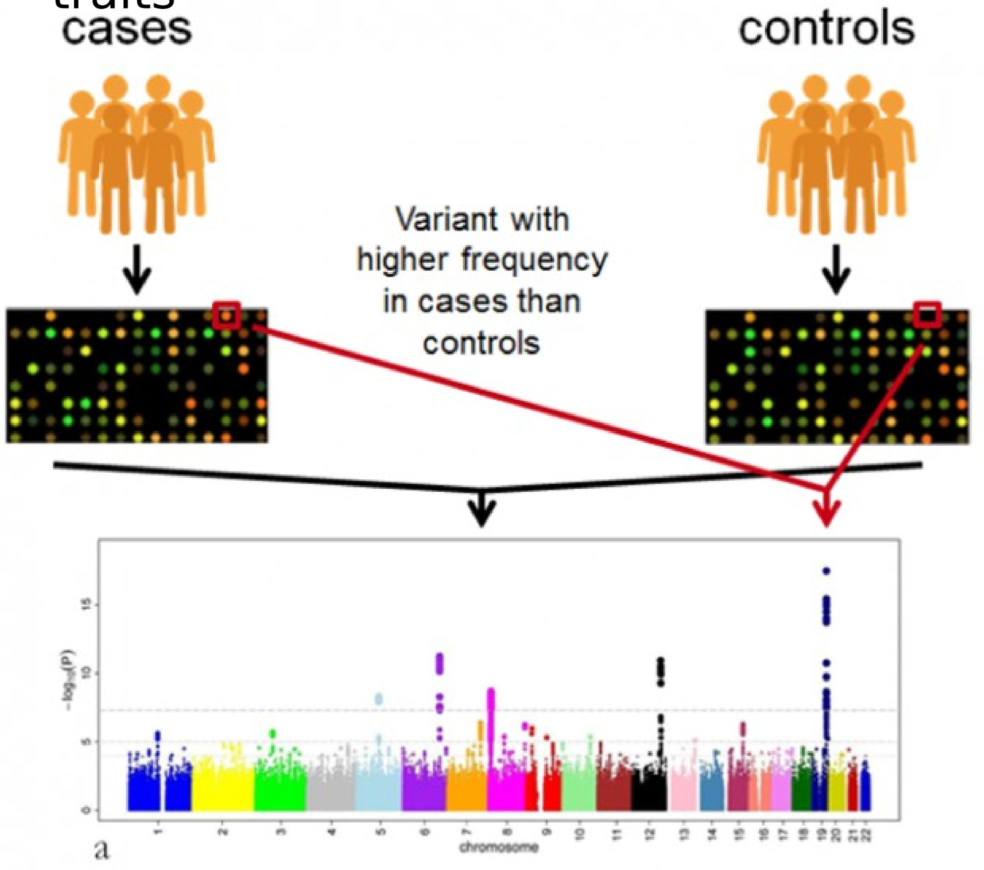
G2PT: Mechanistic genotype-phenotype translation using hierarchical transformers

Ingoo Lee
Postdoctoral Researcher
School of Medicine, University of California, San Diego
Ideker Laboratory

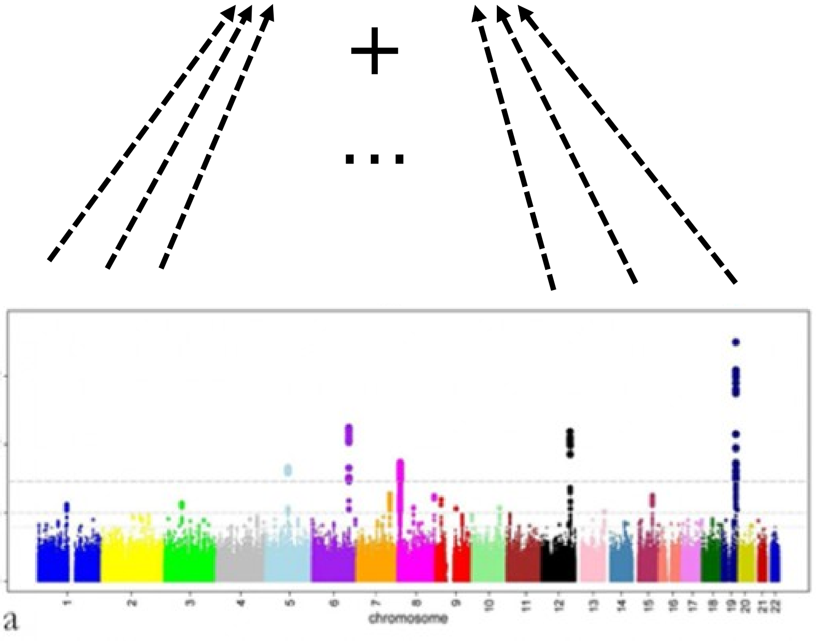
ASPLOS

Genome-Wide Association Study (GWAS)

- **GWAS** identifies genetic variations associated with specific diseases or traits

