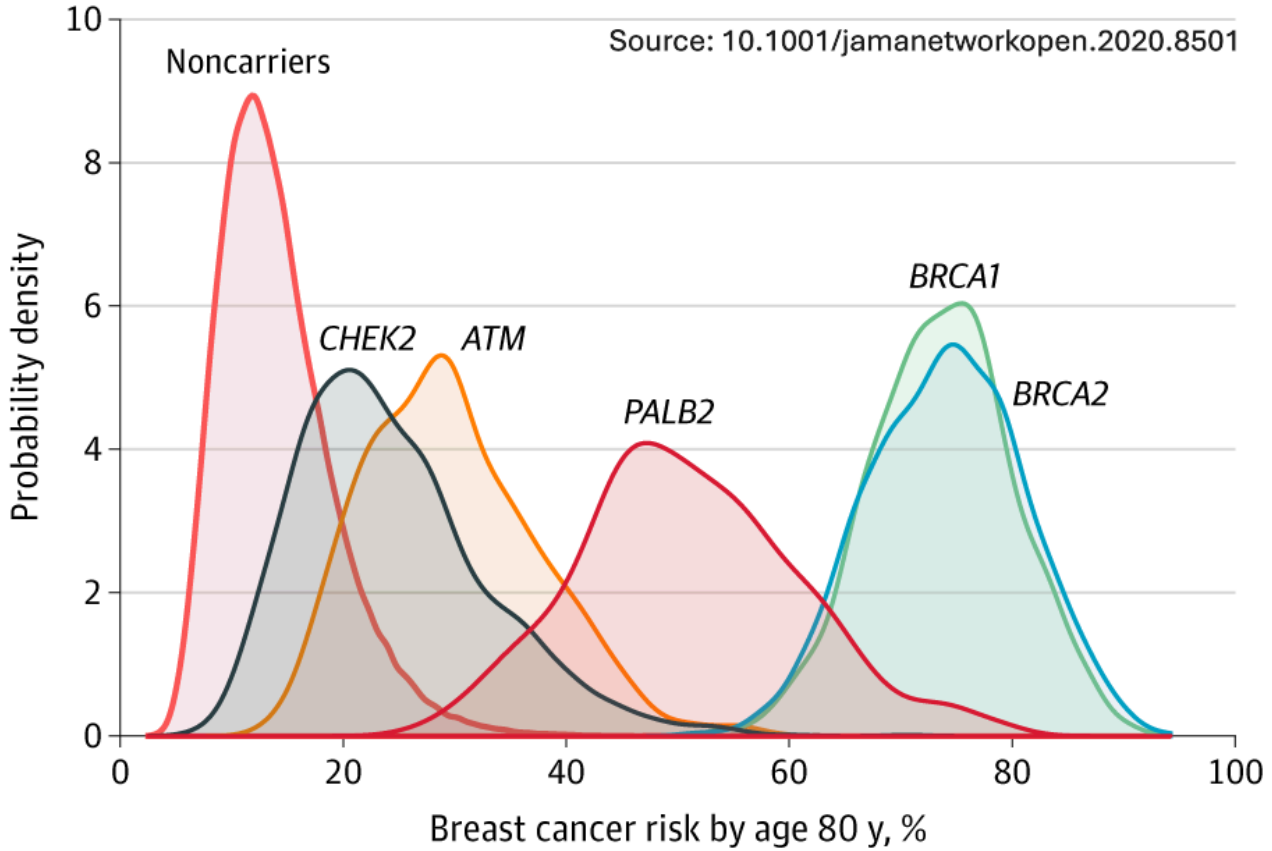
- many common genetic variants contribute to the risk of common diseases
- but, they usually have a very small effect on their own
- however, a (weighted) sum of small effects can result in a very large effect
- this is called a polygenic (risk) score (PGS or PRS)

