Discordant variant analysis

Daniel Coral

GAME unit - LUDC

May 26, 2021

BMI → T2D is heterogeneous

- Cardiometabolic risk varies greatly within the same BMI level.
  - 'Metabolically healthy obesity'
  - 'Favorable adiposity'
  - What factors link BMI gain to metabolic risk?
    - Better stratification
    - New therapeutic targets

How to define this phenotype?

- Obesity with and without disease (case-control)
- Follow-up of individuals with obesity
- Genetics:
  - Multiple loci associated with BMI
  - Availability of associations of each loci across the phenome (PheWAS)
  - Causal inference

Analytical framework

- Starting point: BMI
- Cardiometabolic risk: T2D
  - Strongly associated with BMI
  - Life-threatening condition
- Identify loci highly associated with both conditions
  - Concordant ($\uparrow$ BMI and $\uparrow$ T2D)
  - Discordant ($\uparrow$ BMI and $\downarrow$ T2D).

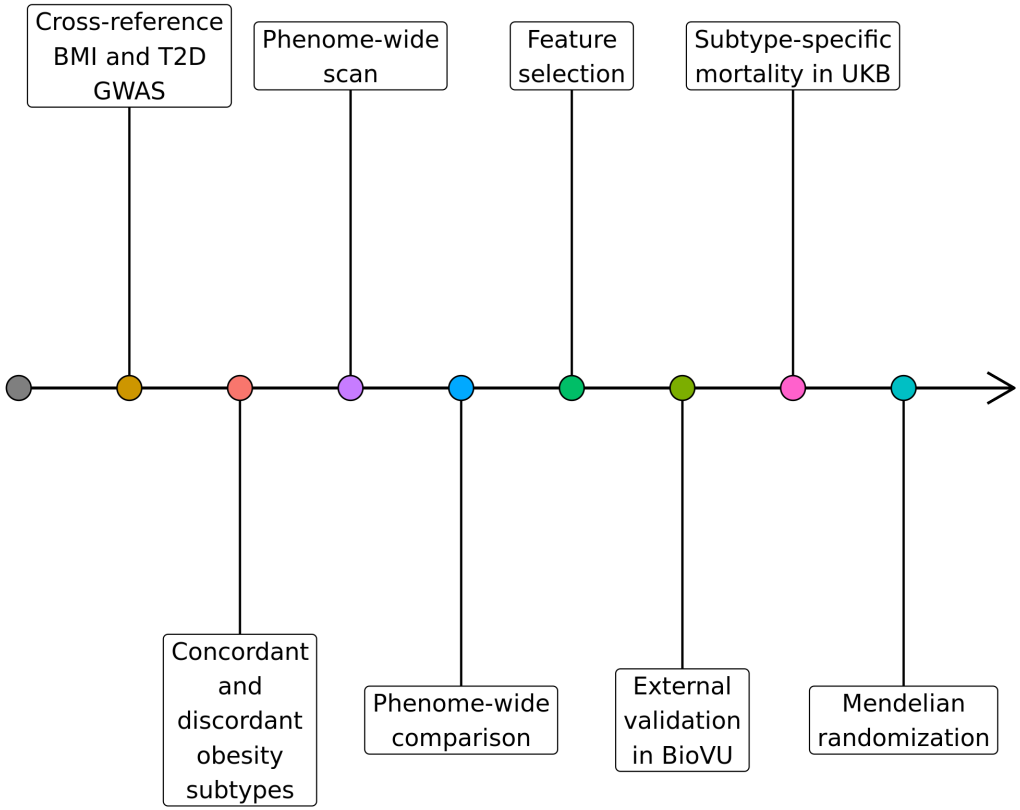
Pseudo case-control setup

- Phenome-wide scan

SNP	Discordant	Trait_1	Trait_2	...	Trait_n
SNP1	Yes				
SNP2	No				
SNP3	Yes				
...	...				
SNPn	No				

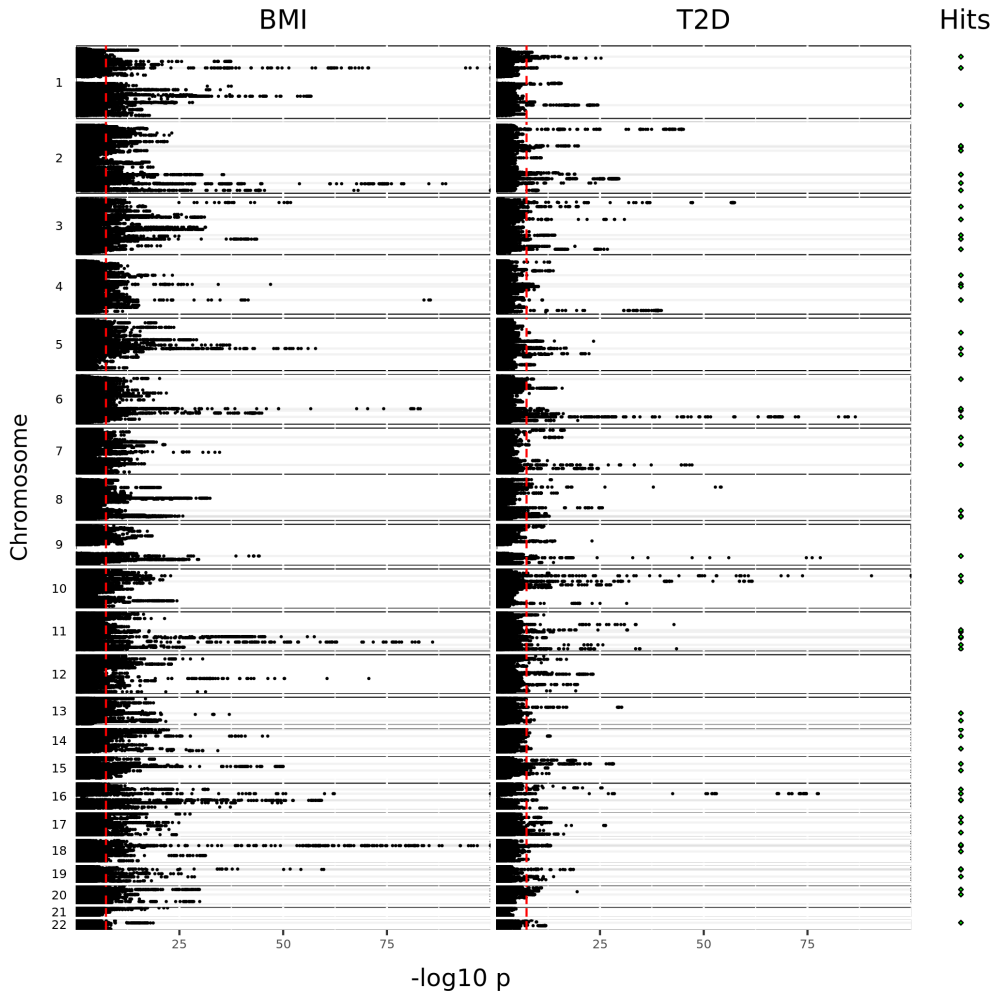
Using these data: $P(Discordant|Traits)$