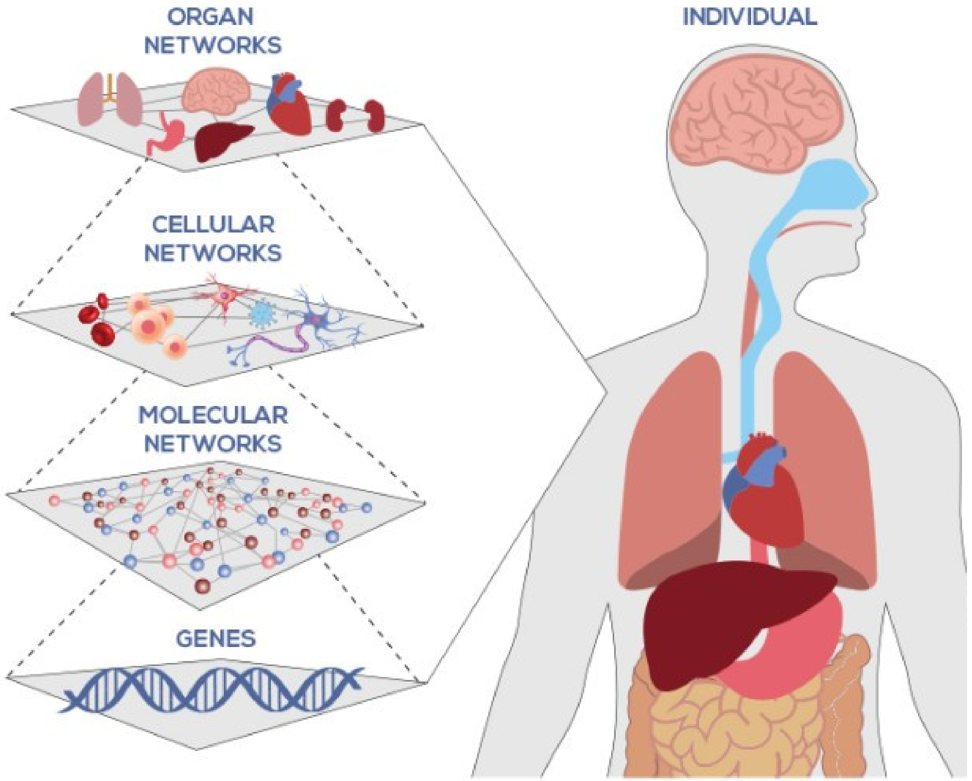


$$y_i = \beta\mu + \boxed{\sum_{j}^{N_{snp}} \alpha_j \cdot SNP_{ij}} + \sum_{k}^{N_{cov}} \gamma_k \cdot Cov_{ik}$$



https://www.ebi.ac.uk/training/online/courses/gwas-catalogue-exploring-snp-trait-associations/what-is-gwas-catalog/what-are-genome-wide-association-studies-gwas/

System biology: A Holistic View to Phenotypes



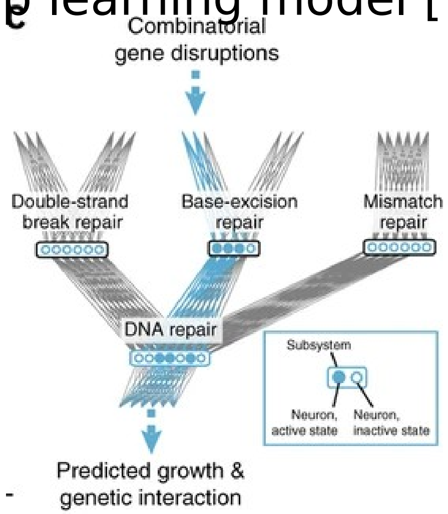
- Phenotypes cannot be interpreted as a sum of individual effects of SNPs.

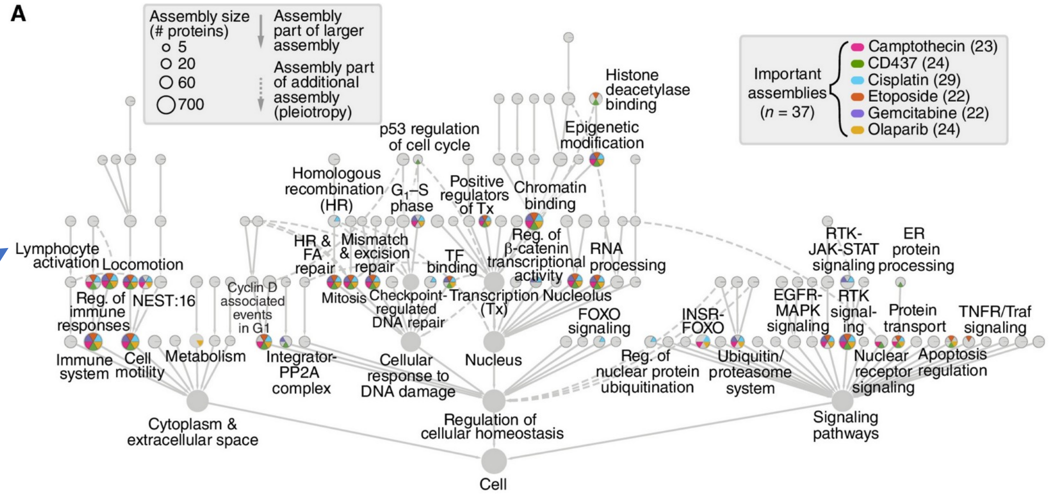
- Biological functions work as systems
  - Not individual SNP
  - Not individual gene

- Systemic analysis can give proper explanation to phenotype

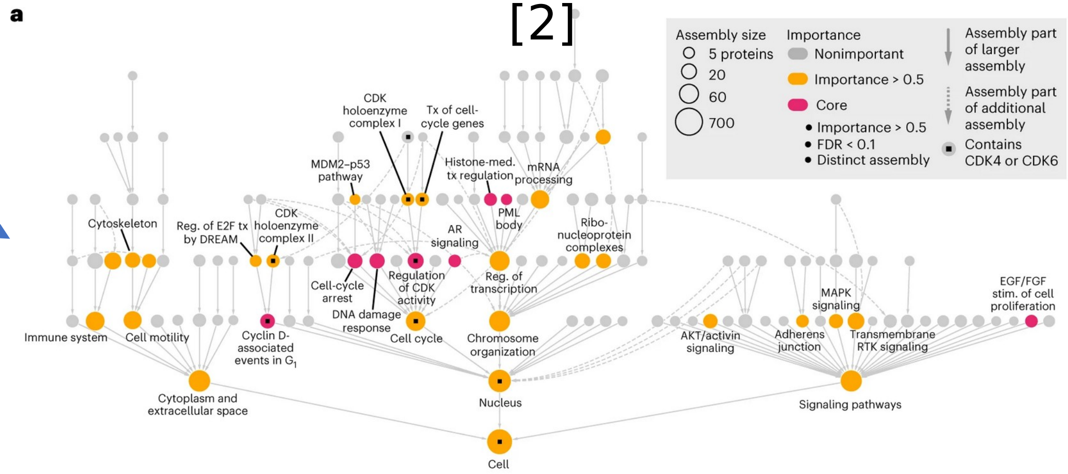
Providing a Prior Knowledge Structure Can Enhance Performance and Interpretability

