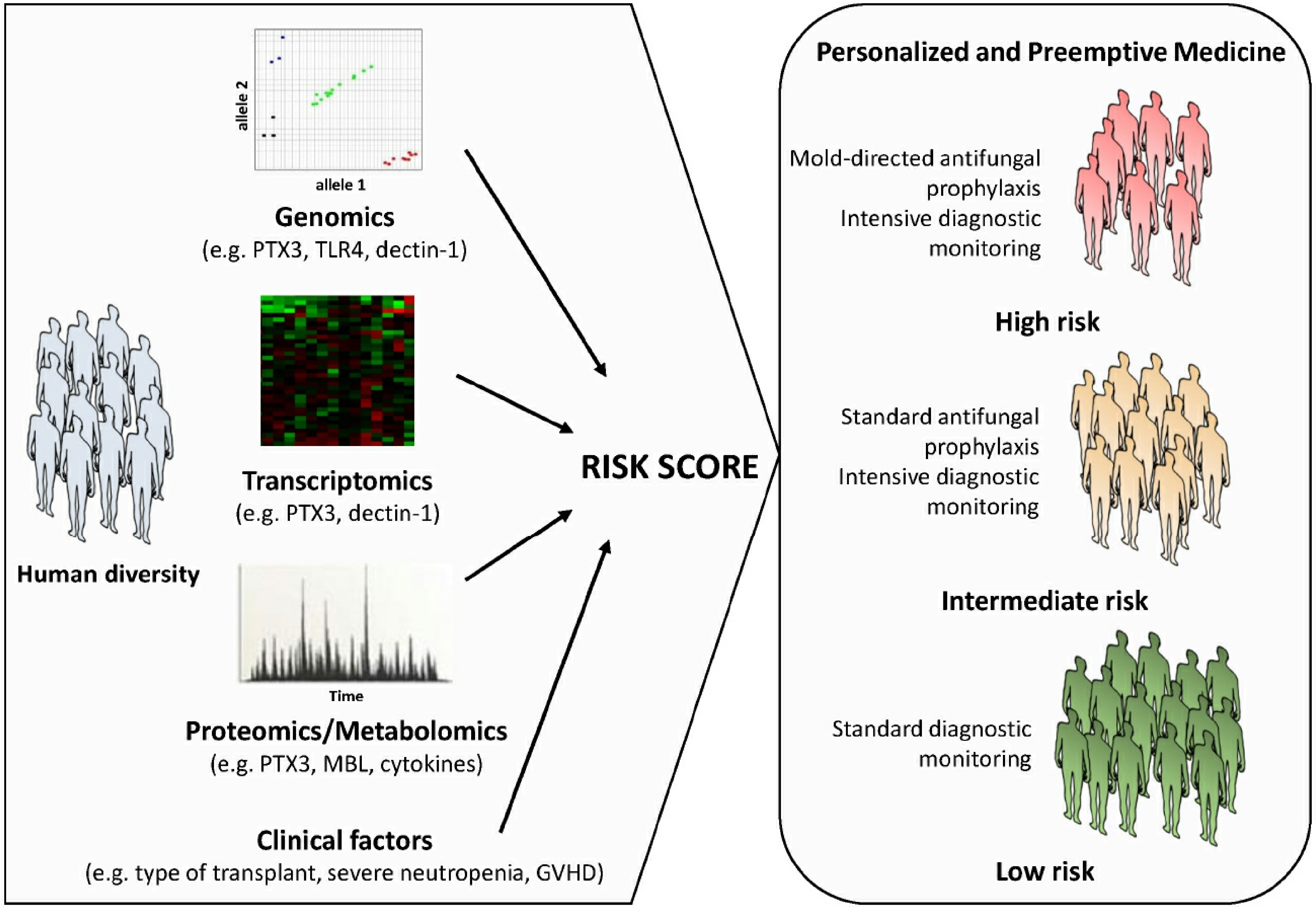
Predicting complex diseases: performance and robustness

Florian Privé (UGA)

supervised by
M. Blum (UGA) & H. Aschard (Institut Pasteur)

January 12, 2018

Personalized genetic medicine:


allele 2
allele 1
Genomics
(e.g. PTX3, TLR4, dectin-1)
Human diversity
Transcriptomics
(e.g. PTX3, dectin-1)
Time
Proteomics/Metabolomics
(e.g. PTX3, MBL, cytokines)
Clinical factors
(e.g. type of transplant, severe neutropenia, GVHD)
RISK SCORE
Personalized and Preemptive Medicine
Mold-directed antifungal prophylaxis
Intensive diagnostic monitoring
High risk
Standard antifungal prophylaxis
Intensive diagnostic monitoring
Intermediate risk
Standard diagnostic monitoring
Low risk

Source: Oliveira-Coelho, Ana, et al. "Paving the way for predictive diagnostics and personalized treatment of invasive aspergillosis." Frontiers in microbiology 6 (2015).