Analysis pipeline



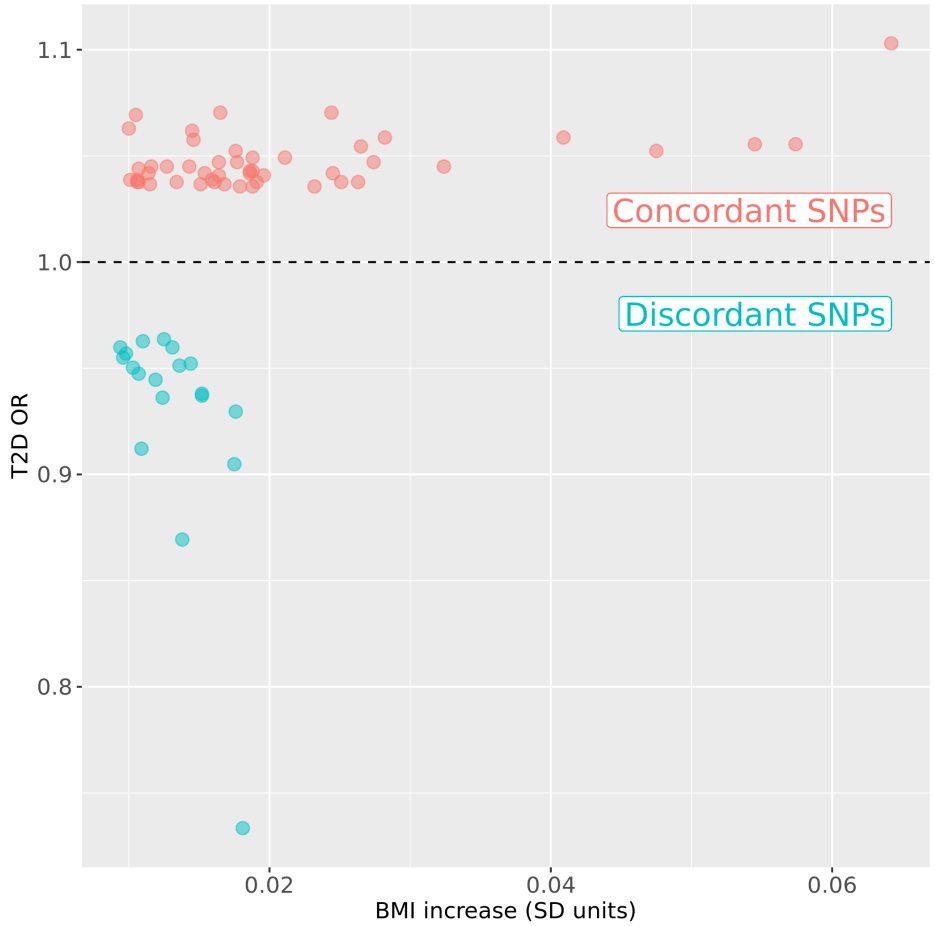
Cross-referencing BMI and T2D GWAS

BMI from Yengo *et al.* (2018) and T2D from Mahajan *et al* (2018).



Assembly of concordant (n = 48) and discordant (n = 19) profiles

Lead variants aligned to the BMI increasing allele and stratified by their β coefficient for T2D.



Phenome-wide scan - Data collection

- MRC IEU GWAS database
  - Associations of lead SNPs or nearest proxy if missing
    - $r^2 < 0.01$ over 500kb window in 1000G EUR
  - Studies in EUR
  - More than 500 individuals
  - Binary traits: more than 25 minor alleles in smallest group
- $\sim$ 3500 traits