Using PGS to modify risk assessment from traditional risk factors

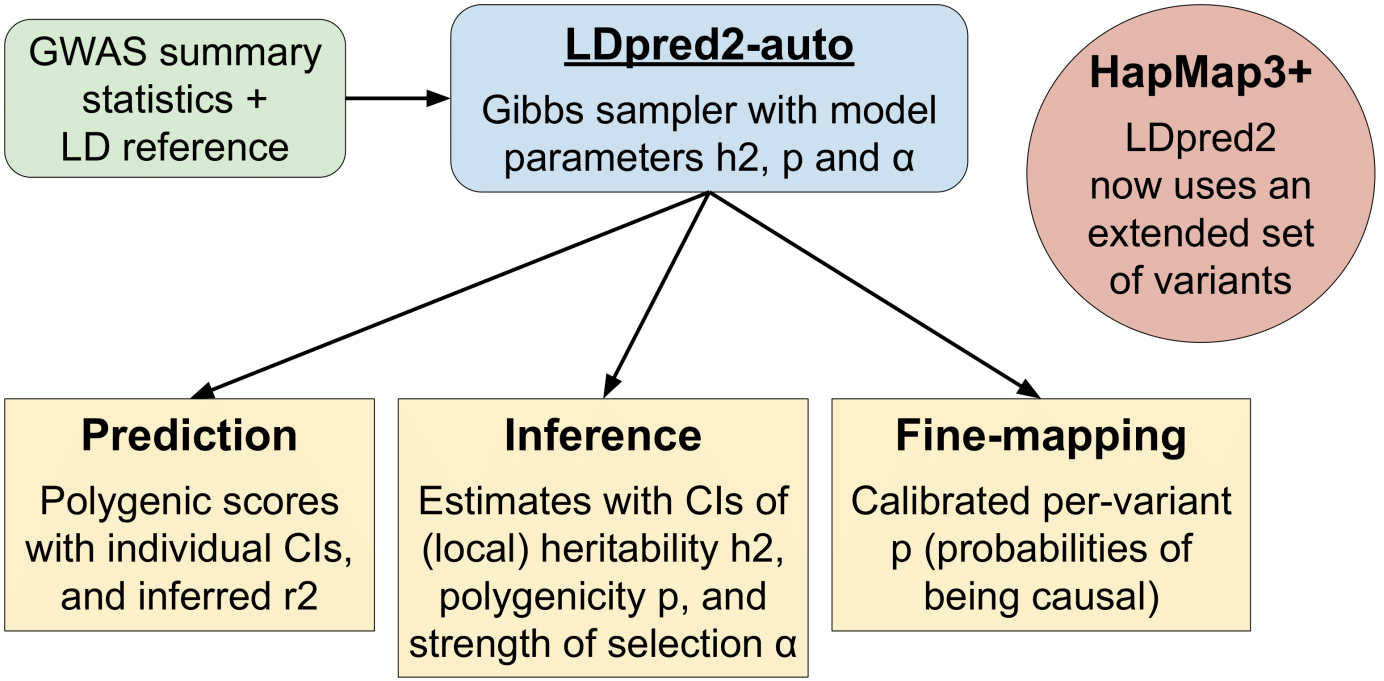


Source: 10.1038/s41440-024-01876-4

Modification of Lifetime Breast Cancer Risk for Pathogenic Variant Carriers and Noncarriers by an 86–Single-Nucleotide Variant Score