Providing
Biological Hierarchy as
Structure of
Deep learning model [1]





Replication Stress Drugs [2]

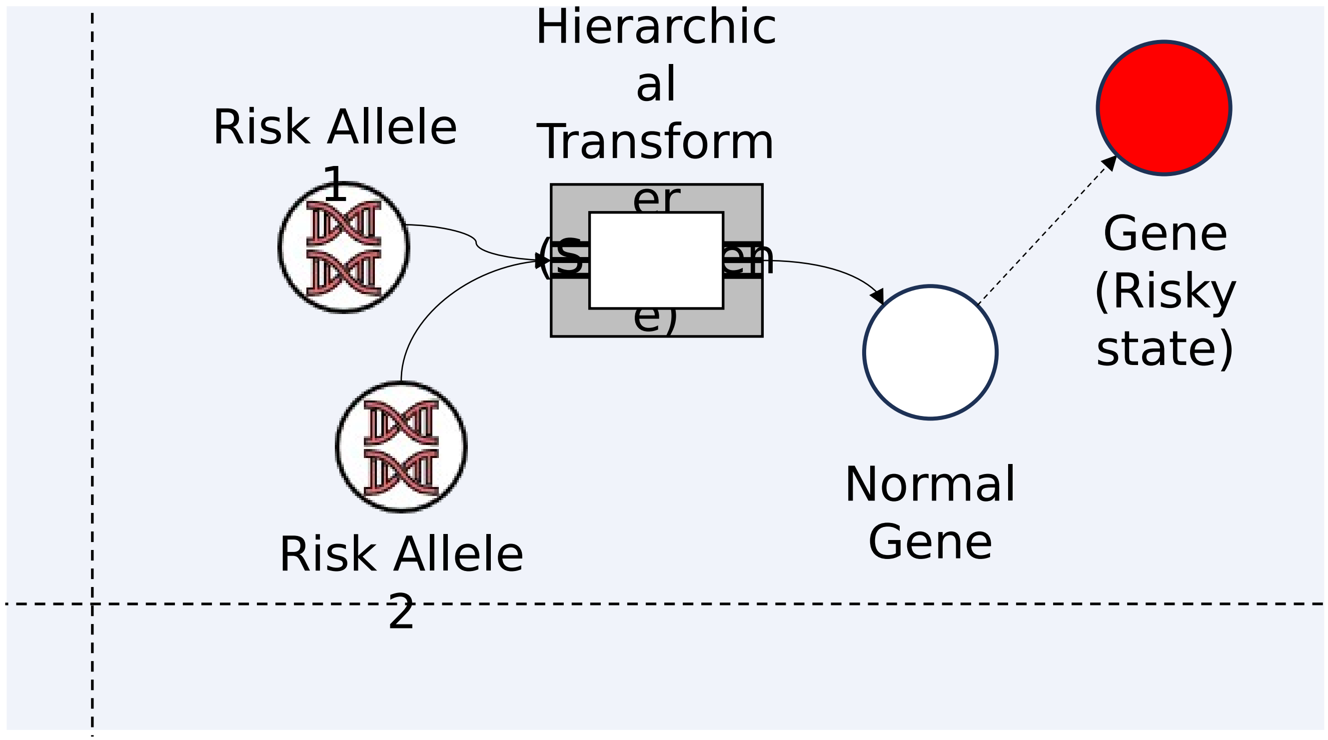


Palbociclib (CDK 4/6 Inhibitor) [3]

1. Kuenzi, Brent M., et al. "Predicting drug response and synergy using a deep learning model of human cancer cells." *Cancer cell* 38.5 (2020): 672-684.
2. Zhao, Xiaoyu, et al. "Cancer mutations converge on a collection of protein assemblies to predict resistance to replication stress." *Cancer Discovery* 14.3 (2024): 508-523.
3. Park, Sungjoon, et al. "A deep learning model of tumor cell architecture elucidates response and resistance to CDK4/6 inhibitors." *Nature Cancer*

Hierarchical Transformer (HiTR)

Biological Embedding Space



Risk Allele 1

Risk Allele 2

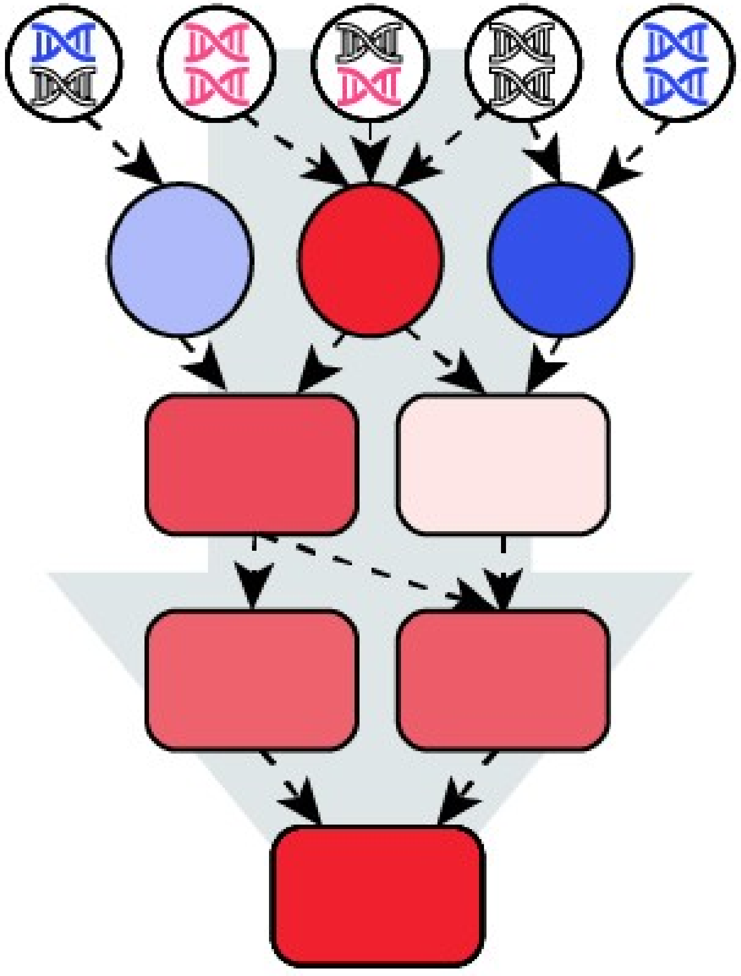
Hierarchical Transformer

Normal Gene

Gene (Risky state)

