

# Discordant variant analysis

Daniel Coral<sup>1</sup>

GAME unit

May 24, 2021

---

<sup>1</sup>daniel.coral@med.lu.se

# Outline

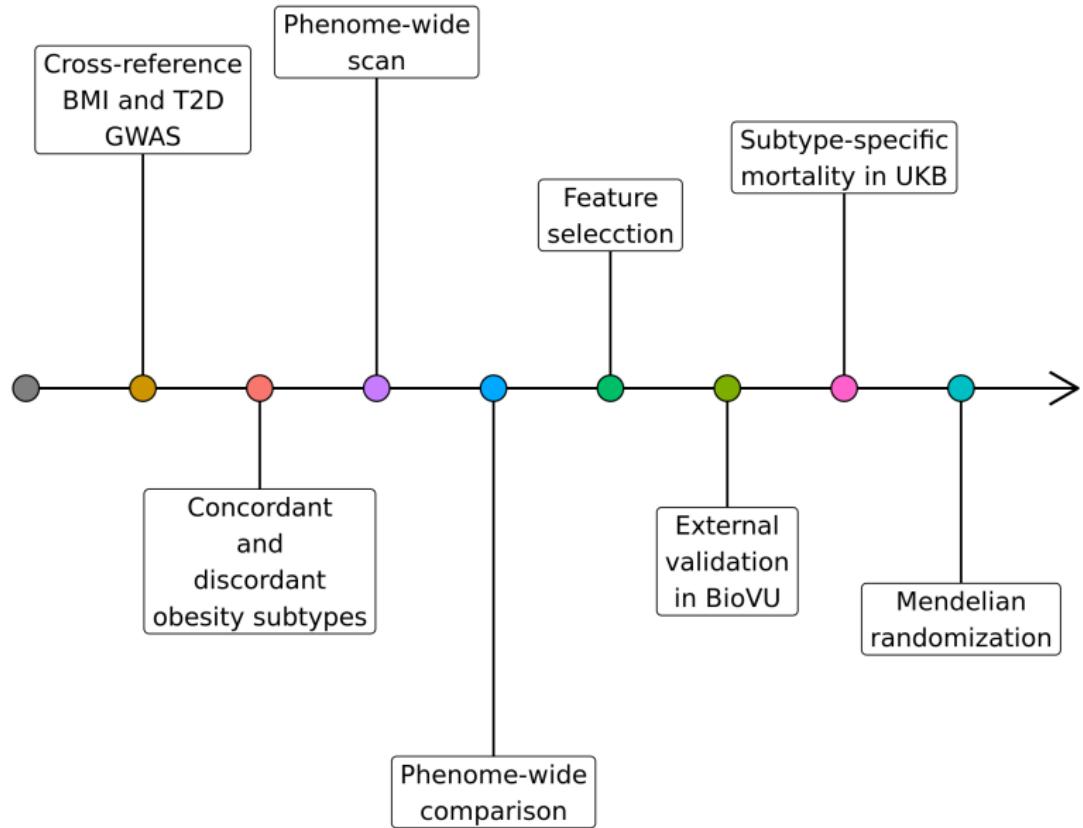
# Background

- The relationship between obesity and cardiometabolic risk is highly heterogeneous.
- Individuals in the same BMI level may display different cardiometabolic risk:
  - ~ 30% of people with a  $\text{BMI} > 30 \text{ kg/m}^2$  display a protective metabolic profile.
  - ~ 30% of people with  $\text{BMI} < 25 \text{ kg/m}^2$  develop conditions usually associated with obesity.
- Better risk stratification and more precise interventions may improve outcomes.
- Obesity and T2D are closely linked.
- T2D is in turn, linked to many cardiovascular diseases, and is life-threatening.
- Factors uncoupling obesity from T2D risk may be used to define clinically relevant subgroups of obesity.

# Aims

- Use genetics to stratify obesity into 2 subtypes:
  - Concordant ( $\uparrow$  BMI and  $\uparrow$  T2D)
  - Discordant ( $\uparrow$  BMI and  $\downarrow$  T2D).
- Perform a phenome-wide scan to detect traits that reinforce this stratification.
- Leverage this information for further stratification with potential clinical meaning.

