—MAGNIFIC—

Maximizing Genetic Findings and Prediction

Florian Privé

CRCN INSERM–CSS6 application

Inserm

La science pour la santé
From science to health

About me

Professional background

- 2013–2016: Engineer in Computer Science & Applied Mathematics
- 2016–2019: PhD in Computational Biology (Grenoble)
- 2019–2021: Postdoc at Aarhus University (Denmark)
- 2022–2025: Senior Researcher (promotion at the same place)

Research focus

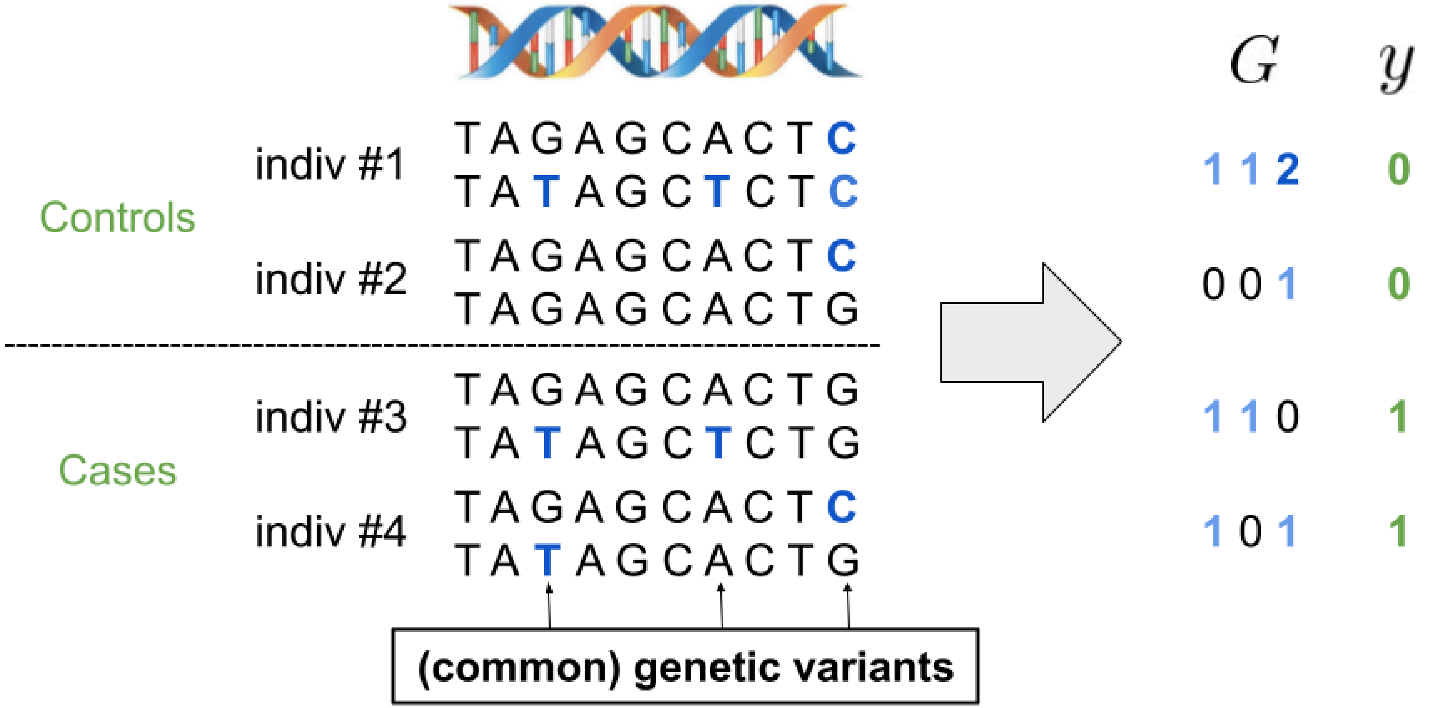
- Statistical human genetics
- Development of statistical methods and R/C++ packages
  for efficient and powerful analyses of large-scale genetic data
- Particularly for deriving **polygenic risk scores (PRS)**

Genetic data

Genome-Wide Association Studies (GWAS)

Polygenic Risk Scores (PRS)

Genetic variants and GWAS



			G	y
Controls	indiv #1	TAGAGCACTC TATAGCTCTC	1 1 2	0
	indiv #2	TAGAGCACTC TAGAGCACTG	0 0 1	0
Cases	indiv #3	TAGAGCACTG TATAGCTCTG	1 1 0	1
	indiv #4	TAGAGCACTC TATAGCACTG	1 0 1	1

(common) genetic variants

Genome-wide association study (GWAS):

association between each genetic variant and the case-control status

$$\text{logit}(\mathbb{P}(y = 1)) = \alpha + \beta_j G_j + \ldots + \epsilon$$

GWAS and polygenic risk scores (PRS)

Studying common diseases, such as heart diseases, cancers, diabetes

Thanks to GWAS, we know that

- many **common** genetic variants are causal $\left(\beta_j \neq 0\right)$
- but, they usually have a **small effect size** β_j on their own

  $\Rightarrow$ <u>a common causal variant is not useful as a risk factor</u>

GWAS and polygenic risk scores (PRS)

Studying common diseases, such as heart diseases, cancers, diabetes

Thanks to GWAS, we know that

- many **common** genetic variants are causal $\left(\beta_j \neq 0\right)$
- but, they usually have a **small effect size** β_j on their own