Phenome-wide comparison

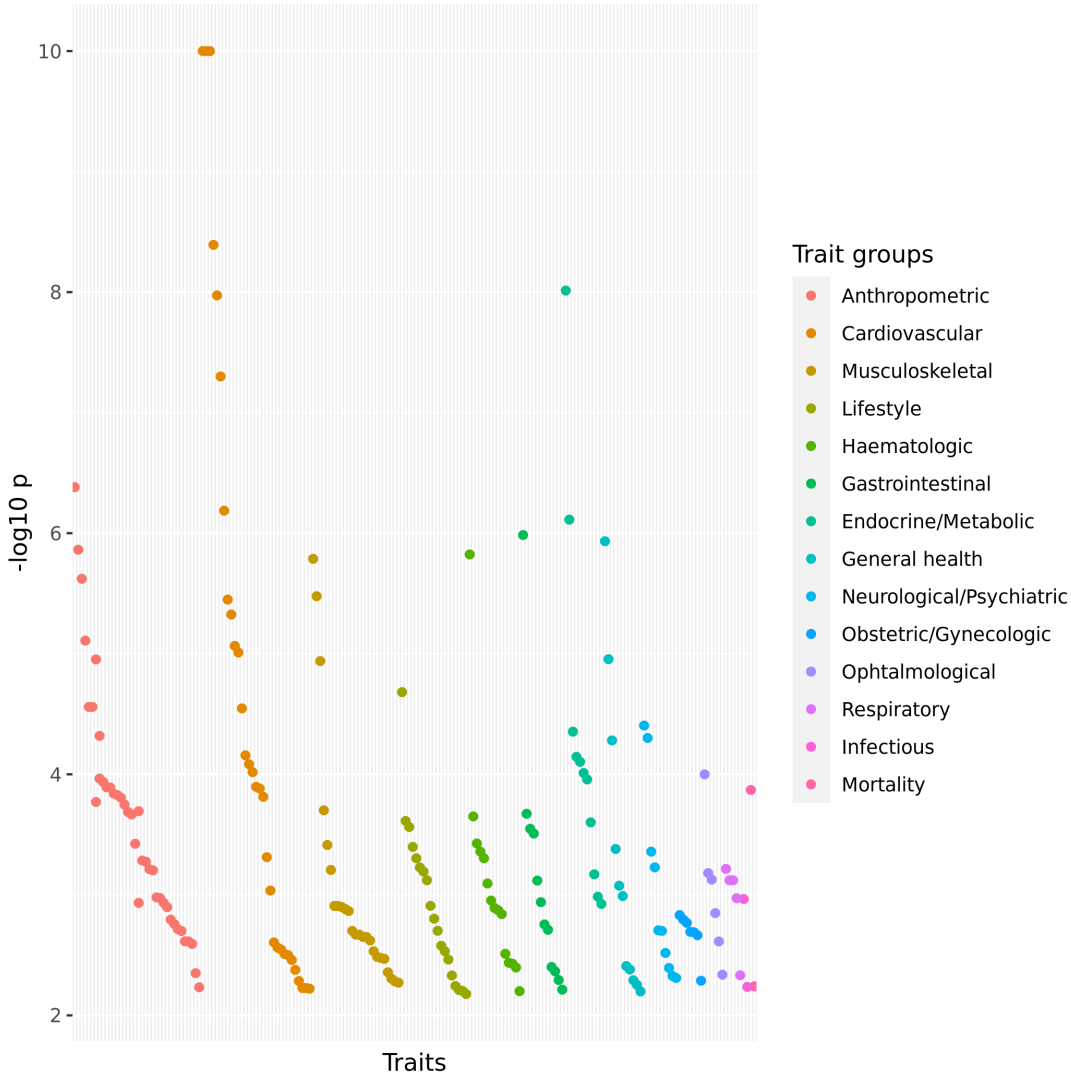
- Two-stage analysis:
  - Univariate comparison
    - Permissive threshold - filter out uninformative traits
  - Selection of traits
    - Hierarchical clustering - Random Forest

Univariate comparison:

- Pooled concordant (β_C) and discordant (β_D) effects for each trait
  - Random effects meta-analysis
- In each trait, are the effects different? ($|\beta_C - \beta_D|$)
- Retain traits:
  - Significant difference (FDR 10%)
  - Any of the two pooled estimates significant (FDR 10%)

Results of univariate comparison (n = 195)

- Significant differences (FDR 10%) between pooled estimates



SNP-Trait matrix

- For each SNP in each trait:

$$Z = \frac{\beta}{SE_\beta}$$

SNP	Discordant	Trait_1	Trait_2	...	Trait_p
SNP1	Yes	Z_{11}	Z_{12}	...	Z_{1p}
SNP2	No	Z_{21}	Z_{22}	...	Z_{2p}
SNP3	Yes	Z_{31}	Z_{32}	...	Z_{3p}
...	...	...	...	...	...
SNPn	Yes	Z_{n1}	Z_{n2}	...	Z_{np}

- High dimensionality
- Multicolinearity
- COVVSURF (Chavent *et al.* 2019)
  - Hierarchical clustering - Random forest

Hierarchical clustering

- At each ascending step:
  1. Starting point: each trait is a cluster
  2. Compute PCA between each pair
     - Obtain PC1 - Captures most variation of members
  3. Group most similar pair into a cluster
     - $\sum_{j=1}^{j} r^2_{x_j, PC_1}$
- Iterate until all traits are clustered together

PCA as measure of similarity

- At every possible partition (*i.e.* number of clusters k):

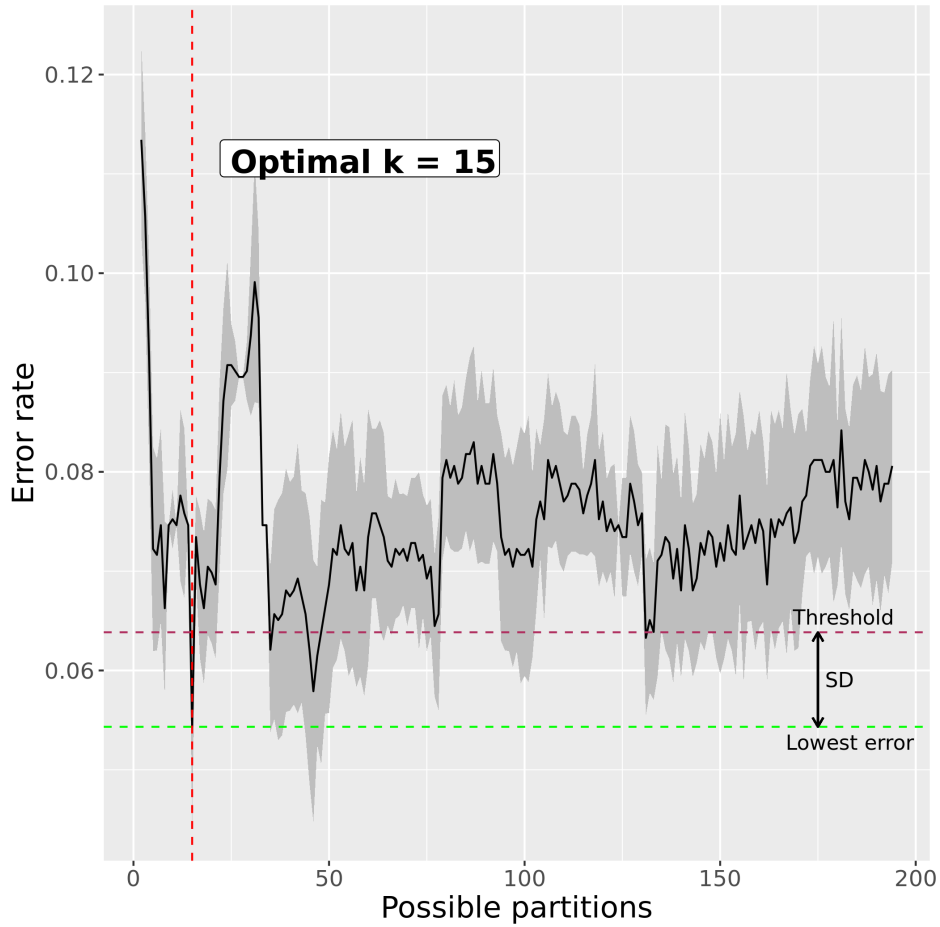
$$\text{Each cluster } (C_1, C_2, ..., C_k)$$
$$\Downarrow$$
$$\text{can be summarized by}$$
$$\Downarrow$$
$$\text{The correspoding PC1s } (PC1_1, PC1_2, ..., PC1_k)$$