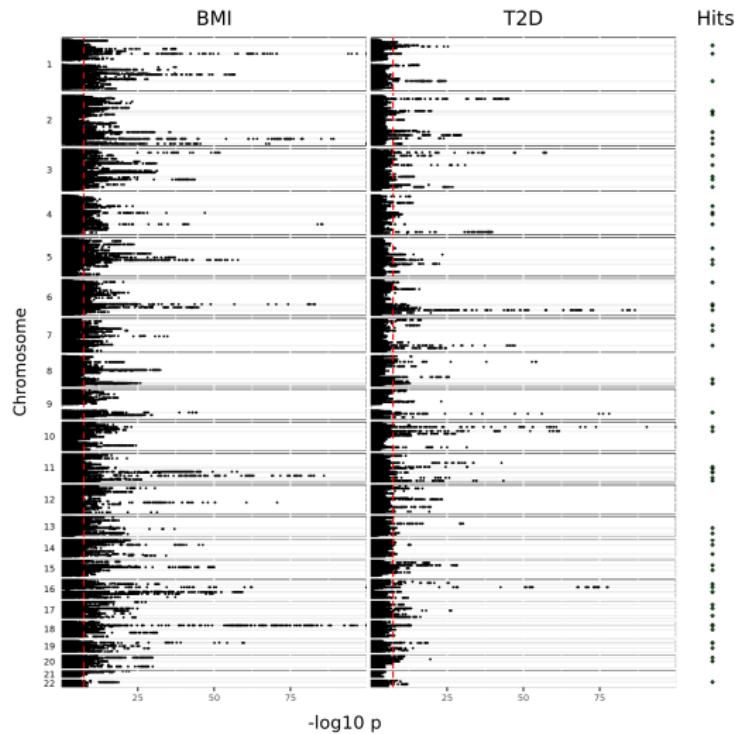
# Analysis pipeline



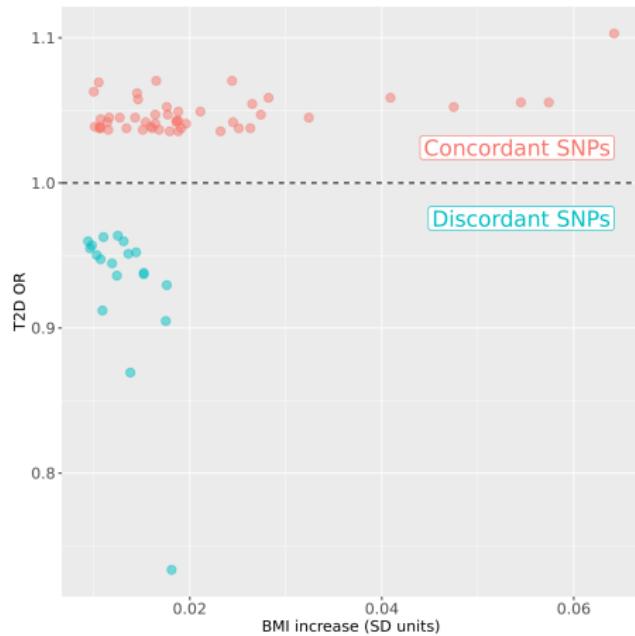
# BMI and T2D GWAS

- BMI: Yengo *et al.* (2018) vs T2D: Mahajan *et al* (2018).
  - Common biallelic variants (MAF > 1%)
  - No INDELs
  - No potentially ambiguous palindromic SNPs (MAF > 30%)
  - More than 20 % difference in MAF with 1000G EUR
  - Clumpling ( $r^2 < 0.01$  over 500kb window in 1000G EUR)

# Cross-referencing



# Assembly of concordant and discordant profiles



**Figure:** Lead variants aligned to the BMI increasing allele and stratified by their  $\beta$  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
- More than 25 minor alleles in smallest group for binary traits
- Studies with information for all reference SNPs
- If multiple studies for a single trait:
  - The study with the highest sample size was selected
- ~ 3500 traits