  $\Rightarrow$ a common causal variant is not useful as a risk factor

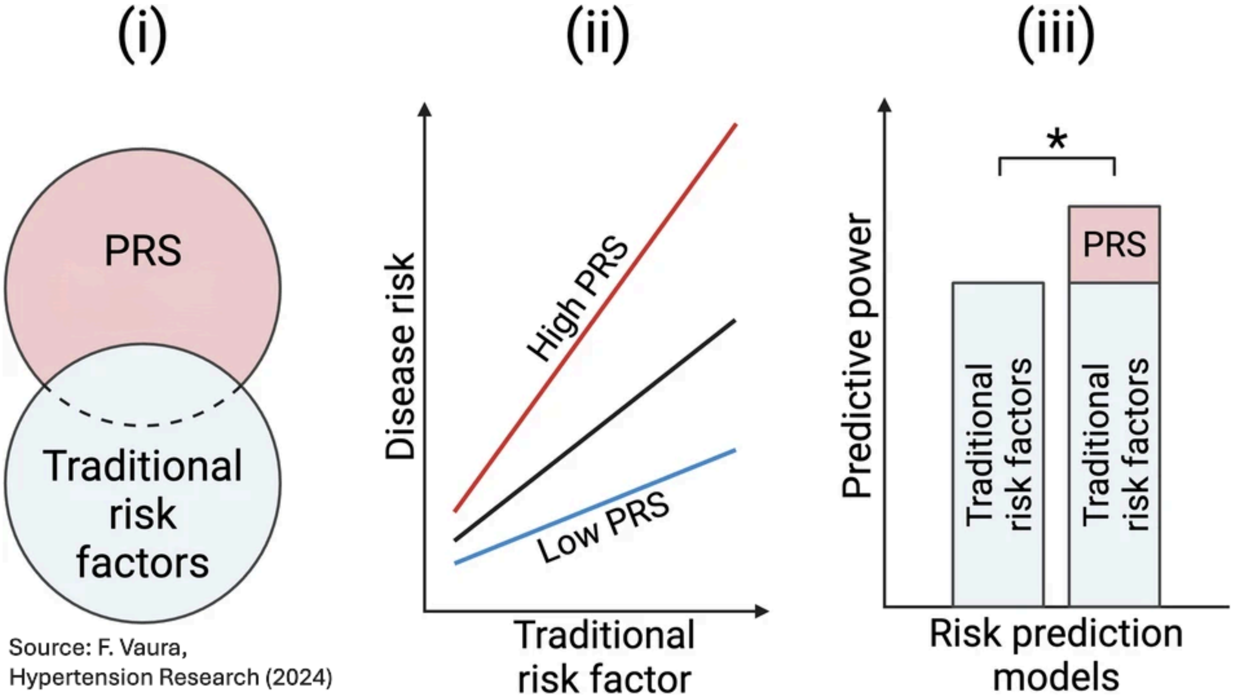
From GWAS data to **polygenic risk scores (PRS)**:

- variants can be aggregated in a joint predictive model: $PRS = \sum_j \hat{\gamma}_j \, G_j$
- by aggregating many small effects, the PRS can have a large effect

  $\Rightarrow$ the PRS can be useful as a risk factor

Public Health: refining risk assessment from traditional risk factors

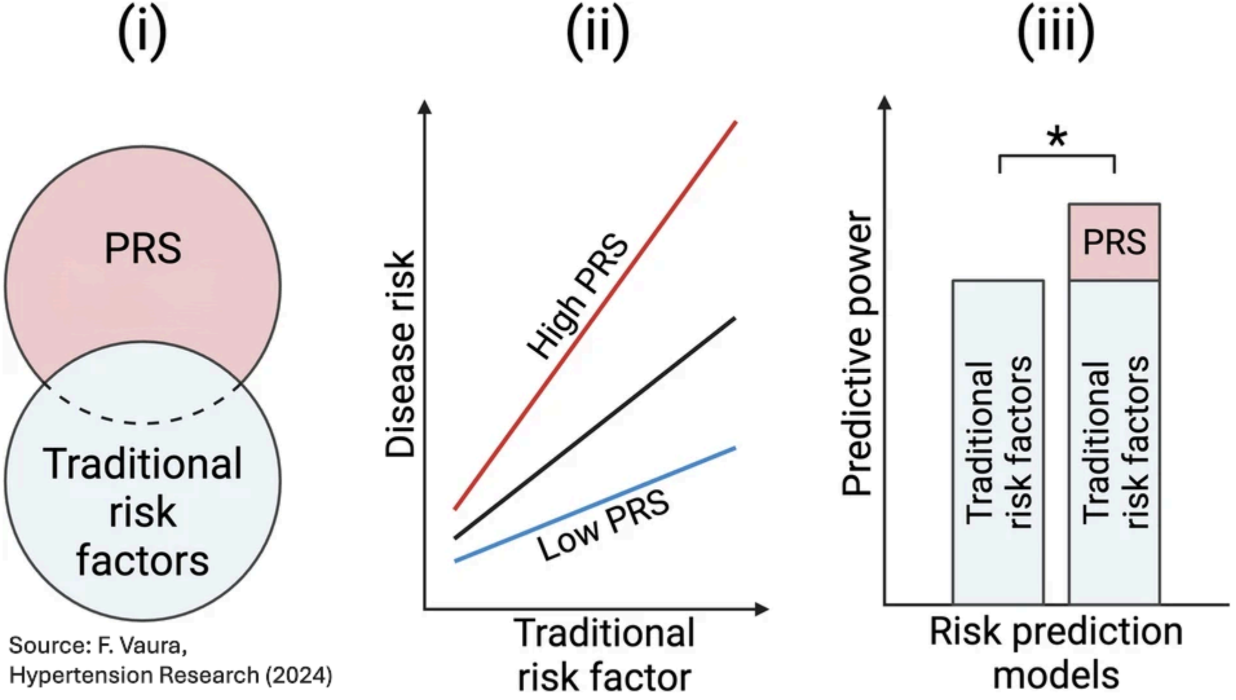
Traditional risk factors: age, smoking, pollution, low SES, diet, physical inactivity, family history, (low-frequency large-effect) genetic mutations, etc