Data

0, 1 or 2 mutations per individual and per locus (position on the genome):

```
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11]
 [1,]    2    2    0    2    1    1    2    0    1     0     0
 [2,]    1    2    1    2    2    1    2    0    0     1     2
 [3,]    2    1    1    2    0    1    2    0    0     2     0
 [4,]    2    2    0    2    0    1    2    0    2     1     1
 [5,]    1    2    2    2    0    0    1    2    1     1     2
 [6,]    2    1    2    2    0    1    2    2    2     1     1
 [7,]    2    1    1    2    0    0    1    1    1     0     2
 [8,]    1    2    1    1    1    1    2    1    0     1     1
 [9,]    1    2    1    2    0    1    1    2    0     1     1
[10,]    2    2    0    2    1    1    1    0    2     1     0
[11,]    2    2    1    2    0    0    2    1    2     0     0
[12,]    1    0    2    2    2    1    2    2    1     0     2
```

Size of current datasets (UK Biobank): 500,000 individuals and (at least) 800,000 loci.

Goal

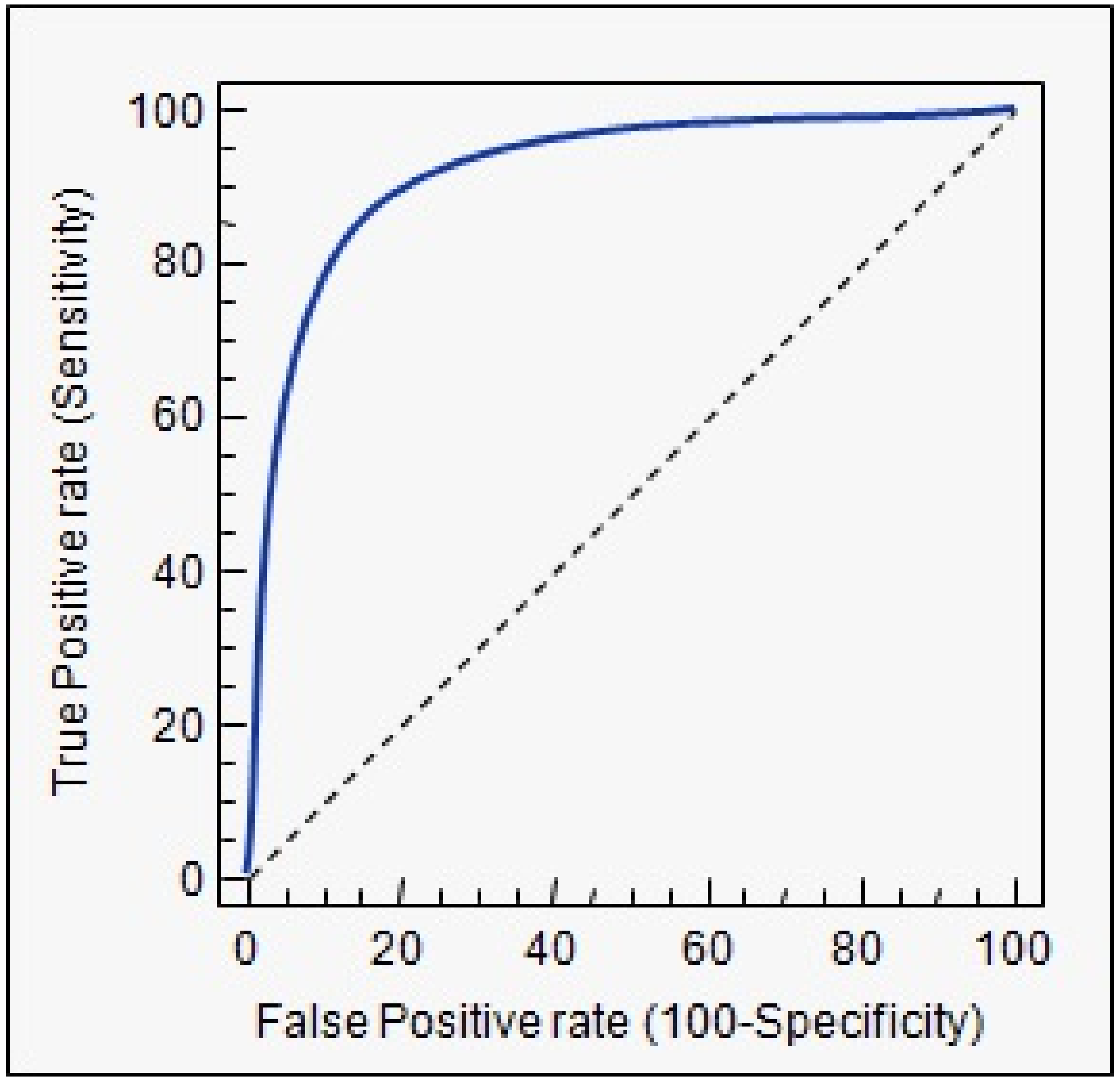
$$\boxed{\text{Disease} \sim \text{DNA mutations} + \cdots}$$