# Phenome-wide comparison

- Two-stage analysis
  - Univariate comparison - Random effects meta-analysis
  - Clustering and selection of traits - Random forest

# Phenome-wide comparison - Univariate analysis

- Pooled concordant ( $\beta_C$ ) and discordant ( $\beta_D$ ) effects for each trait
  - ① Standardized beta coefficients:  
$$SE = 1/\sqrt{2 * MAF * (1 - MAF) * (n + Z^2)}$$
  - ② Random effects meta-analysis (Paule-Mandel  $\tau$  estimator)
- Difference between pooled estimates ( $D$ )

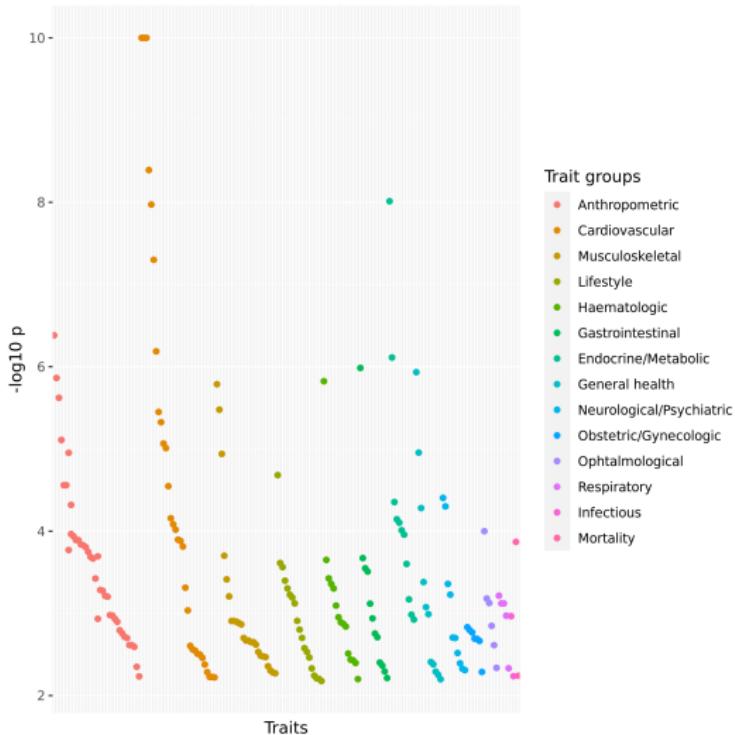
$$D = |\beta_C - \beta_D|$$

$$SE_D = \sqrt{SE_C^2 + SE_D^2}$$

$$Z = \frac{D}{SE_D} \sim \mathcal{N}(0, SE_D^2)$$

- Selection of traits: (  $\beta_C \neq 0$  |  $\beta_D \neq 0$  ) &  $D \neq 0$ 
  - FDR 10%
- 195 traits

# Results of univariate comparison



**Figure:** Significant P values (FDR 10%) for difference between pooled estimates  
(D)

# Phenome-wide comparison - SNP-Trait matrix

- For each SNP in each trait:

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

- Produces a 67 SNP x 195 traits matrix with:
  - High dimensionality
  - Multicollinearity
- COVVSURF (*Chavent et al.* 2019)
  - Hierarchical agglomerative clustering - PCA
  - Random Forest

# PCA

- PCs are linear combinations of original variables
  - Redundant variables can be summarized
  - The contribution of a variable in a PC = Squared loading (Pearson's  $r^2$ )
- Maximum possible information (variance) is summarized in PC1

# Random Forest (RF)

- Multiple decision trees - Classification or regression
  - Grown using a random subset of columns and rows of the original data (in-bag)
    - Replacement
  - Tested using out-of-bag (OOB) data
    - Estimation of error rate / accuracy
- The output includes:
  - Probability (*votes*)/value assigned by trees
  - Importance score
    - Error rate increase if variable is absent
  - Proximity matrix
    - N times two observations occupy the same terminal node
    - From Breiman and Cutler:

$1 - prox(n_i, n_j)$  are squared distances in a Euclidean space

# Clustering of traits using RF

- Dimensionality reduction into clusters of traits
- Relevant for distinguishing between the two sets of SNPs
- Non-parametric, data-driven
- Process:
  - ① Agglomerative clustering using PCA results in:
    - Clustering tree with  $k = 1, 2, 3, \dots / p$  possible partitions
    - At every  $k$ : the 1PCs summarize each cluster
  - ② Among every  $k$ : 1PC as predictors for RF
  - ③ Model with minimum OOB error rate determines the optimal  $k$ .
  - ④ Nested RF models
    - Starting with model with only the most important cluster
    - Ending with model including all clusters selected in ③
    - Final model - minimum OOB error rate

# RF models

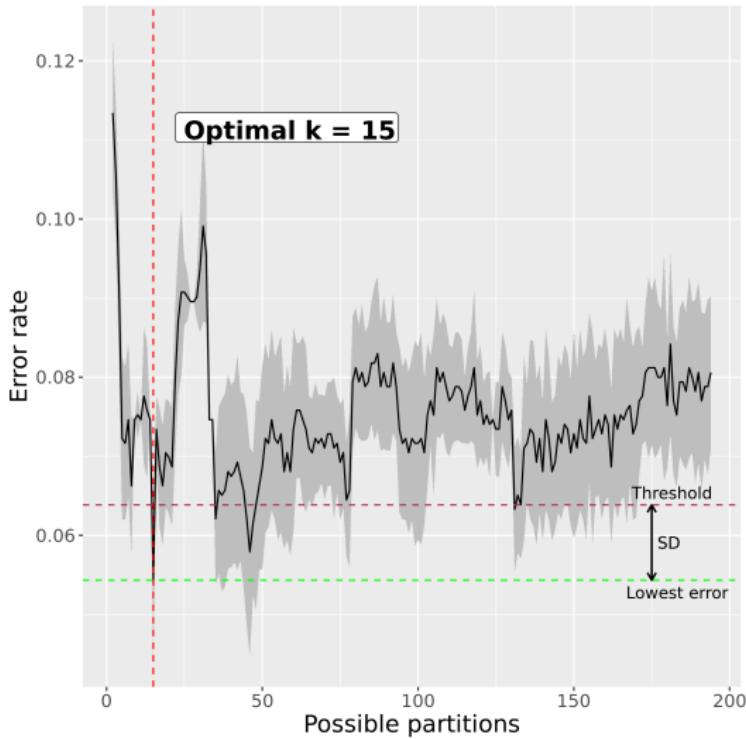


Figure: Error rate of RF models across every  $k$

# Clusters



Figure: Clustering tree with optimal partition

# Trait selection - First stage

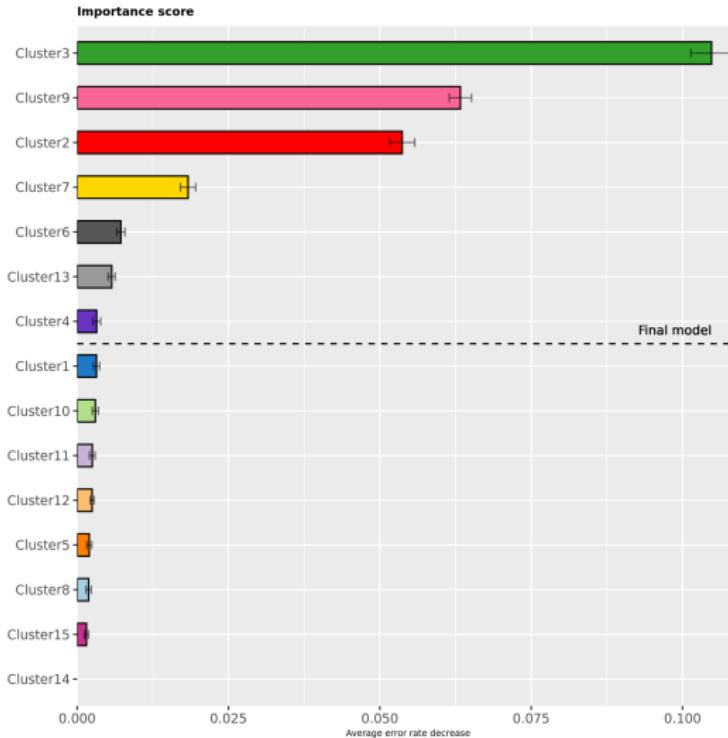


Figure: Importance score of clusters

# Trait selection - Final model

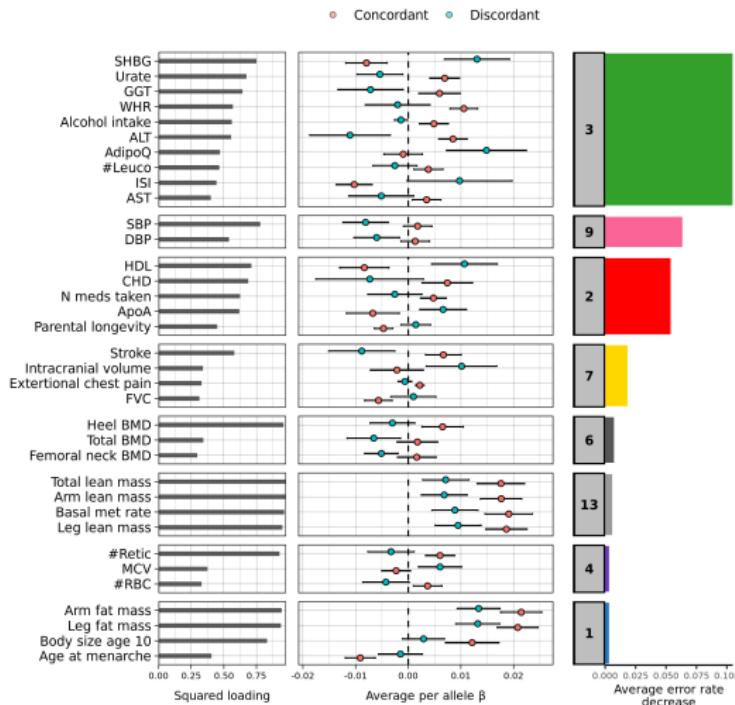


Figure: Importance score, squared loading and pooled estimates of each subtype