LDpred2-auto: a widely-used PGS (and inference) method

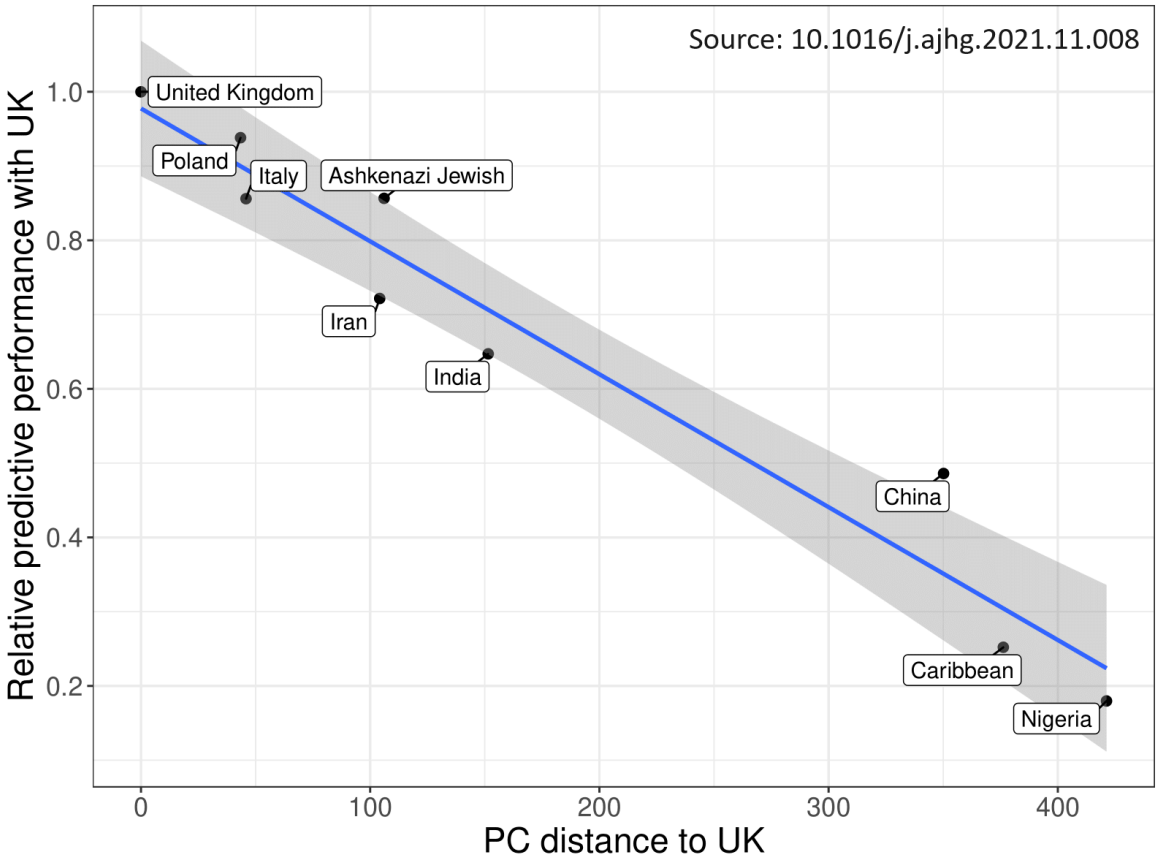


Privé, F., Arbel, J., & Vilhjálmsson, B.J. (2020). LDpred2: better, faster, stronger. *Bioinformatics*.

Privé, F., Albiñana, C., Arbel, J., Pasaniuc, B., & Vilhjálmsson, B.J. (2023). Inferring disease architecture and predictive ability with LDpred2-auto. *The American Journal of Human Genetics*.