Recap Random Forest (RF)

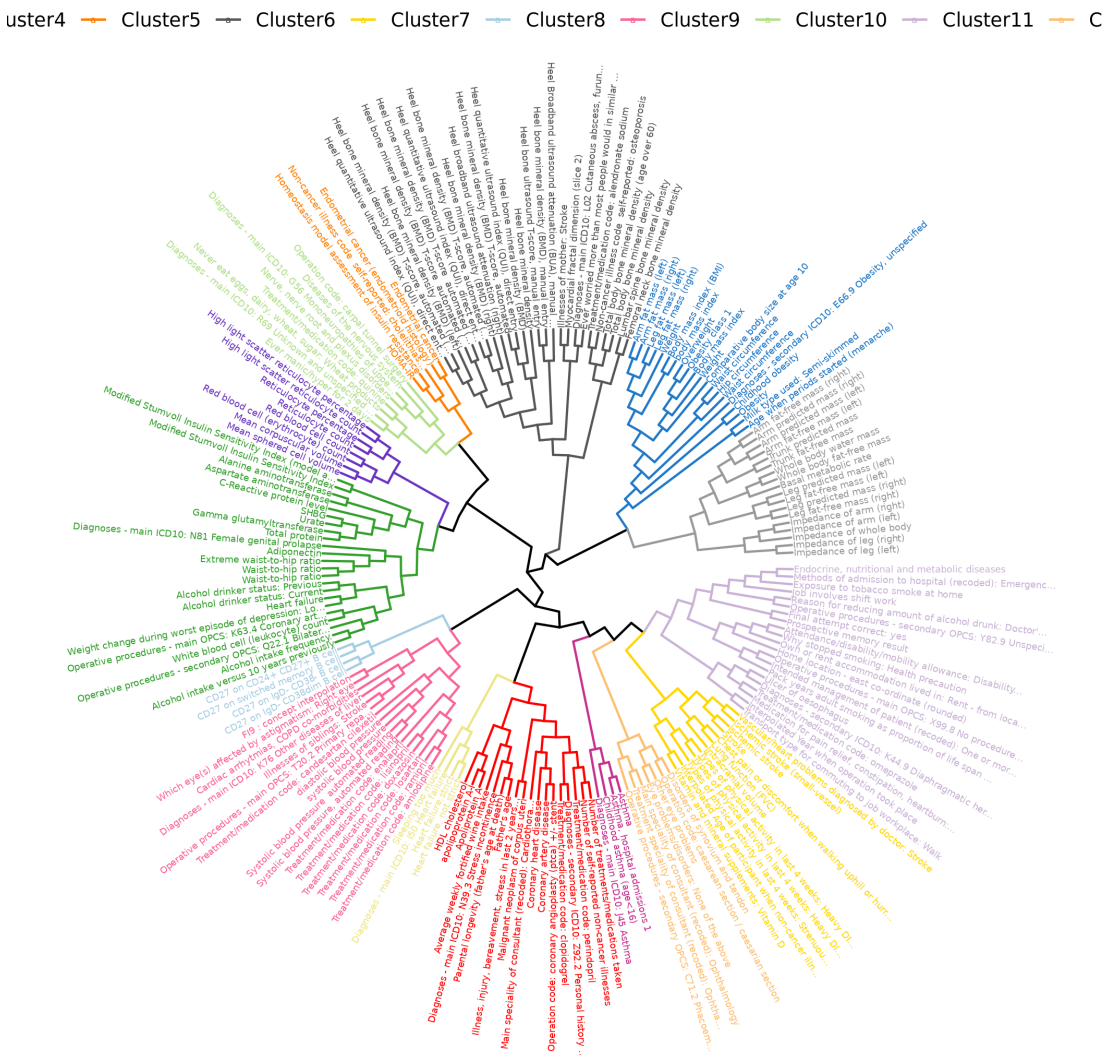
- Robust non-parametric classifier
- Multiple decision trees:
  - Grown using a random subset of data (in-bag)
  - At each split, selects the trait that minimizes variance in child nodes
  - Estimation of error rate / accuracy
    - Average out-of-bag (OOB) error of trees

Optimizing partition using RF

- At each k, take $PC1_1, PC1_2, ..., PC1_k$ as predictors for random forest
- Compute error rate
- Select k where the model has the lowest error rate