Source: F. Vaura, Hypertension Research (2024)

Public Health: refining risk assessment from traditional risk factors

Traditional risk factors: age, smoking, pollution, low SES, diet, physical inactivity, family history, (low-frequency large-effect) genetic mutations, etc



Source: F. Vaura, Hypertension Research (2024)

PRS clinical utility in a **clinical trial**: A. Fuat et al, Eur. J. of Preventive Cardiology (2024)

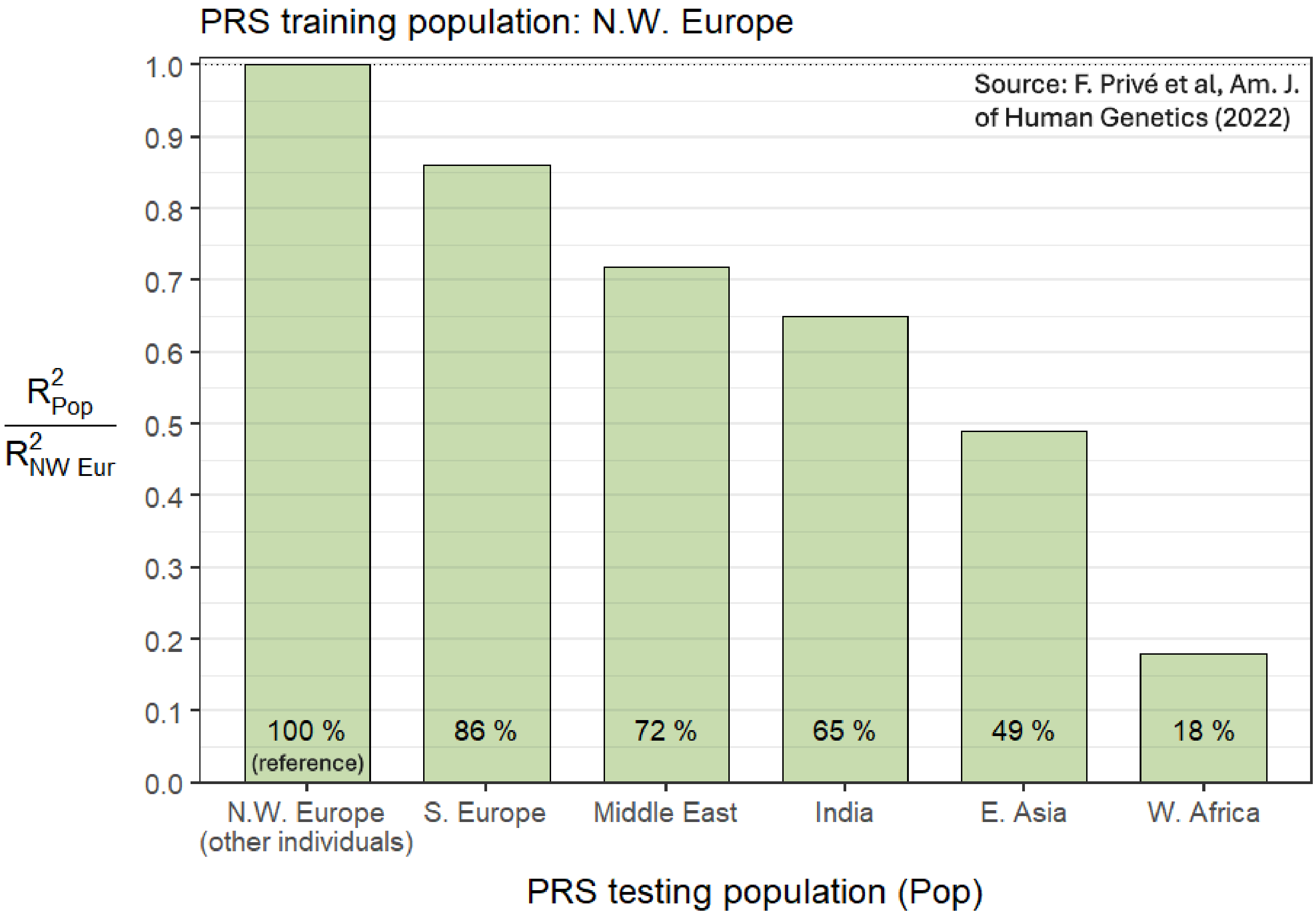
Refining breast cancer genetic risk using a PRS **in France**: Y. Jiao et al, Eur. J. of Cancer (2023)