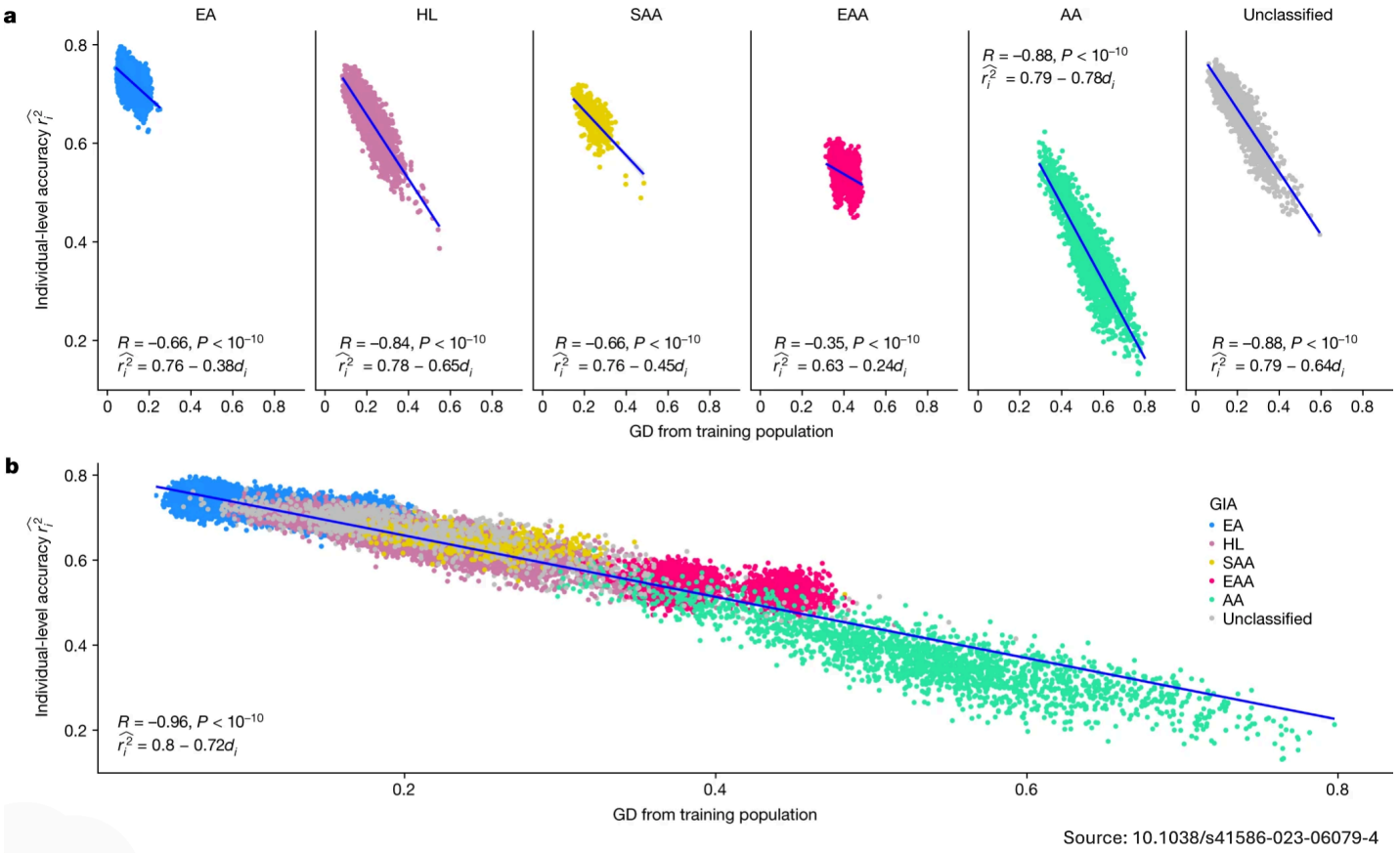
A major limitation of polygenic scores: their poor portability across ancestries

Predictive performance drops with genetic distance to training



Privé, F., Aschard, H., Carmi, S., Folkersen, L., Hoggart, C., O'Reilly, P.F., & Vilhjálmsson, B.J. (2022). Portability of 245 polygenic scores when derived from the UK Biobank and applied to 9 ancestry groups from the same cohort. *The American Journal of Human Genetics*.

Also at the individual level (instead of the group level)



Ding, Y., Hou, K., Xu, Z., Pimplaskar, A., Petter, E., Boulier, K., **Privé, F.**, Vilhjálmsson, B.J., Loohuis, L.O. and Pasaniuc, B. (2023). Polygenic scoring accuracy varies across the genetic ancestry continuum in all human populations. *Nature*.

How to explain this drop in performance?

It has been recently shown that

- causal variants are mostly similar across many populations
- their effect sizes are also very similar

How to explain this drop in performance?

It has been recently shown that

- causal variants are mostly similar across many populations
- their effect sizes are also very similar

The issue:

- in practice, we often don't use causal variants
- instead, we use tagging variants,
  highly correlated with the causal variants **in the training population**

How to explain this drop in performance?

It has been recently shown that

- causal variants are mostly similar across many populations
- their effect sizes are also very similar

The issue:

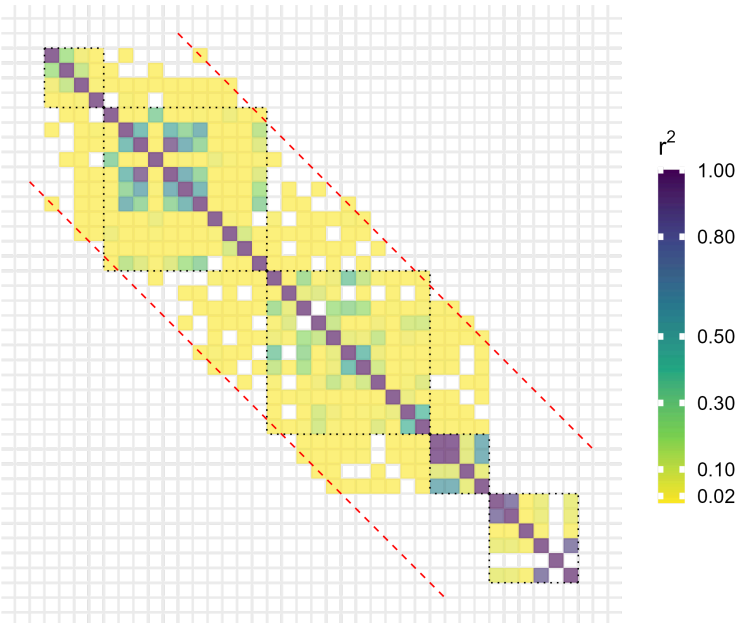
- in practice, we often don't use causal variants
- instead, we use tagging variants,
  highly correlated with the causal variants **in the training population**
- but correlations between tagging and causal variants **varies across populations**
- which reduces the predictive power of tagging variants **in other populations**.

The solution:

Precisely identifying causal variants

WP1: Using millions of genetic variants (problem)

- Causal variants need to be in the input data
  (around 10M common variants)
- The methods use the LD matrix (pairwise correlations between variants)
- This is a sparse matrix: banded or block-diagonal
- Takes 30 GB for 1M variants $\longrightarrow$ would take 3000 GB for 10M variants

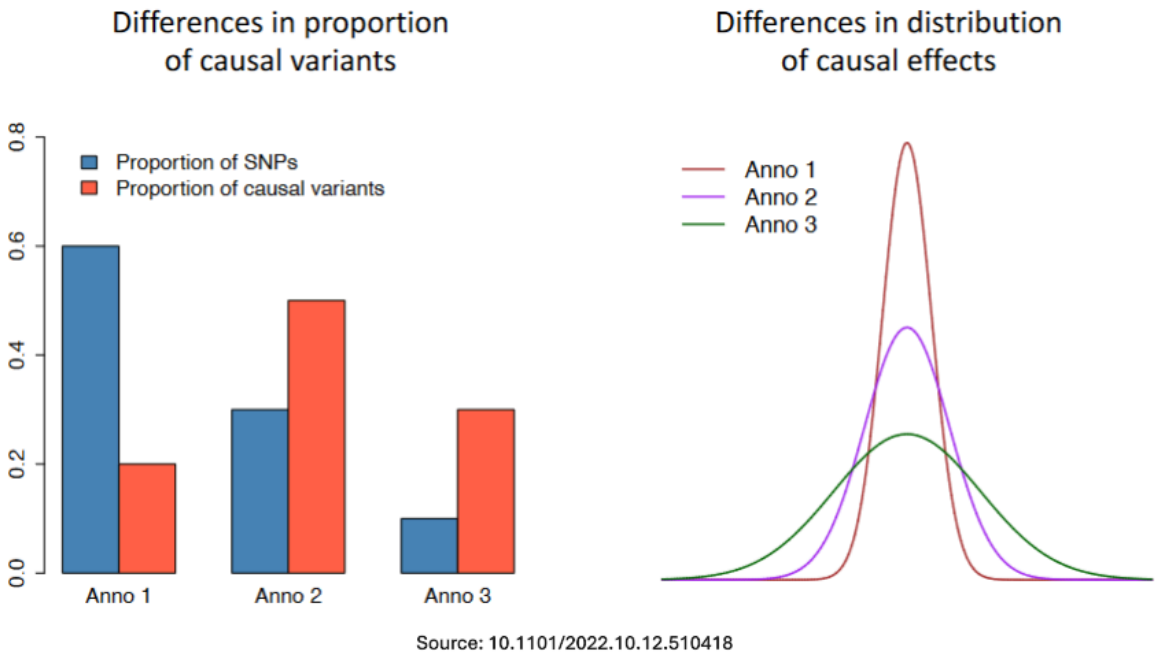


WP1: Using millions of genetic variants (possible solutions)