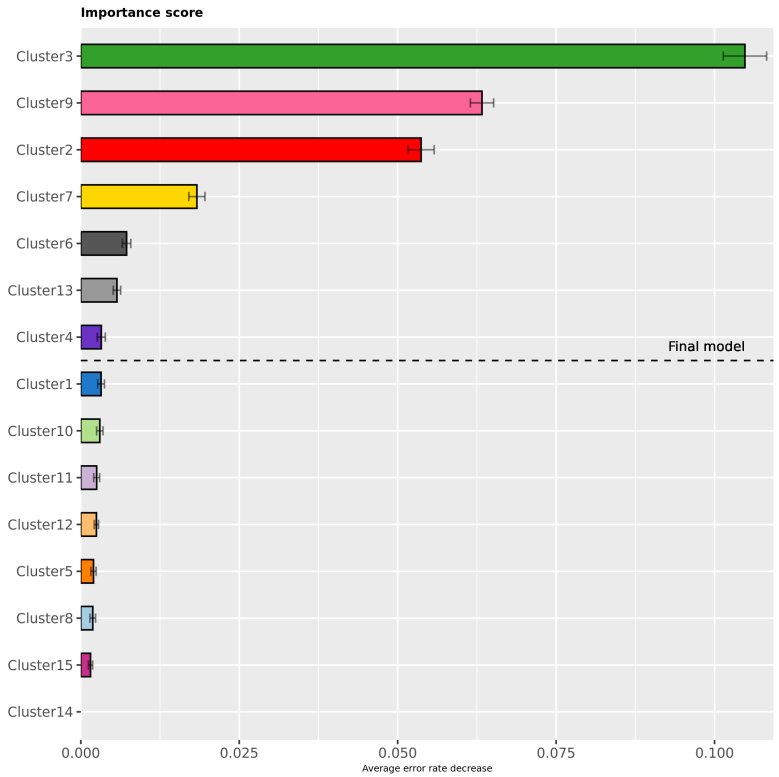


Clusters of traits



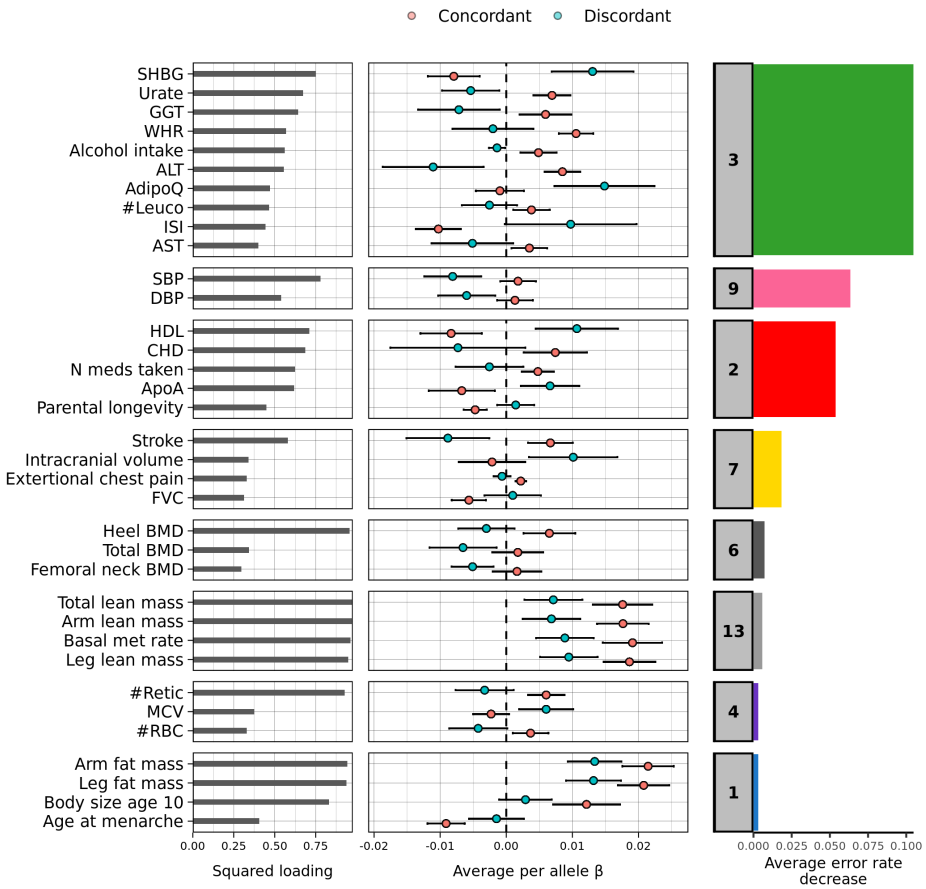
Importance score

- Imp_{v_p} = Average OOB error rate increase of trees when $v_p$is absent
- Nested models » Final model - minimum OOB error rate.