A major limitation of PRS:

their poor portability across populations

risks exacerbating health disparities

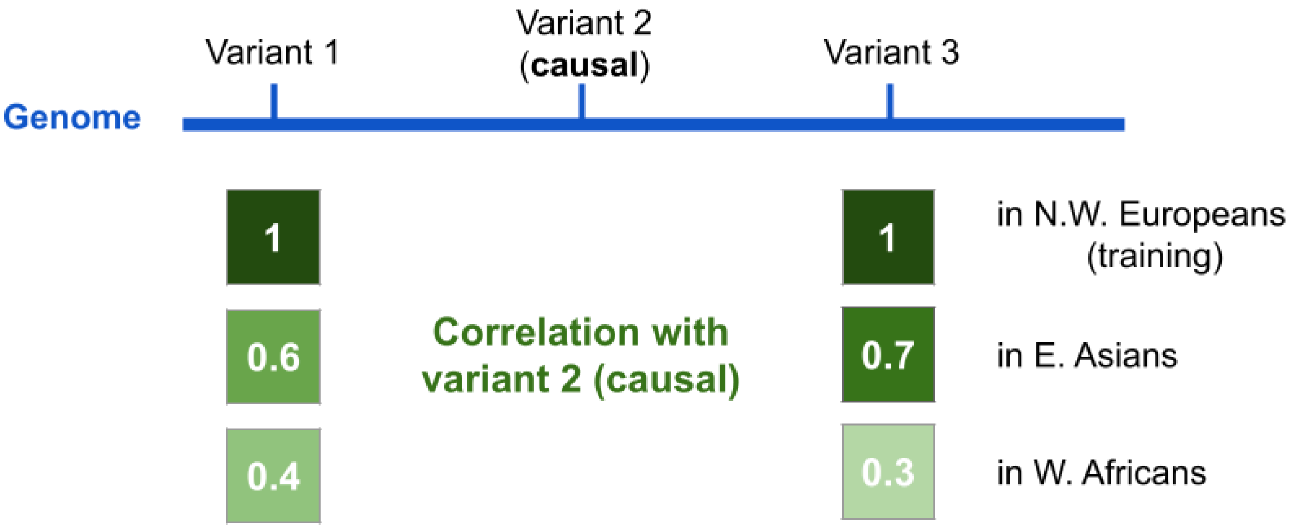
PRS performance drops with distance from training population



Source: F. Privé et al, Am. J. of Human Genetics (2022)

Explanation: we often don't use causal variants in practice

A causal variant largely shares its effect size across populations
But…



Correlation with variant 2 (causal)

Variant 1	Variant 3	
1	1	in N.W. Europeans (training)
0.6	0.7	in E. Asians
0.4	0.3	in W. Africans

The correlated variants are <u>as predictive</u> as the causal variant
<u>in the training population only</u>

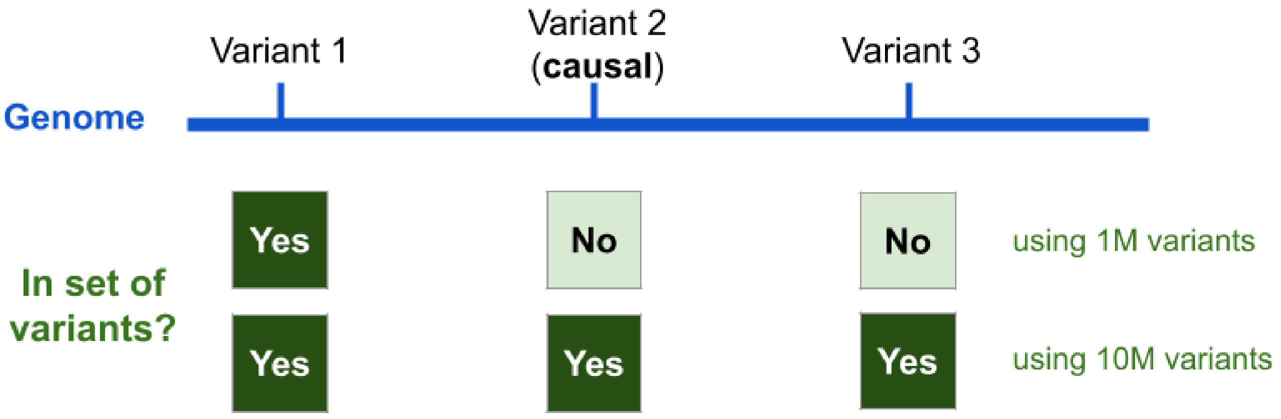
The solution $\Rightarrow$ my proposed project:

identifying causal variants and using them

in polygenic risk scores (PRS)

Scaling methods to using 10M genetic variants (WP1)