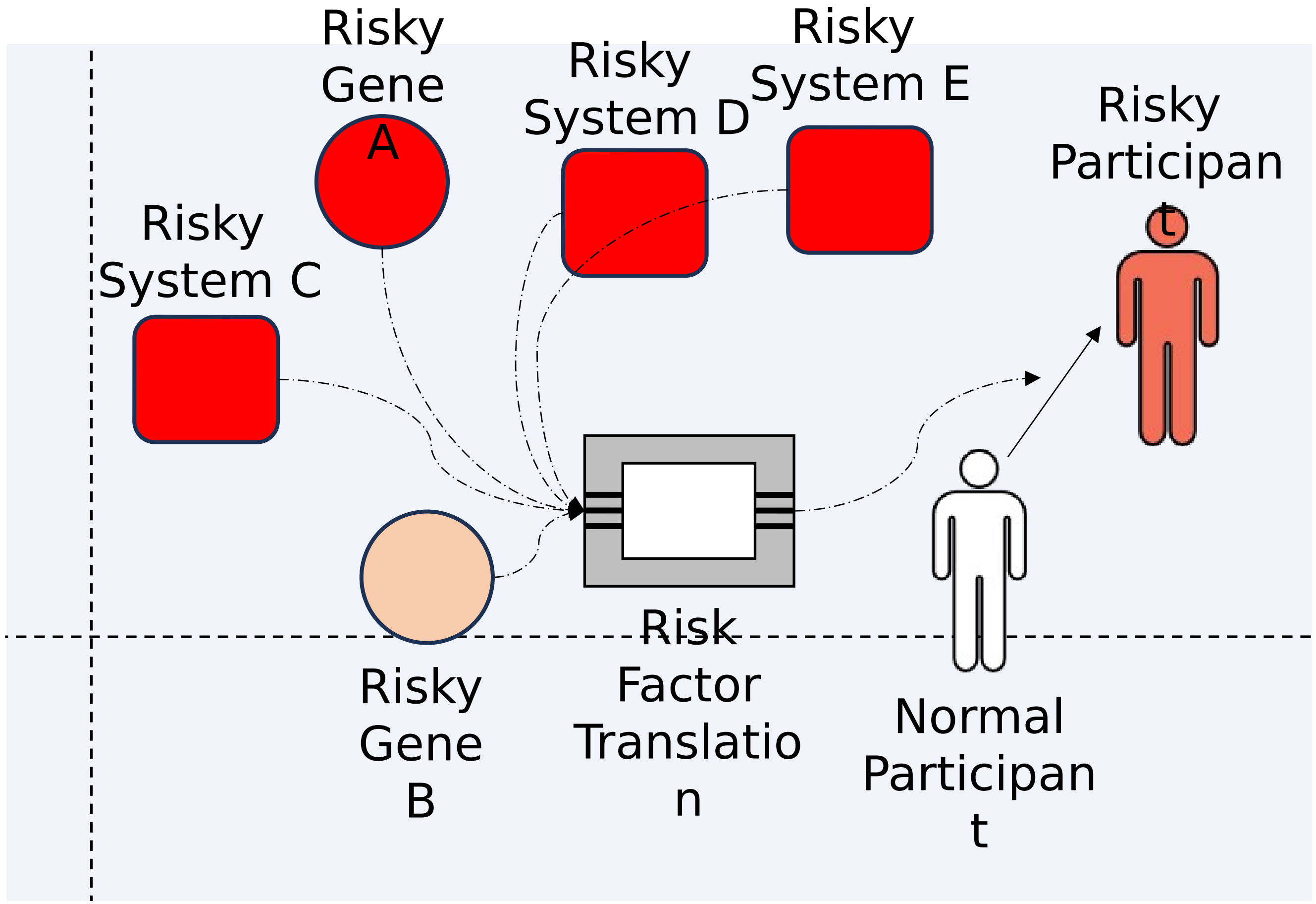
- Hierarchical structure-guided transformer
  - Propagating risk effect through biological hierarchy by **updating** state of connected entities