- compact sparse format (divide size by 2)
- storing correlations with two bytes only (divide size by 4)
- matrix seriation $\longrightarrow$ reordering variants to make blocks smaller
- eigendecomposition or compression of the matrix
- adapt methods to use very sparse *inverse* covariance matrices

WP2: Leveraging functional annotations

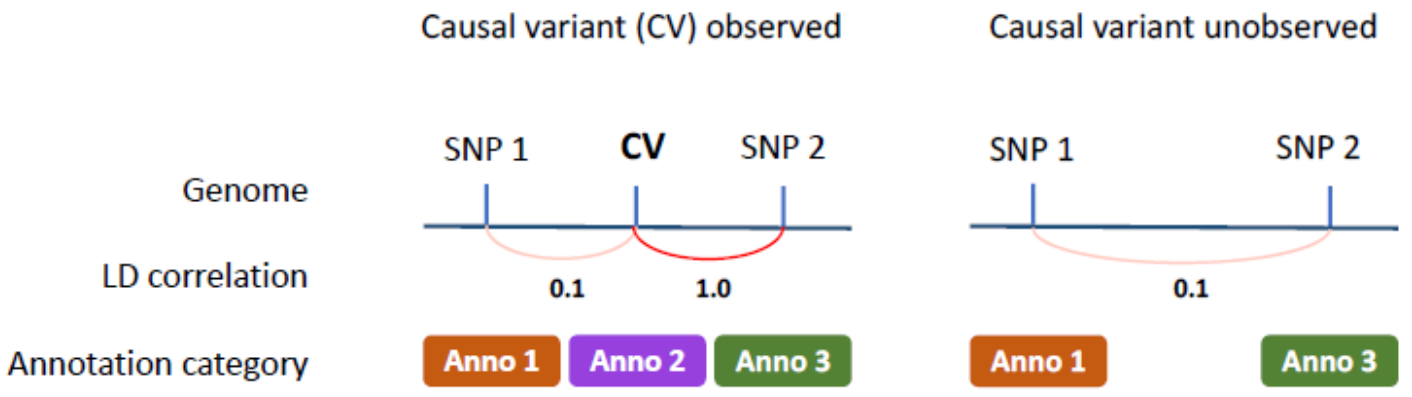
How can functional categories be useful?



Can be used to modify prior probabilities of being causal in LDpred2-auto.

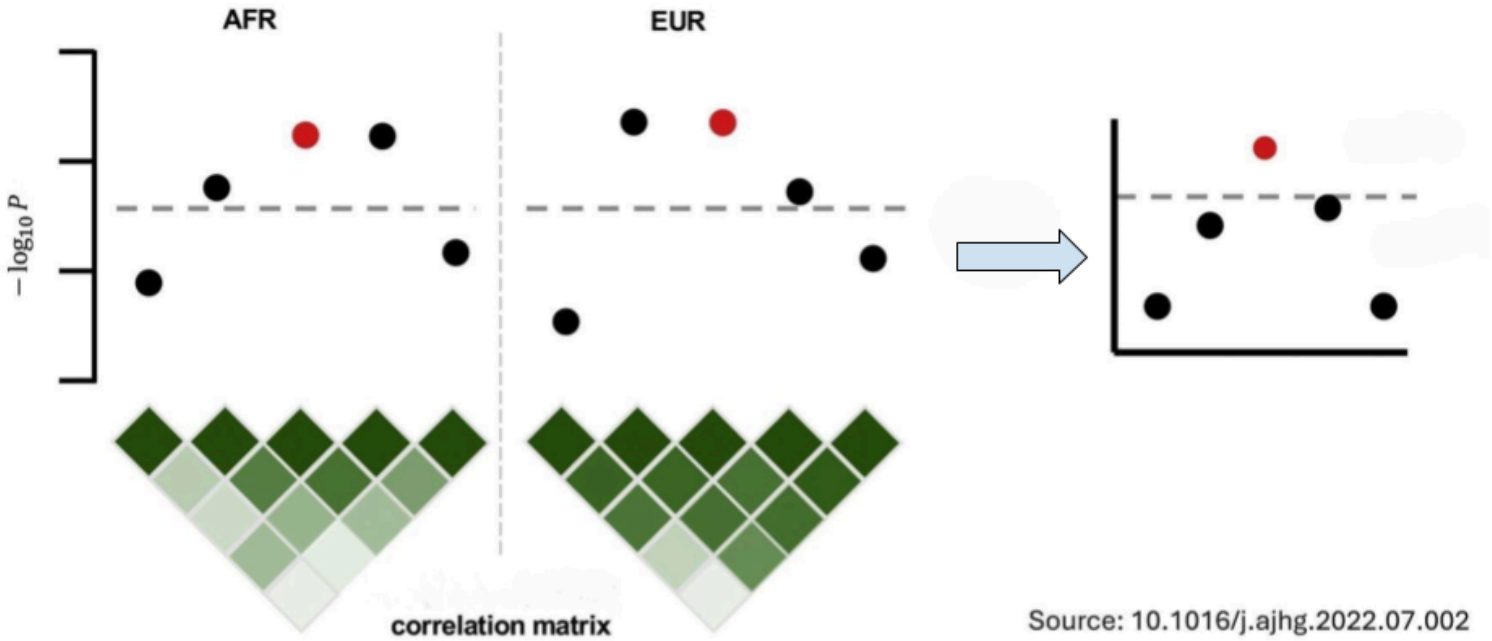
Functional annotations: the importance of using millions of variants

