ANALYSIS OF HUMAN DISEASE WITH EXOME SEQUENCE ASSOCIATION DATA

Uliana Kniaziuk, Timofei Ryko,
Olga Bochkareva & Sofia Buyanova

Although many genotype-phenotype associations are known, it is often difficult to understand what function of the protein was altered by a particular mutation, which molecular processes will be affected and how this contributes to the phenotype. Subsetting only SNPs affecting amino acids in ligand binding sites can potentially help to make meaningful assumptions about metabolic or regulatory processes affected in genetic diseases, and make predictions for therapy. We developed a pipeline, which allows us to subset ligand-binding amino acids, analyzed resulting SNPs to check whether the results are meaningful, and suggest mechanisms for several diseases based on our data.

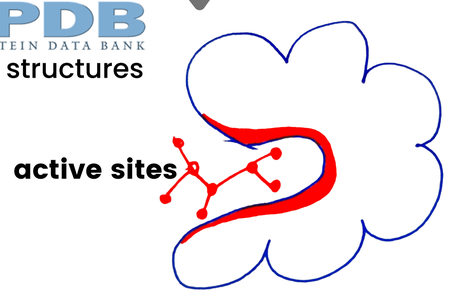
Pipeline developing

nature biobank

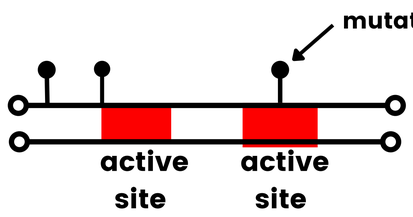
Data: Analysis of exome sequencing of 454,787 UK Biobank participants [1]
Identified associations of SNPs with 4000 traits (p-values)

Annotation: genes, protein

RCSB PDB PROTEIN DATA BANK
3D structures



Subset the variants in ligand-binding sites



New p-value threshold

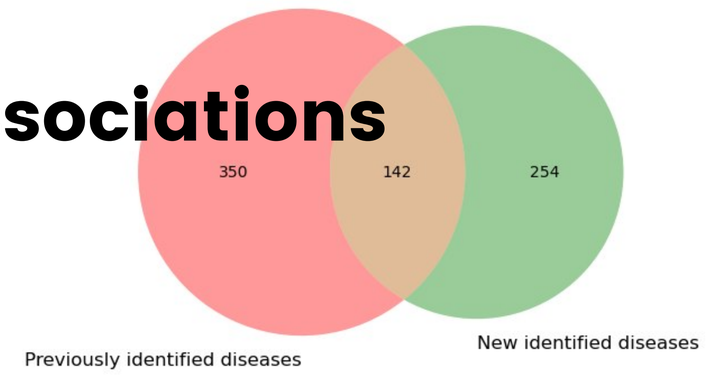
New associations

Which cells are mostly affected by the identiified variants? → THE HUMAN PROTEIN ATLAS

Which signaling and methabolic pathways are affected by these variants? → KEGG

What are clinical consequences of these variants? → ClinVar

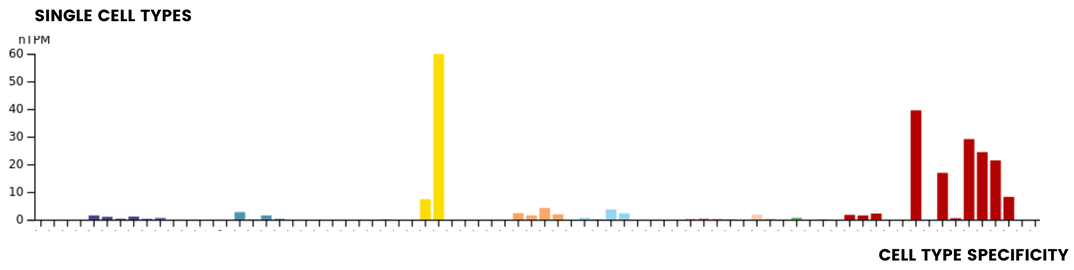
How to explain observed interactions with ligand? → STRING + Research articles