Propagation of Risk by Hierarchical Transformer



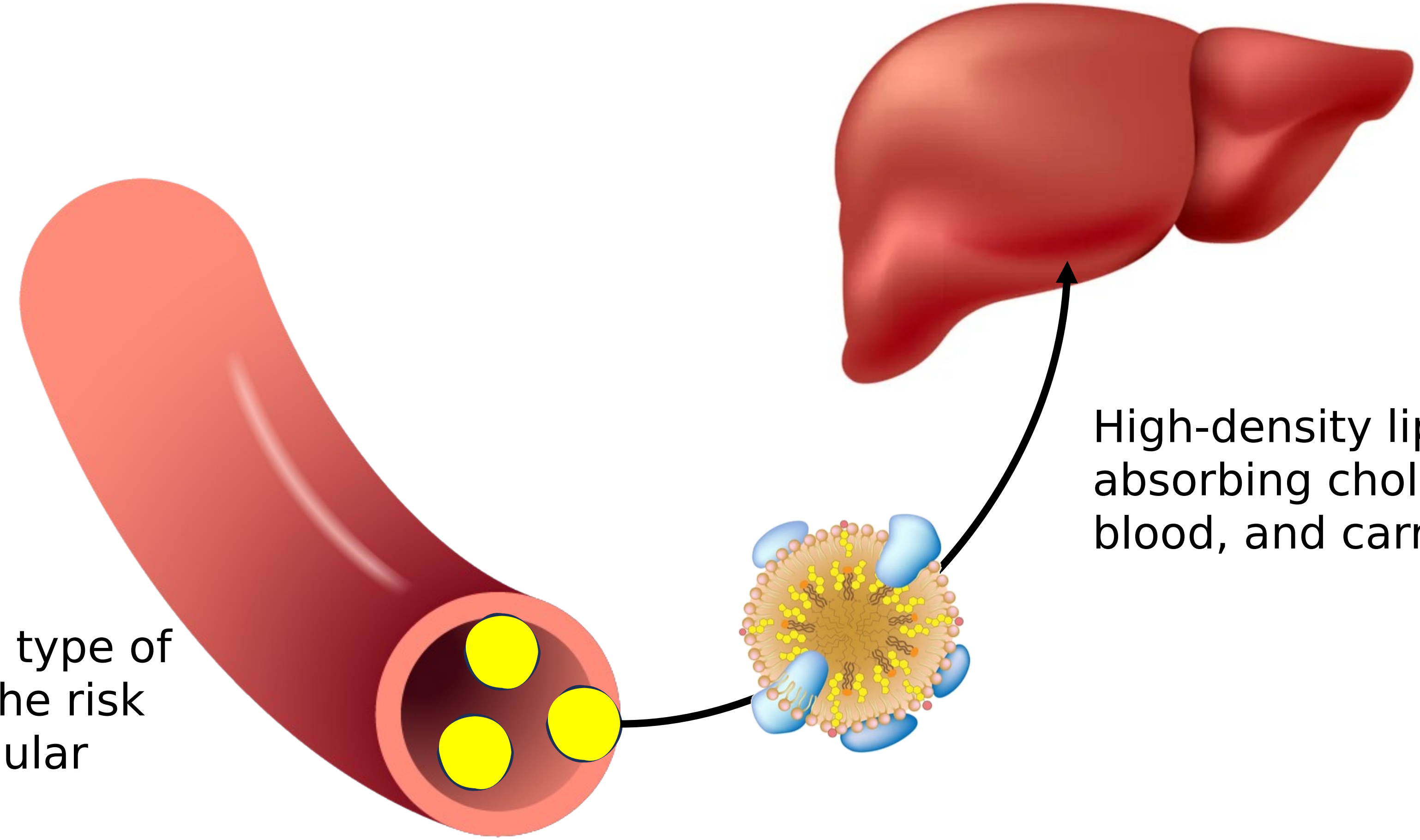
- Updating biological entities (genes, systems) through hierarchy
  - Propagating Risk to parent entities
  - Alterations of system and genes are represented as update of embeddings

Risk Factor Translation:
Method to Enhance Interpretability



- In the Risk Factor Translation phase:
    - Transformer with one head aggregates all altered states of systems and genes to predict phenotype
    - Attention mechanism prioritizes systems and genes

Phenotype In interest: TG/HDL ratio



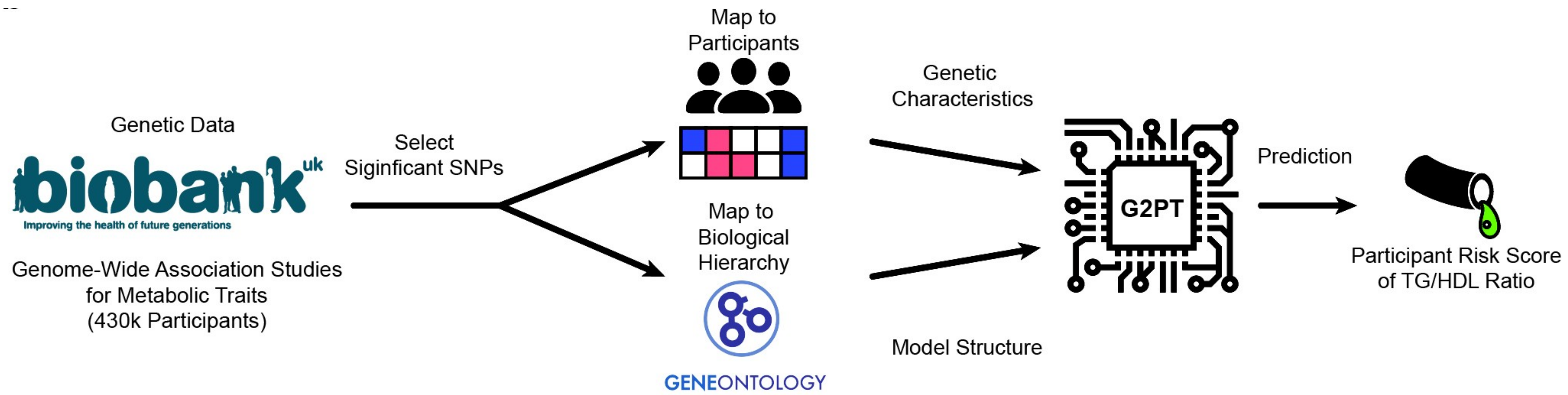
High-density lipoprotein: absorbing cholesterol from blood, and carries to the liver.