We must use many variants (WP1) to keep annotations useful



Source: 10.1101/2022.10.12.510418

WP3: Leveraging multi-ancestry data



Again, we need to make methods more scalable (WP1), to use multiple datasets here.

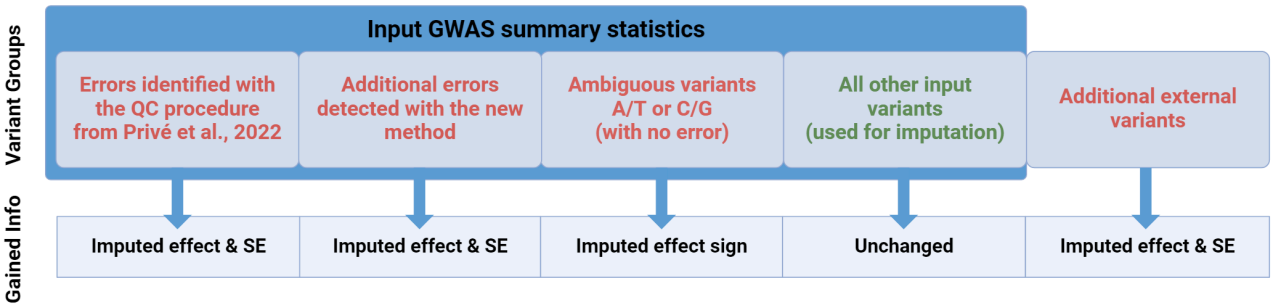
WP4: Ensuring the quality and coverage of the training data

- there are lots of problems with the input data (GWAS summary statistics)
- which can causes lots of misspecifications and biases in the methods

Privé, F., Arbel, J., Aschard, H., & Vilhjálmsson, B. J. (2022). Identifying and correcting for misspecifications in GWAS summary statistics and polygenic scores. *Human Genetics and Genomics Advances*.

WP4: Ensuring the quality and coverage of the training data

- there are lots of problems with the input data (GWAS summary statistics)
- which can causes lots of misspecifications and biases in the methods



- I propose to implement a QC and imputation method (synergistic)
- and to provide a set of highly refined GWAS summary statistics

Privé, F., Arbel, J., Aschard, H., & Vilhjálmsson, B. J. (2022). Identifying and correcting for misspecifications in GWAS summary statistics and polygenic scores. *Human Genetics and Genomics Advances*.

Outputs from this project

- Better polygenic scores for all populations
- Identification of causal variants to better understand disease etiology
- More robust genetic findings thanks to better QC and imputation

Outputs from this project

- Better polygenic scores for all populations
- Identification of causal variants to better understand disease etiology
- More robust genetic findings thanks to better QC and imputation

Why me?

- Background in Mathematics, Statistics, and Computer Science
- Already developed many state-of-the-art methods in past 8.5 years (recognized internationally)
- Published 9 first-author, 2 sole-author and 4 co-senior-author papers
- Already found many collaborators for these work packages

Thank you for your attention

privefl