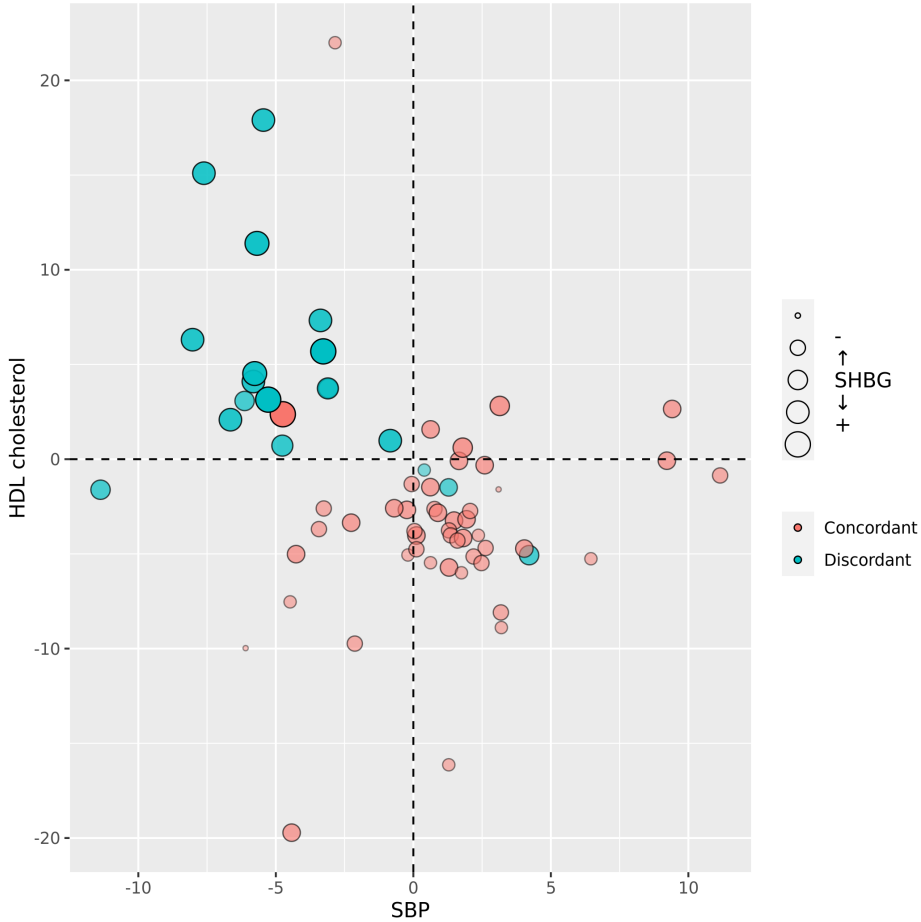


Final model

- The relevance of a cluster is given by the importance score
- Within each cluster, the relevance of a trait is given by its squared loading to the PC1 of the cluster