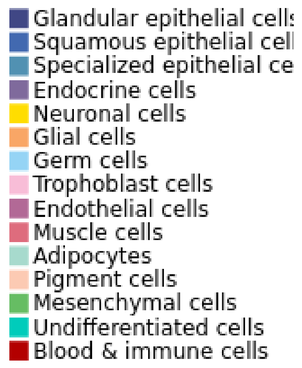


Pipeline evoluation

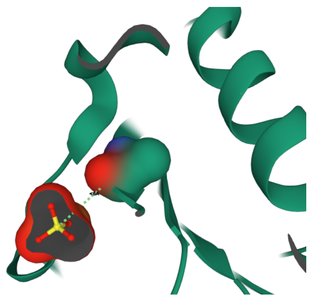
- **Case studies demonstrate that found associations are biologically relevant**

FLT3
FMS RELATED RECEPTOR TYROSINE KINASE 3





ligand and active site



Pathway

- **MAP04640**
  HEMATOPOIETIC CELL LINEAGE
- **MAP05221**
  ACUTE MYELOID LEUKEMIA

Disease

- **H00003**
  ACUTE MYELOID LEUKEMIA
- **H02412**
  ATYPICAL CHRONIC MYELOID LEUKEMIA

- **Proposed subset method reveals new associations with diseases**

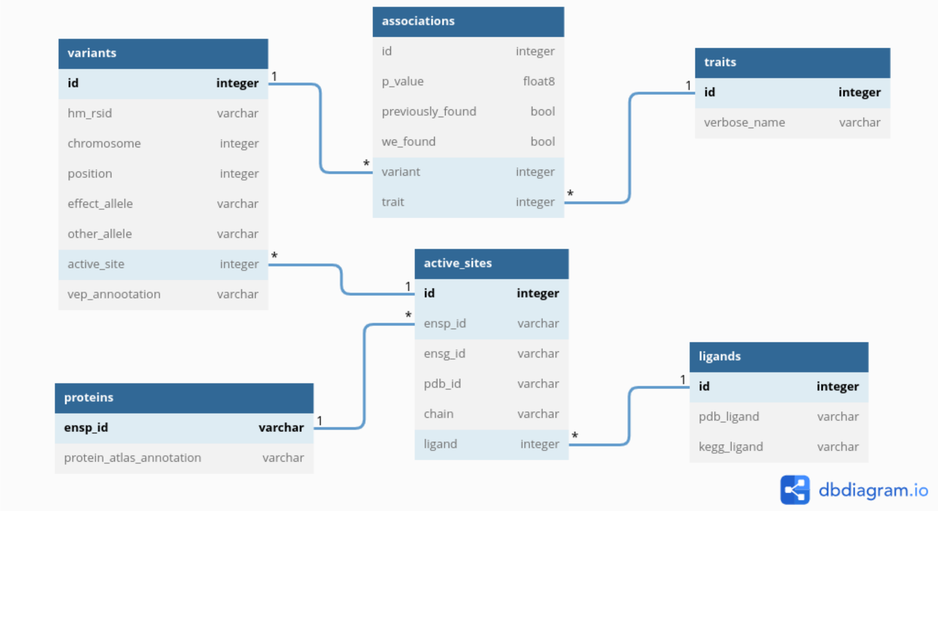
LYZ
Lysozyme
ENSP00000261267

Trait:
Monocyte count
Nature article

Trait:
Neutrophill count
Our study
could not be identified before

Gene expressed in both cell types, and traits are related

- **Proposed database structure**



Conclusions

We created a first version of a pipeline, which helps find associations which are biologically interpretable. We used existing expression, protein interaction data, mice knockout data and existing variation data to verify that we get meaningful results. During our research we found multiple ways to improve SNP prioritization and pipeline automation and are planning to implement these ways.

Acknowledgments: Peter Vlasov, Evgeny Akkuratov