# Main variables

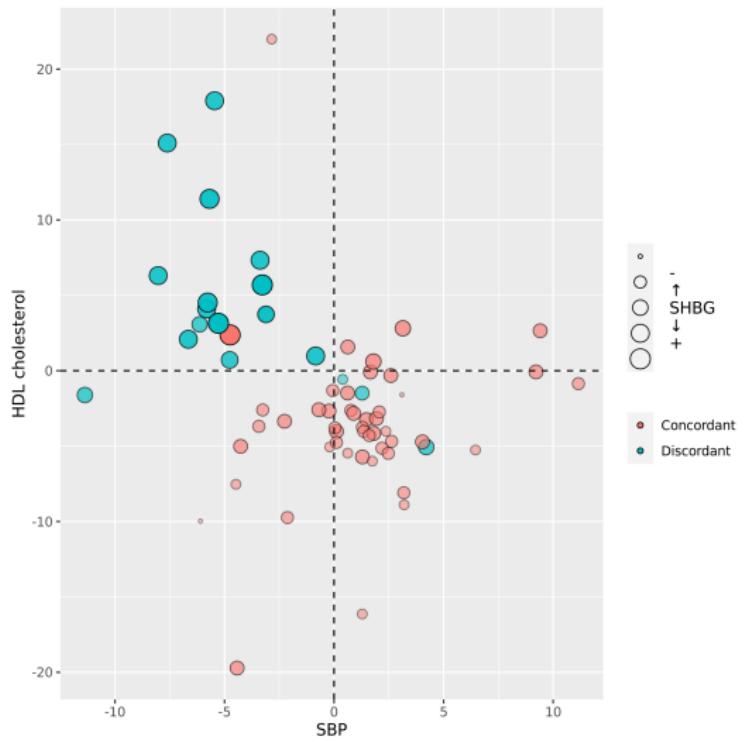


Figure: SNP effects for the main three variables

# Proximity matrix

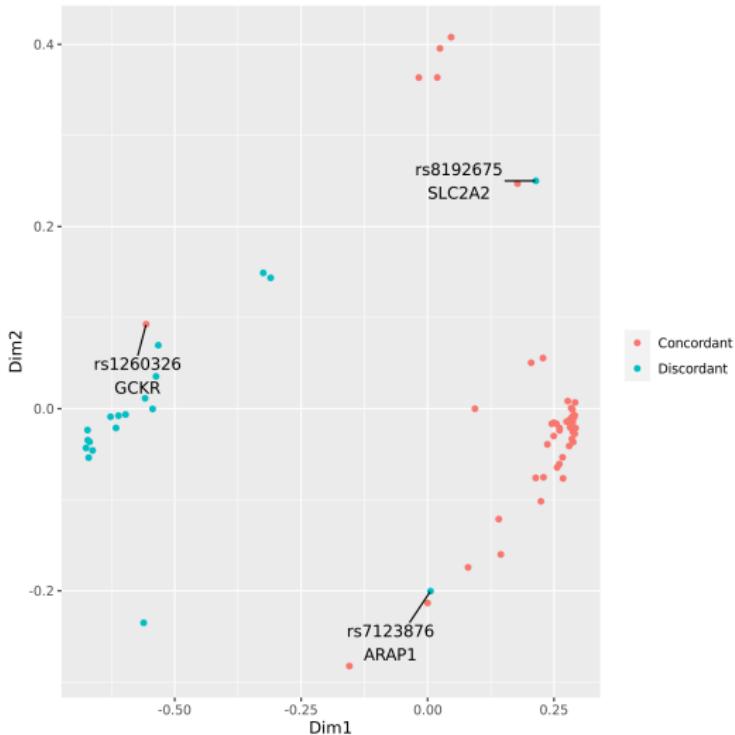


Figure: Multi-dimensional scaling of RF proximity matrix

# Individual SNP effects on traits selected

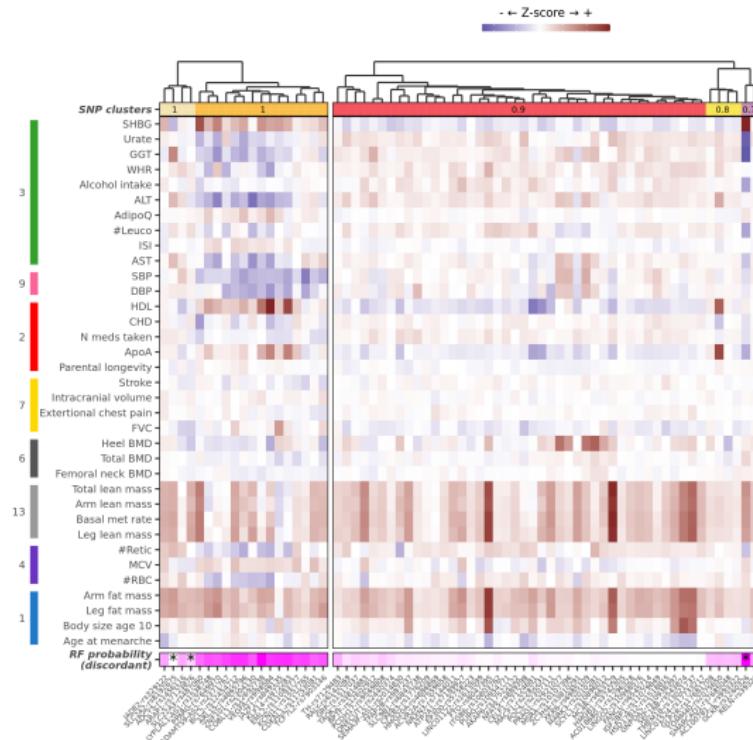


Figure: Heatmap of main variables and RF assignment

# External validation - BioVU PheWAS

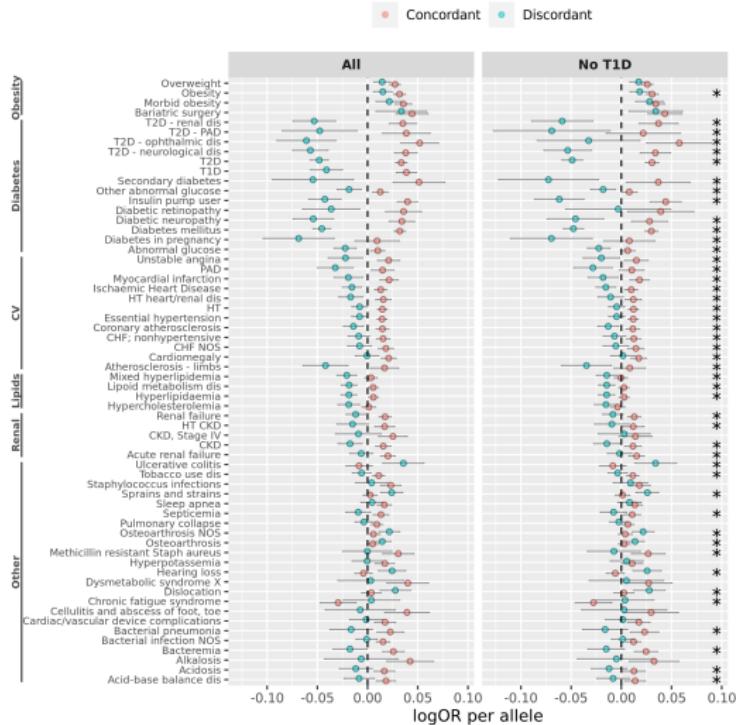


Figure: Concordant and discordant effects in BioVU. Stars represent 5% FDR significance after exclusion of individuals with T1D