●

Triglyceride: a type of fat, increase the risk for cardiovascular disease

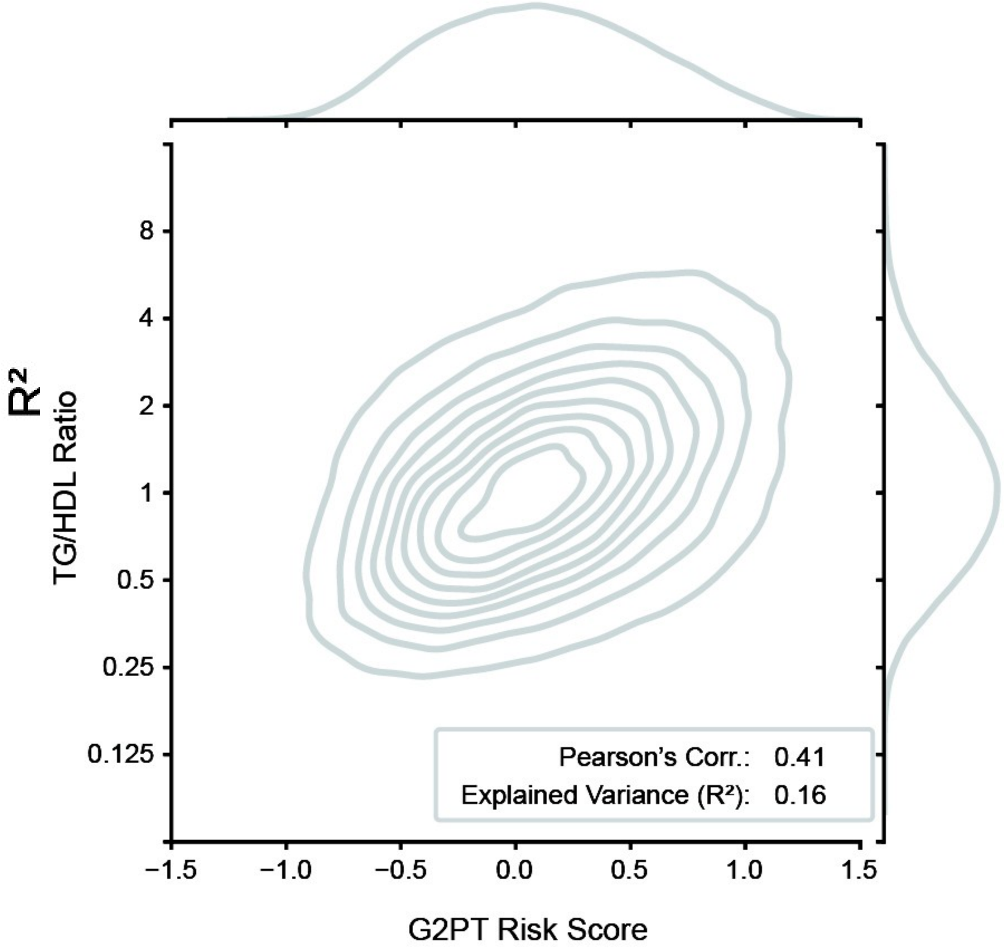
High TG/HDL ratio is risk factor for cardiovascular disease