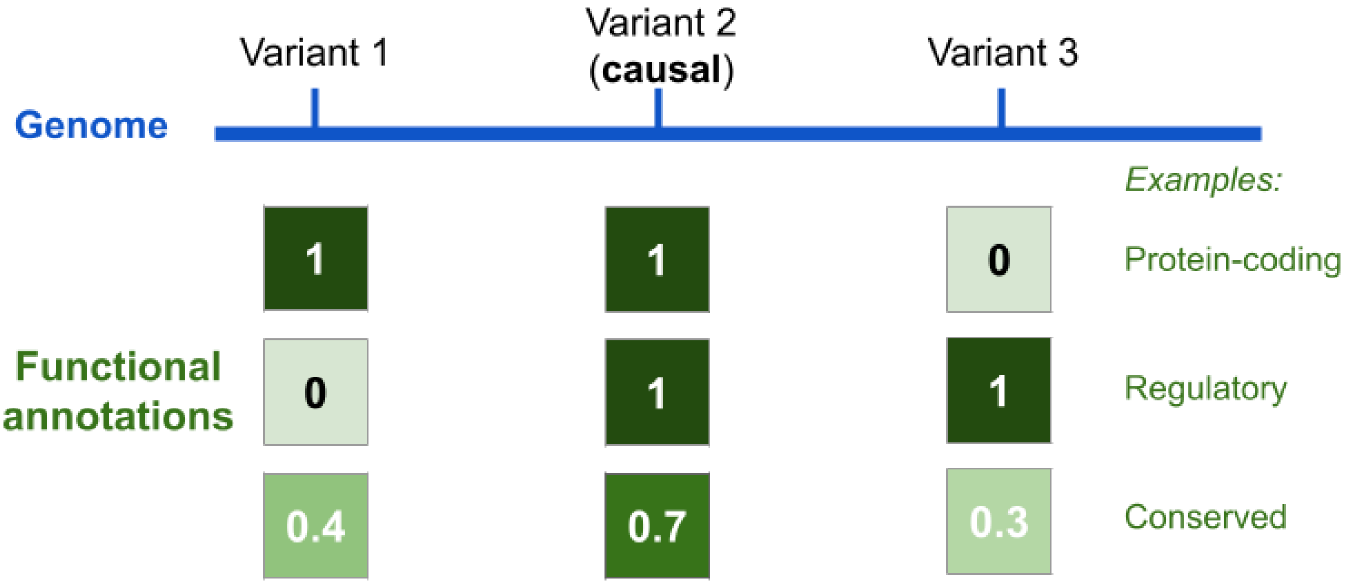
- there are ~10M common variants
- but most PRS methods (including mine) use ~1M variants (mostly for computational reasons and due to redundancy)



We need to use 10M to make sure most causal variants are present

💻 I will optimize both methods and data structures to use 10M variants 💻

Prioritizing causal variants thanks to functional annotations (WP2)

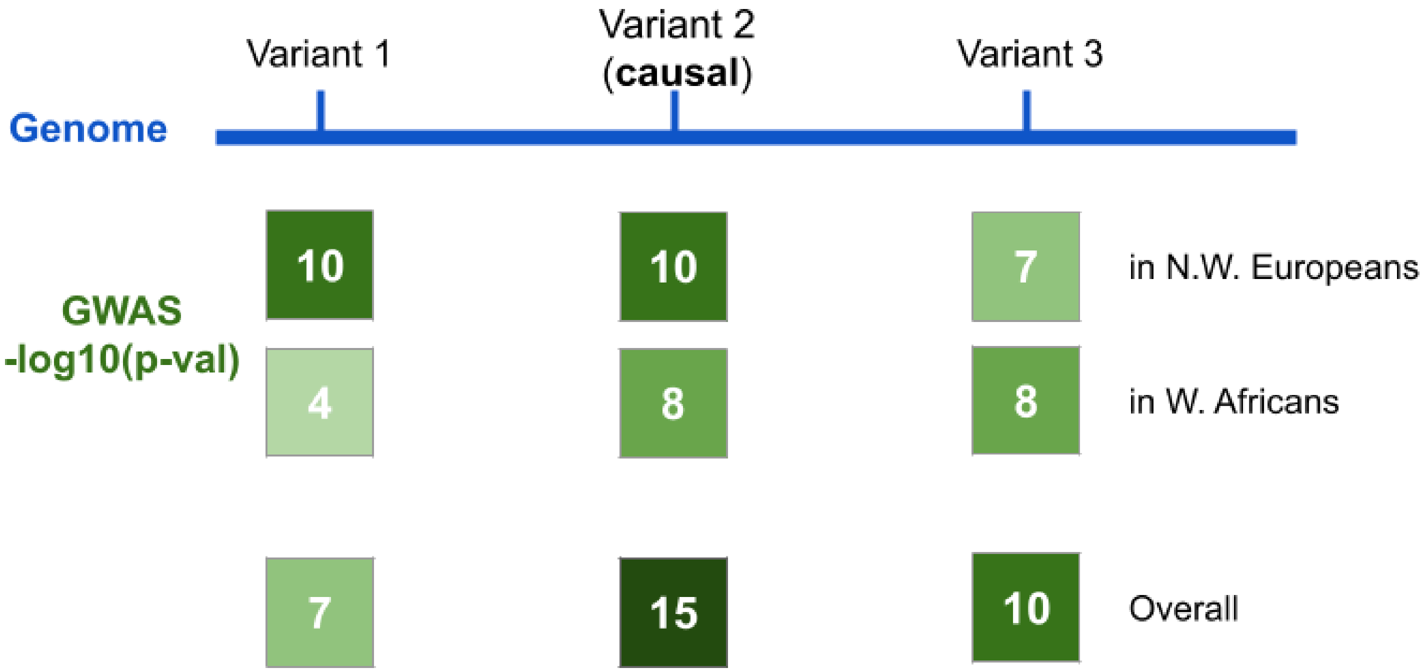


	Variant 1	Variant 2 (**causal**)	Variant 3	*Examples:*
Functional annotations	1	1	0	Protein-coding
	0	1	1	Regulatory
	0.4	0.7	0.3	Conserved