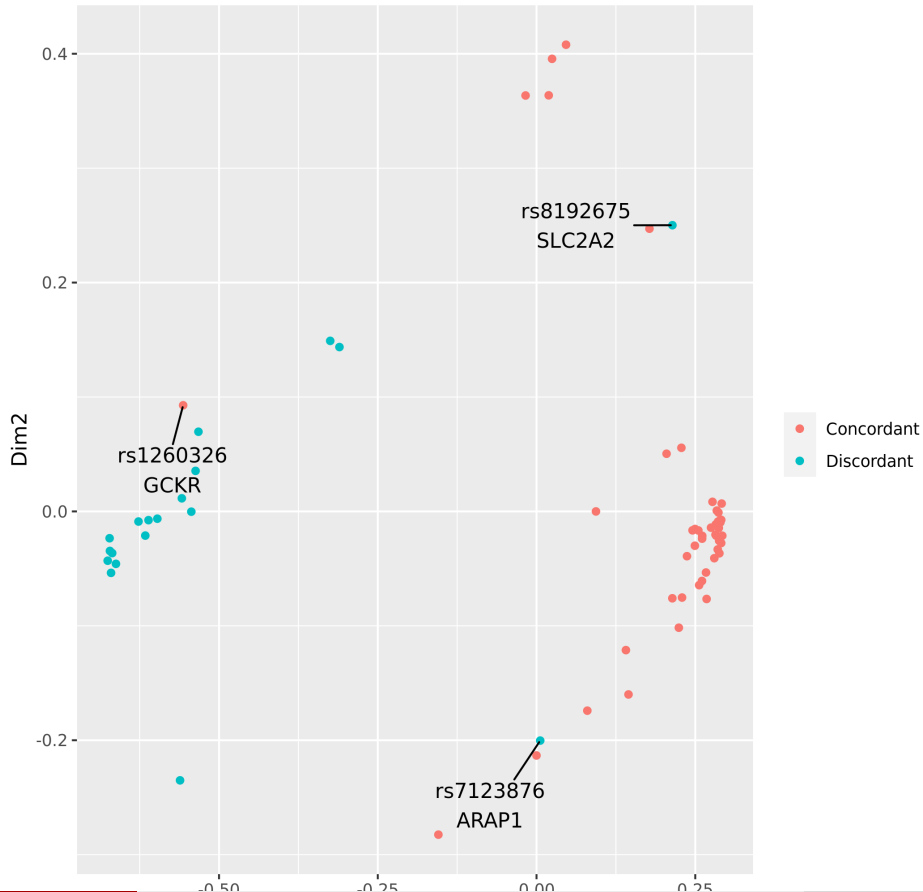


The 3 main variables

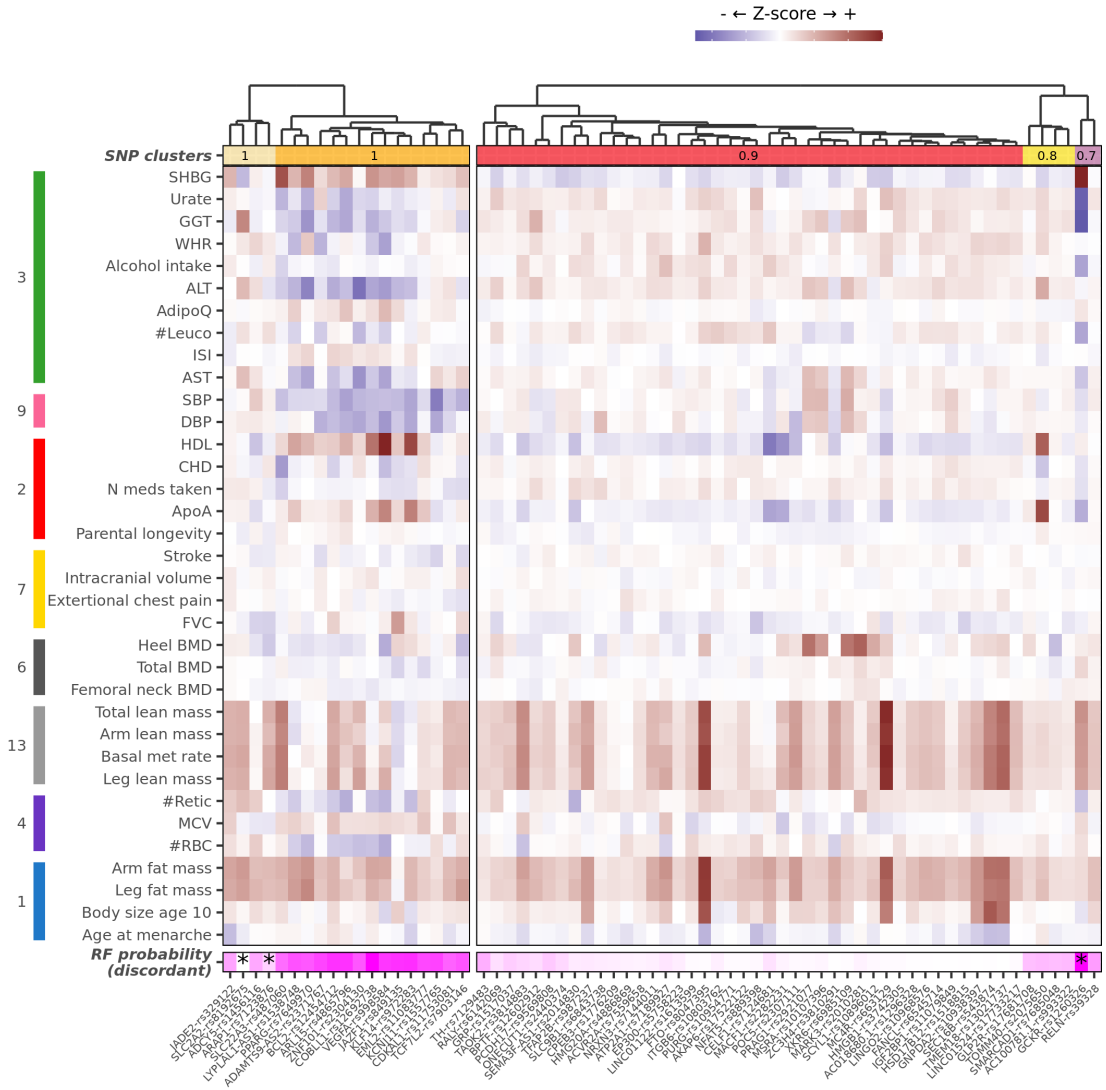


RF Proximity matrix

- RF proximity: N times two observations occupy the same terminal node
- SNP x SNP proximity matrix
- First two dimensions:

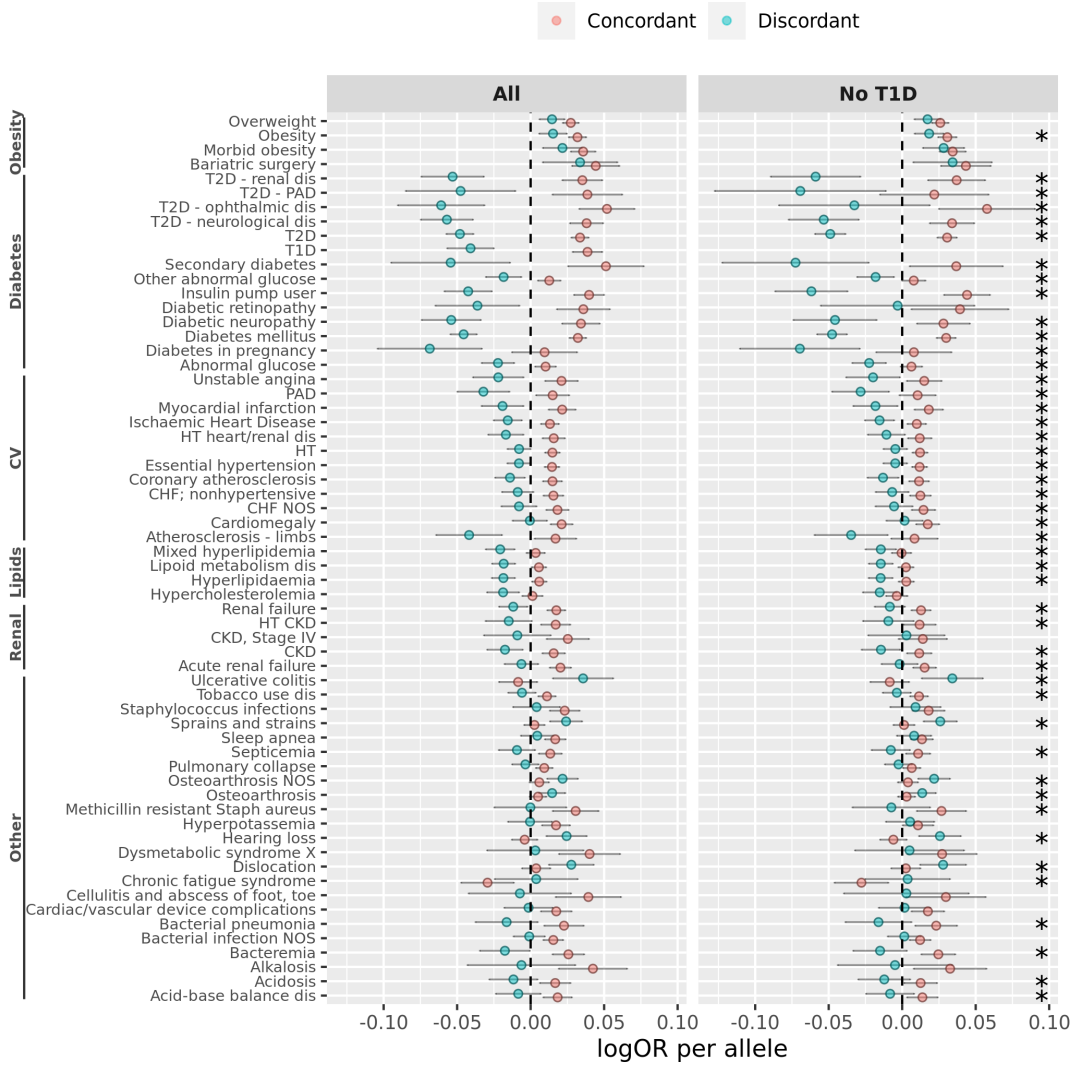


Heatmap of SNP effects on traits selected



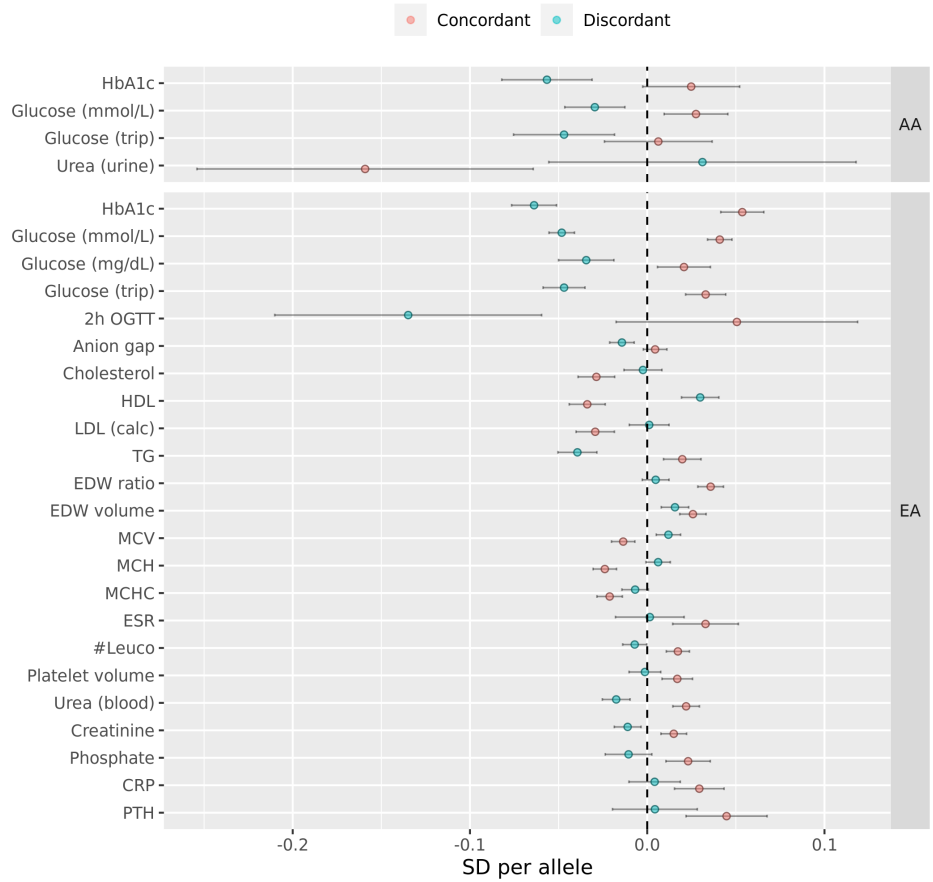
External validation - BioVU PheWAS

- Associations with diagnostic codes (FDR 5%)

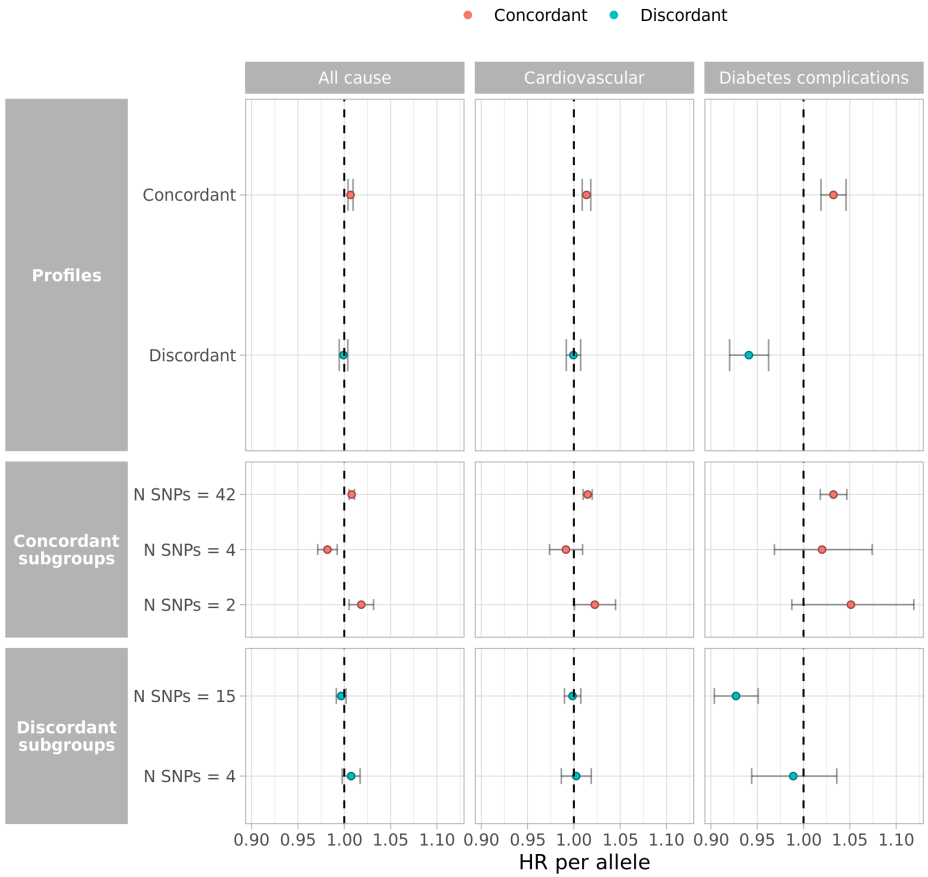


External validation - BioVU LabWAS

- Associations with median values of laboratory measurements per individual in BioVU (FDR 5%)



Mortality in UK Biobank - Concordant vs Discordant SNPs



Mendelian Randomization