G2PT pipeline

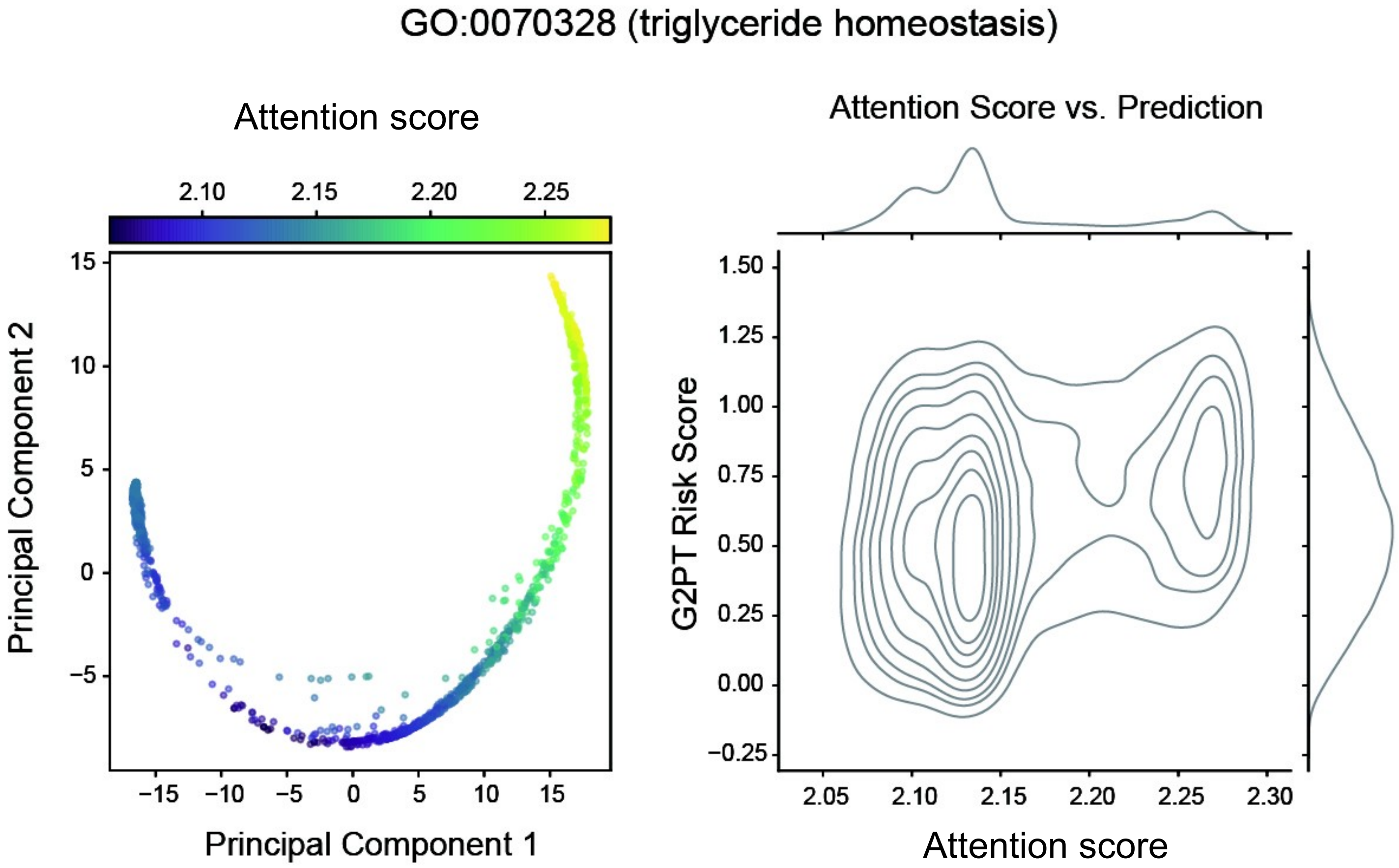


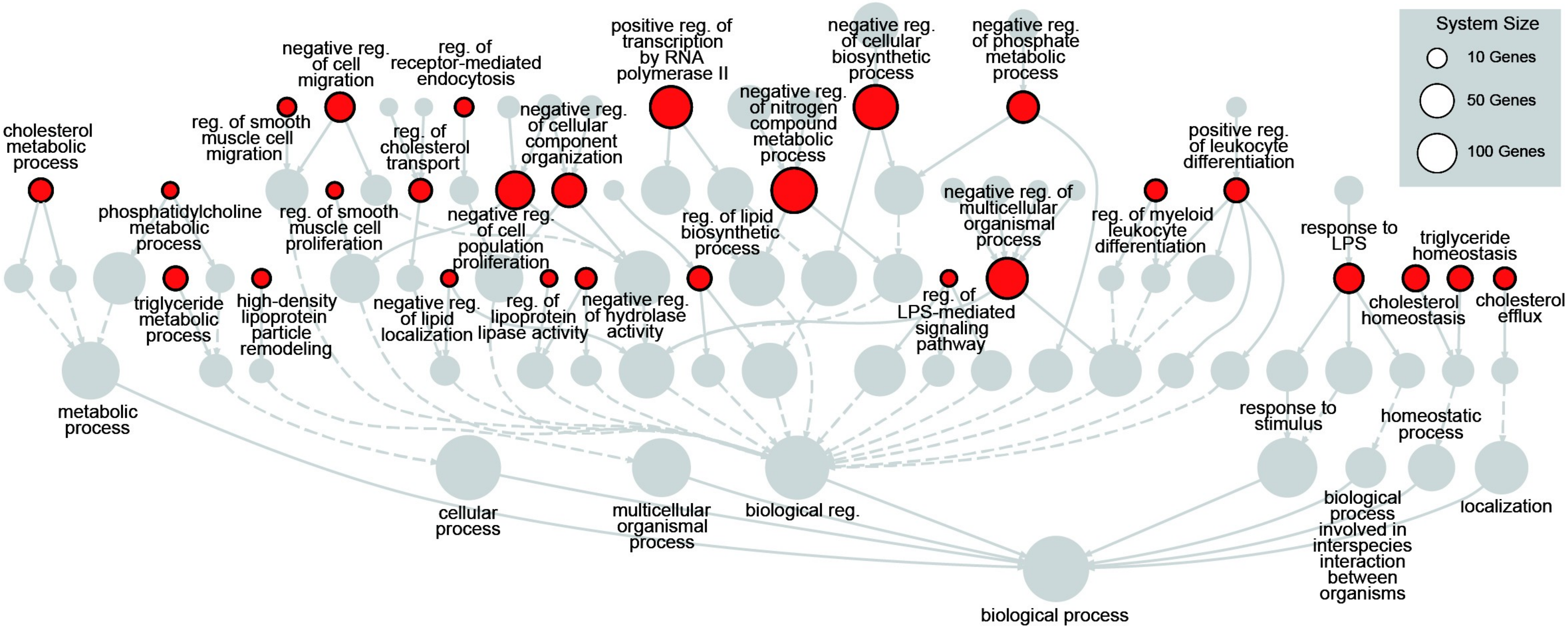
G2PT outperform previous GWAS model

- Explained Variance (R2) measures genetic heritability for TG/HDL ratio
- G2PT outperforms previous mixed Bayesian linear model (BOLT-LMM)



System Matters for Prediction!