Variants in some functional categories are more likely to be causal

💻 I will integrate this information into my Bayesian PRS methodology 💻

Prioritizing causal variants thanks to multi-ancestry data (WP3)



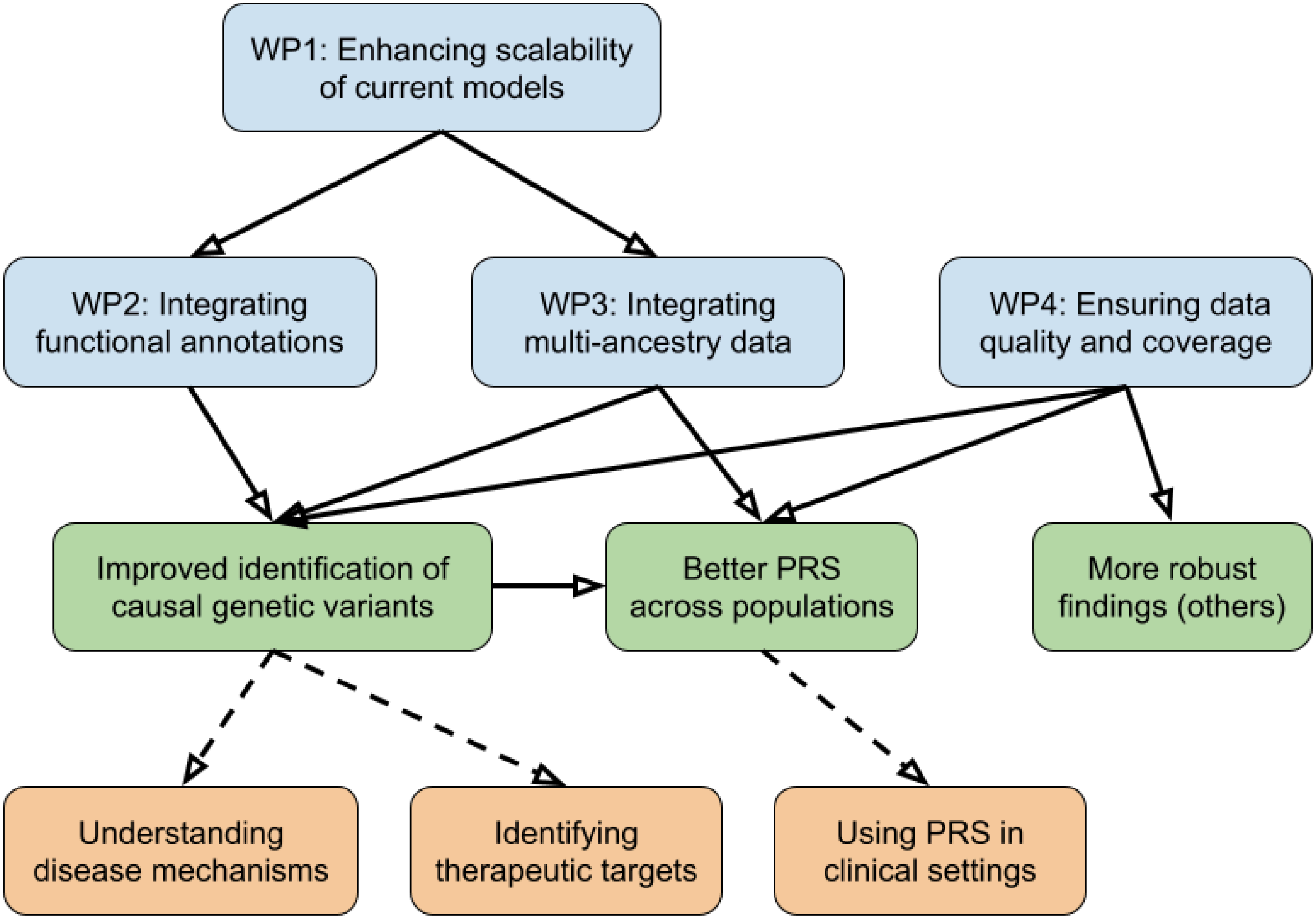
Genome

GWAS -log10(p-val)

Variant 1	Variant 2 (causal)	Variant 3	
10	10	7	in N.W. Europeans
4	8	8	in W. Africans
7	15	10	Overall

💻 I will integrate multi-ancestry data into my Bayesian PRS methodology 💻

MAGNIFIC: Maximizing Genetic Findings and Prediction


WP1: Enhancing scalability of current models

WP2: Integrating functional annotations

WP3: Integrating multi-ancestry data

WP4: Ensuring data quality and coverage

Improved identification of causal genetic variants

Better PRS across populations

More robust findings (others)