# External validation - BioVU LabWAS

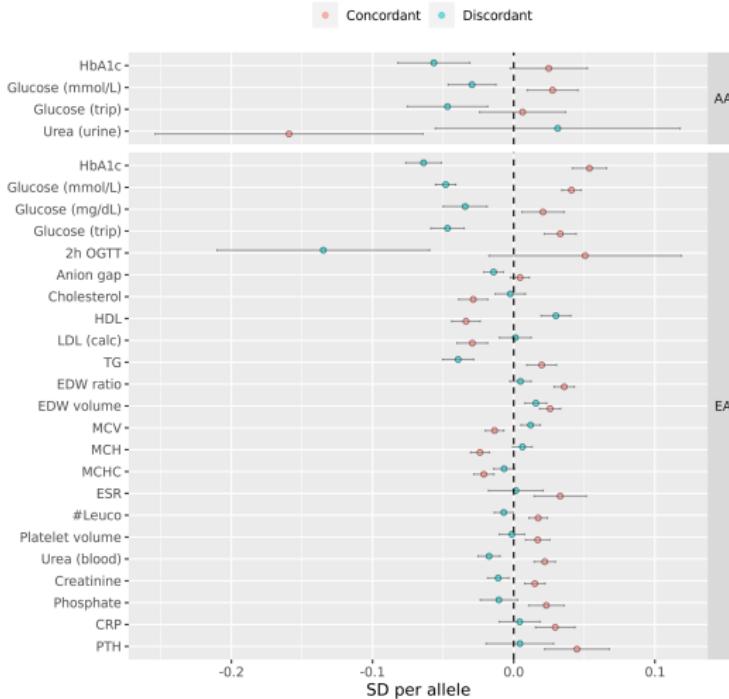


Figure: Concordant and discordant weight gain in BioVU

# Mortality in UK Biobank - Concordant vs Discordant SNPs

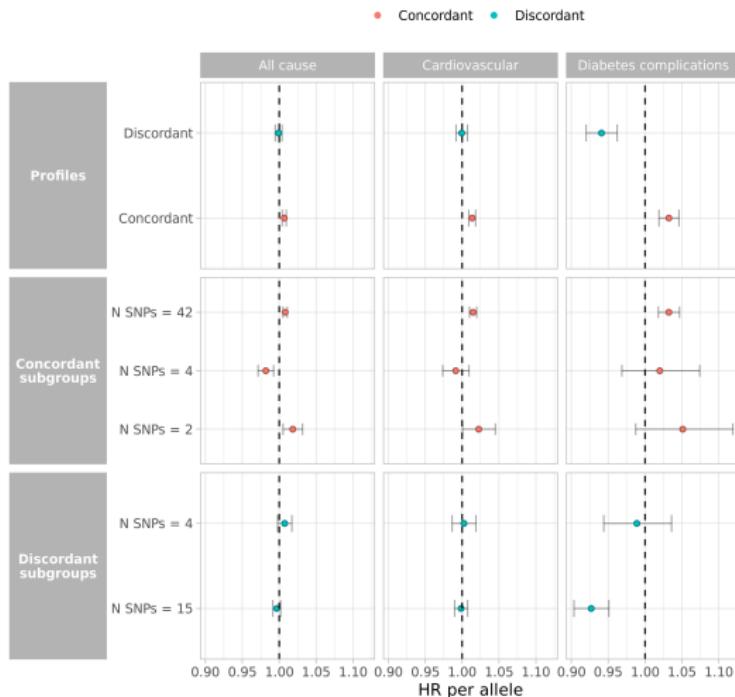
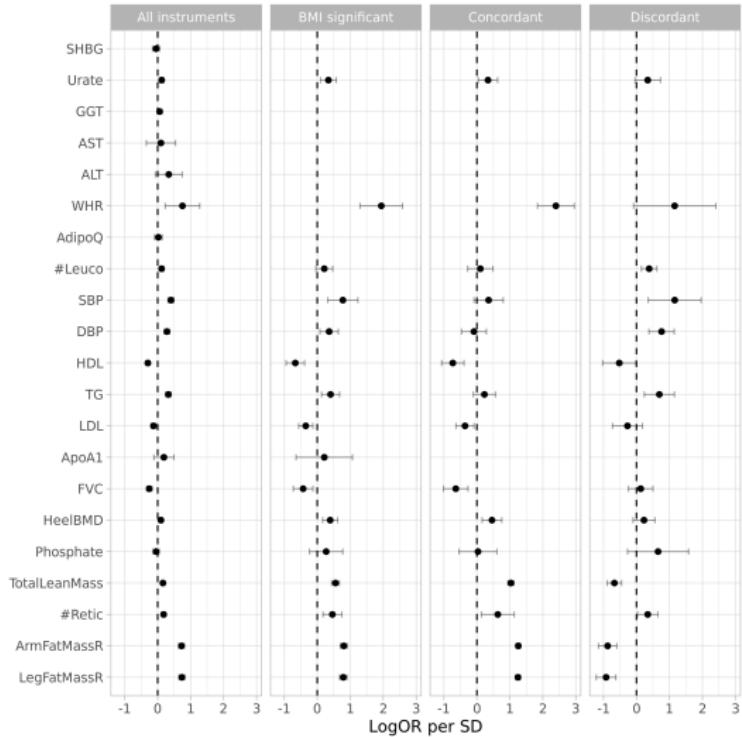


Figure: Survival analysis of concordant and discordant PRS

# Mendelian Randomization



# Conclusions

- Concordant and discordant SNPs differ mainly in:
  - Liver enzymes - Central adiposity
  - Blood pressure
  - Lipids
- And differ less strongly in:
  - BMD
  - RBC counts
  - Overall adiposity
- The two subgroups are not homogeneous
- The causal pathways might differ

# Strengths and limitations

- Mechanisms uncoupling obesity from T2D
- Agnostic approach
- Reverse causality and confounding
- Only EUR
- Thresholds chosen affect:
  - Genetic factors chosen
  - Clusters built
- Smaller sets of instruments in stratified MR

# Possible future applications

- Other levels of the phenome
- Random forest proximity matrix to find more subtypes