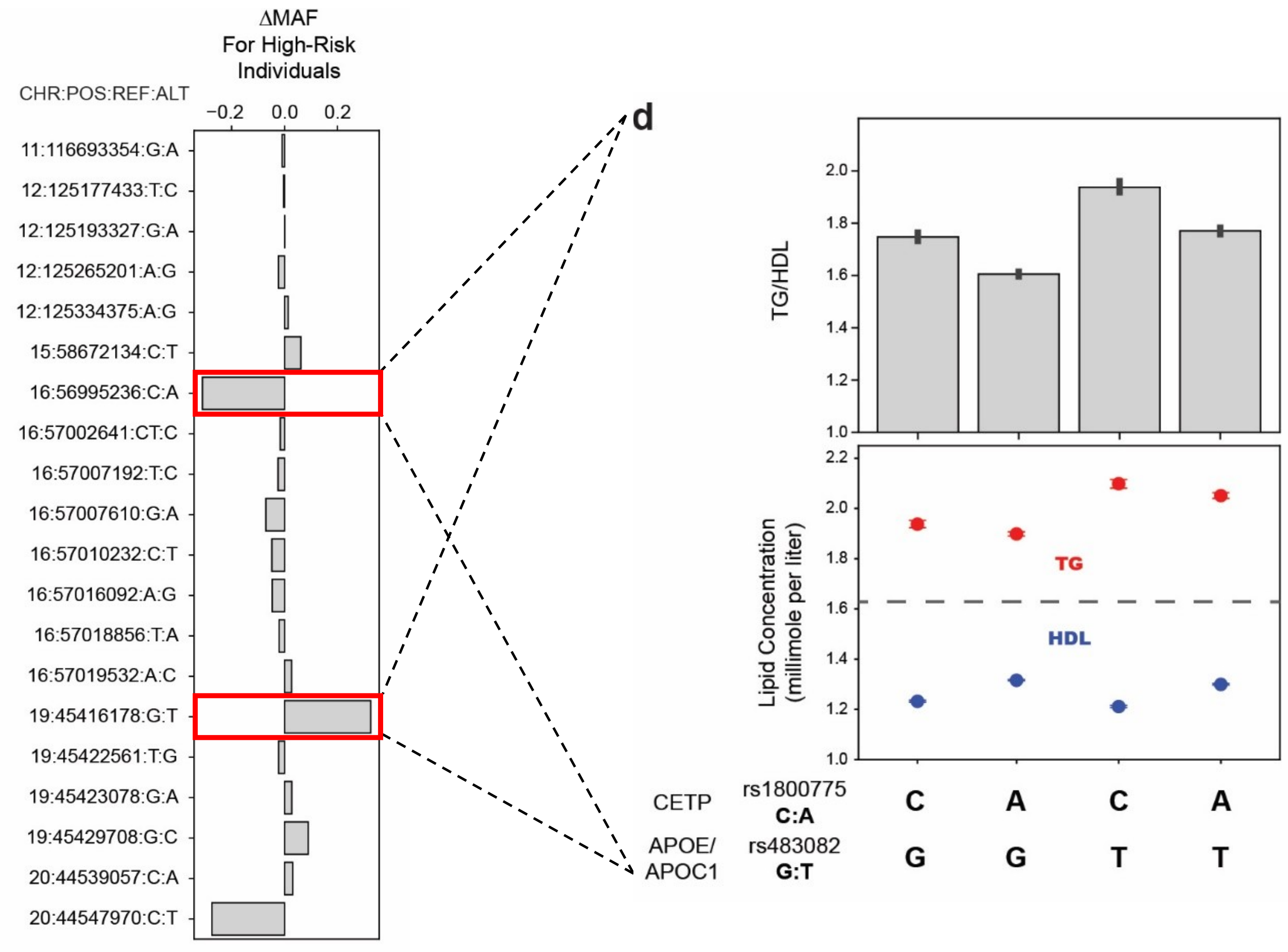

System Size
10 Genes
50 Genes
100 Genes
negative reg. of cell migration
reg. of receptor-mediated endocytosis
positive reg. of transcription by RNA polymerase II
negative reg. of cellular biosynthetic process
negative reg. of phosphate metabolic process
reg. of smooth muscle cell migration
reg. of cholesterol transport
negative reg. of cellular component organization
negative reg. of nitrogen compound metabolic process
cholesterol metabolic process
positive reg. of leukocyte differentiation
phosphatidylcholine metabolic process
reg. of smooth muscle cell proliferation
negative reg. of cell population proliferation
reg. of lipid biosynthetic process
negative reg. of multicellular organismal process
reg. of myeloid leukocyte differentiation
response to LPS
triglyceride homeostasis
triglyceride metabolic process
high-density lipoprotein particle remodeling
negative reg. of lipid localization
reg. of lipoprotein lipase activity
negative reg. of hydrolase activity
reg. of LPS-mediated signaling pathway
cholesterol homeostasis
cholesterol efflux
metabolic process
response to stimulus
homeostatic process
cellular process
multicellular organismal process
biological reg.
biological process involved in interspecies interaction between organisms
localization
biological process

Can G2PT detect genetic interaction?

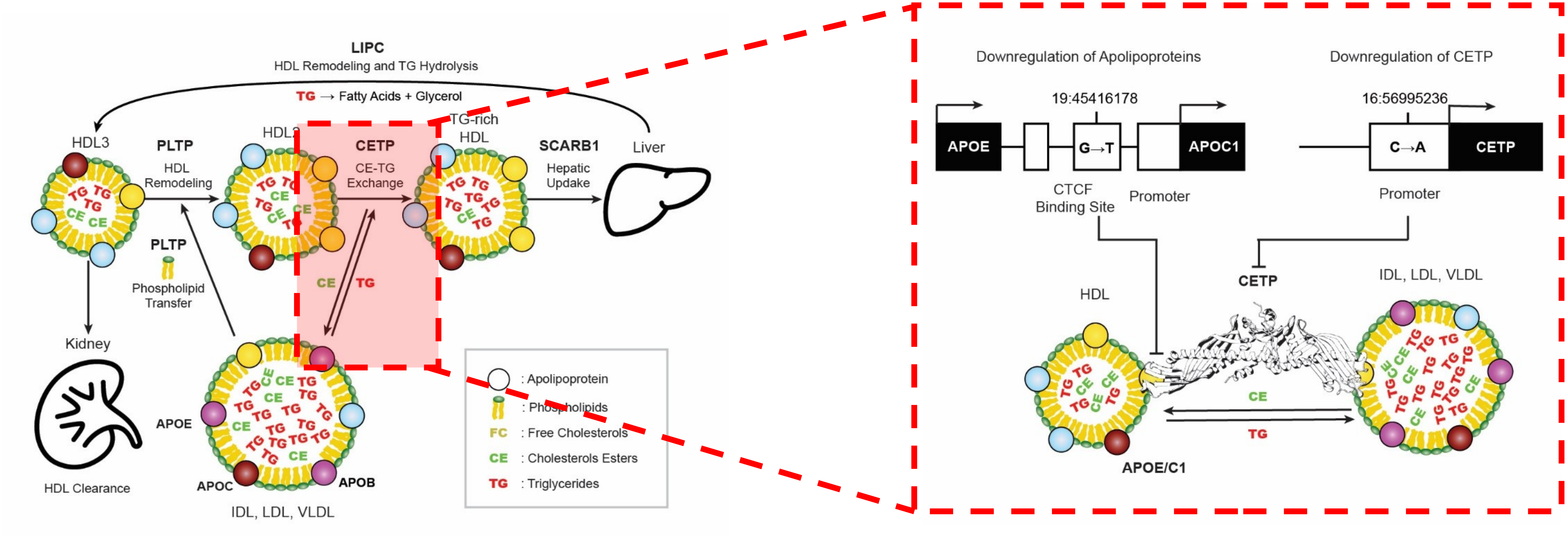
- Two SNPs in different genes (CETP, APOE/C1) show high MAF difference in high-risk subset
  - 16:56995236:C:A (Chi $p<10^{-101}$)
  - 19:45416127:G:T (Chi $p<10^{-188}$)
- Two SNPs show mutual exclusivity (Fisher’s exact $p<10^{-11}$)

- CETP and APOE/C1 get high attention from system

- SNP in CETP decrease TG/HDL level
- SNP in APOE/C1 increases TG/HDL level



Epistasis between CETP and APOE/C1



Acknowledgements

- Trey Ideker lab
  - Dr. Trey Ideker
  - Zach Wallace
  - Sarah Wright

- Amit Lab
  - Amit Majithia