Understanding disease mechanisms

Identifying therapeutic targets

Using PRS in clinical settings

Feasability

- **I have developed many efficient & powerful methods** in past 9 years
  - LDpred2, widely used for constructing PRS + often ranked best
  - bigstatsr and bigsnpr, R(cpp) packages for large-scale analyses
- I have published 28 papers with 2800 citations in total, including **2000 for my 11 first-author papers**
- My **funding strategy** to recruit people:
  - ANR JCJC
  - ATIP-Avenir
- I have **co-supervised several young researchers**
  - two PhD students who graduated (**co-last author on 4 papers**)
  - ongoing: two PhD students, one research assistant, one postdoc
- I have found **several collaborators** for these work packages (Broad, UCLA, Oxford, Helsinki, Pasteur, INRIA, etc)

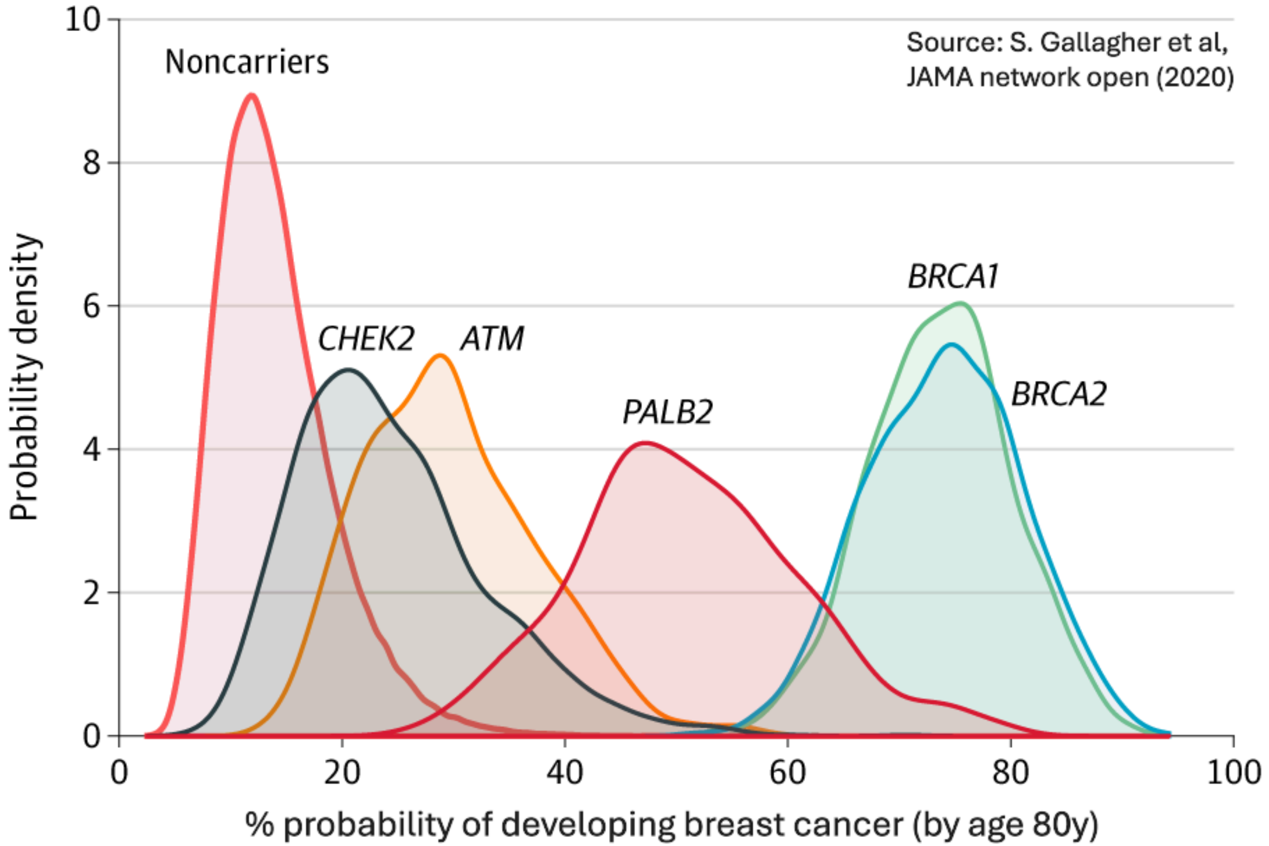
Integration into INSERM U1220 in Toulouse

- **Host Team**: "GenoFun: Functional impact of Genomic variations on disease", a Bioinformatics team at IRSD, INSERM U1220
- **Collaboration**:
    - **Sarah Djebali (CR INSERM)**: Expert in functional genome annotation, supporting integration of annotations
    - **Jean Monlong (CR INSERM)**: Specialist in pangenomes and structural variants, expanding from simply using single-nucleotide polymorphisms (SNPs)
    - **Other lab members**: validation of causal variants using experimental models (e.g., mice, organoids)
- Technical support and computational resources via **Genotoul compute cluster** (5000 cores, 83 TB RAM, 7.5 PB storage)
- **Collaborative Environment**: Toulouse bioinformatics, biostatistics, mathematics and informatics network (INRAE, CNRS, INSERM, Uni)

Thank you for your attention

Florian Privé

Refining breast cancer genetic risk using a 86-variant PRS



Source: S. Gallagher et al, JAMA network open (2020)