Methods

- statistical learning methods
- clever implementations for handling large datasets
- two R packages (Privé et al., 2017)

Assessing predictive performance

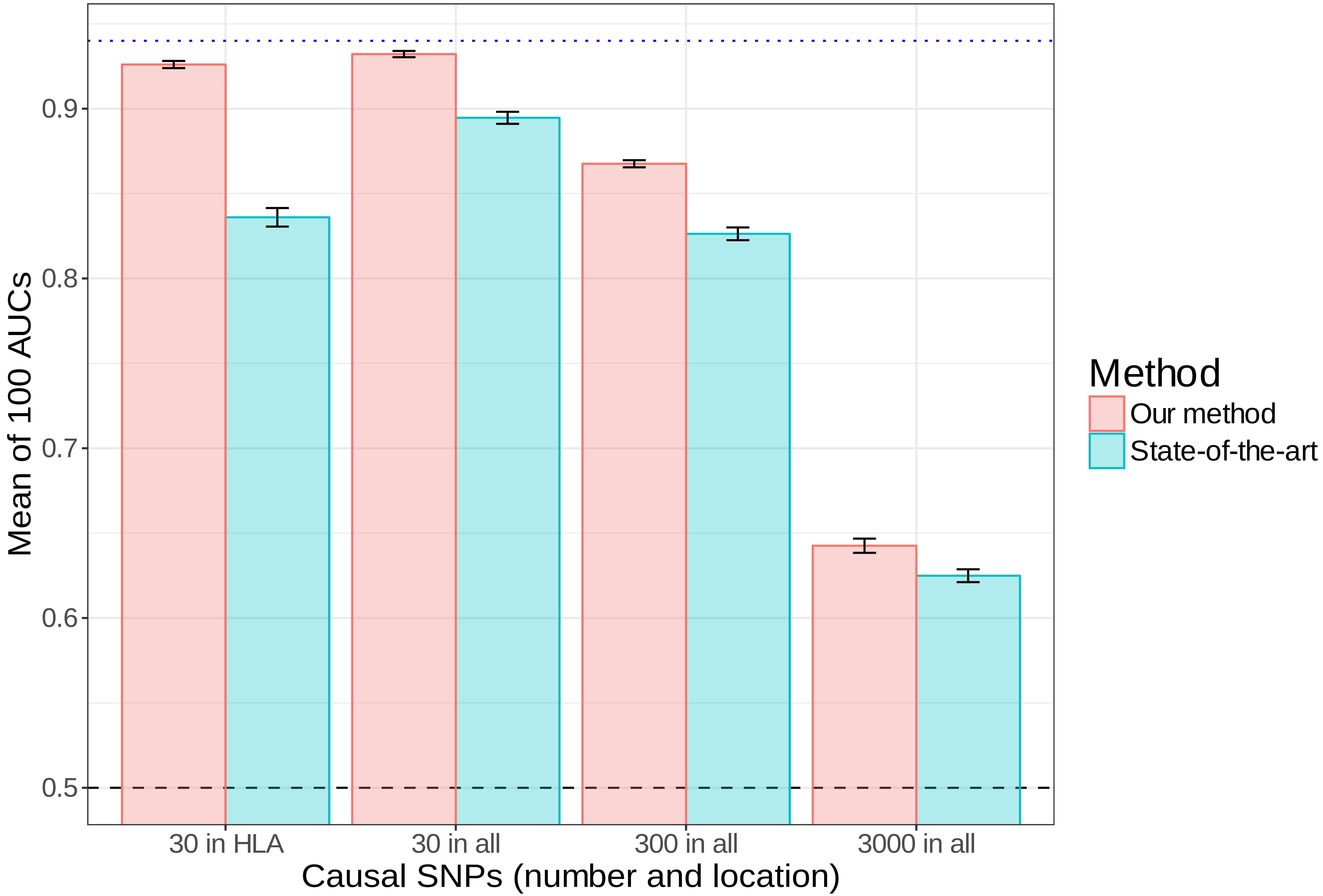
AUC (Area Under the ROC Curve) is often used.



Example of ROC curve.

$$\mathrm{AUC} = P(S_{\text{case}} > S_{\text{control}})$$

Results (simulating different disease architectures)



Thanks!

Twitter and GitHub: @privefl

Slides created via the R package **xaringan**.