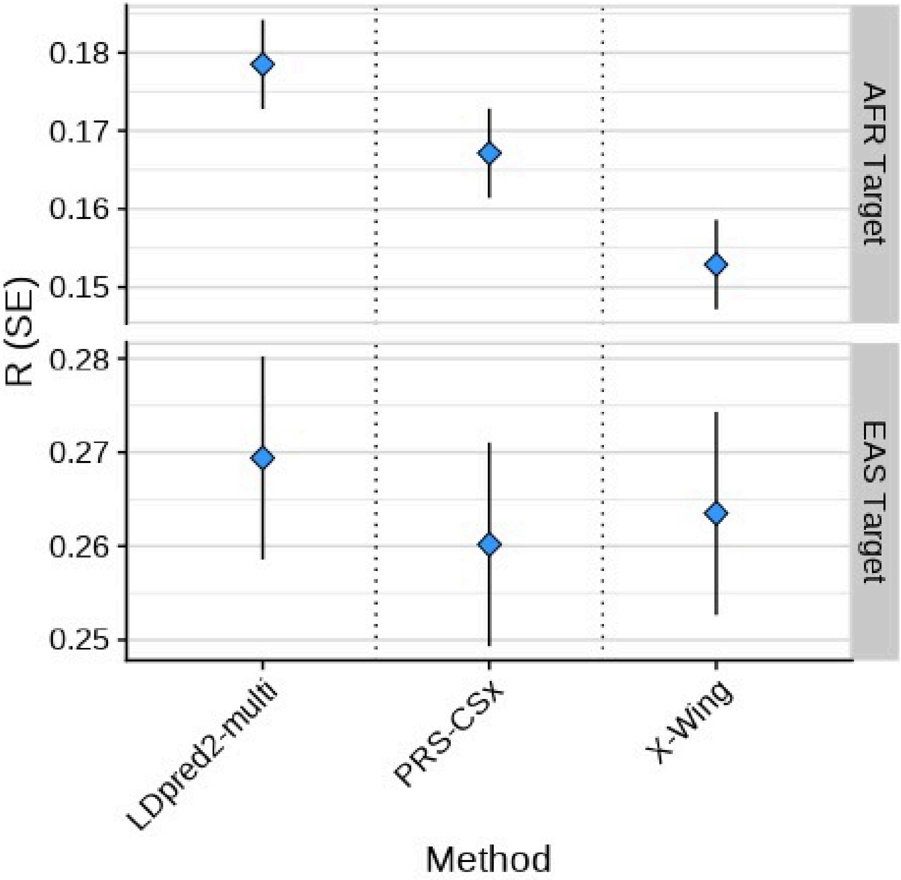
WP1: Using millions of genetic variants (possible solutions)

The main bottleneck is storing and using the matrix of correlations between variants.

Possible solutions:

- quantization: storing correlations with two bytes only (divide size by 4)
- compression on top of quantization
- matrix seriation → reordering variants to make blocks smaller
- eigendecomposition
- adapt methods to use very sparse *inverse* covariance matrices

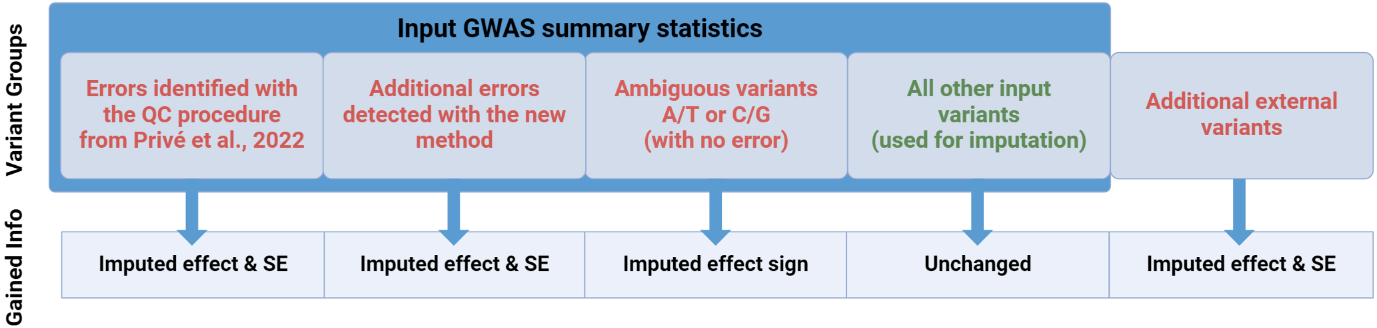
LDpred2 vs some state-of-the-art multi-ancestry PRS methods



O. Pain (2025). Leveraging Global Genetics Resources to Enhance Polygenic Prediction Across Ancestrally Diverse Populations. *medRxiv*

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


Variant Groups
Gained Info
Input GWAS summary statistics
Errors identified with the QC procedure from Privé et al., 2022
Additional errors detected with the new method
Ambiguous variants A/T or C/G (with no error)
All other input variants (used for imputation)
Additional external variants
Imputed effect & SE
Imputed effect & SE
Imputed effect sign
Unchanged
Imputed effect & SE

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

F. Privé et al (2022). Identifying and correcting for misspecifications in GWAS summary statistics and polygenic scores. *Human Genetics and Genomics Advances*.

Scientific animation

- 10 oral presentations (+ 2 planned) at international scientific conferences, including 1 invited
- invited to 16 seminar or lecture presentations
- reviewed a total of 61 different manuscripts, for 30 different journals
- external reviewer for Amsterdam UMC Fellowship 2022
- member of the Scientific Committee of EMGM Brest 2025
- reviewer for the Scientific Program Committee of ESHG 2025