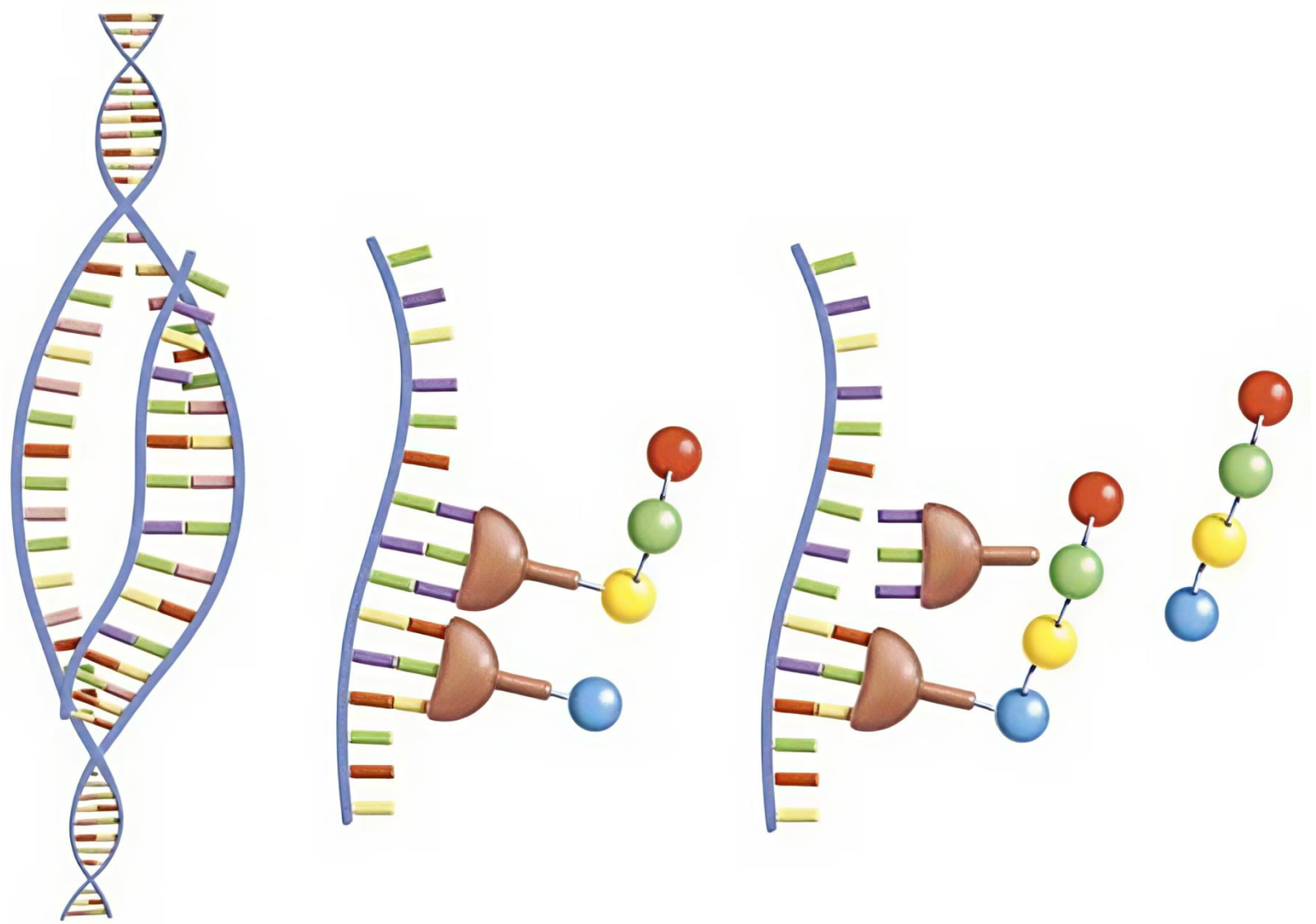
Modern statistics in medical and genetic research

Supervisors: Nemes András Botond, Dr. Németh László, Dr. Firneisz Gábor

Somogyi Dalma

2026. 06. 04.

Background



https://www.sciencesource.com/1241399-protein-synthesis-stock-image-rights-managed.html

Goal

- Epistasis detecting with interaction analysis
- Eventually we want to use it on database on Gestational Diabetes Mellitus

Methods

1. • Database on 4000 individuals' 3000 SNPs with 15 significant interactions
2. • Grouping SNPs into LD-blocks
3. • Representative vectors for each block
4. • Block-level interactions with representatives
5. • SNP-level interactions

Methods

- Dataset: $X \in \{0, 1, 2\}^{4000 \times 3000}$
- Target: $Y \in \{0, 1\}^{3000}$

LD-based blocks

- Creating blocks based on linkage disequilibrium (LD)

$$r^2 = \frac{(p_{AB} - p_A p_B)^2}{p_A(1-p_A)p_B(1-p_B)}$$

(for alleles A and B with allele frequencies p_A, p_B, p_{AB})

- The samples corresponding to the whole dataset is X, and to the jth block is X^j $(j=1, \ldots, m)$

Block representatives

1st method: Principal Component Ananlysis (PCA)

$$PC_1^j = \text{argmax}_{\|u\|=1}\left(\| X^j u \|_2^2 \right)$$

$$PC_k^j = \text{argmax}_{\substack{\|u\|=1, \\ u \perp \left\{PC_i^j\right\}_{i=1}^{k-1}}} \left(\| X^j u \|_2^2 \right)$$

$$k = 2, \ldots, 3000$$

Block representatives

2nd method: ElasticNet regression

$$\text{logit}(\mathbb{P}(Y = 1|X^j)) = X^j\beta^j$$

$$\widehat{\beta^j} = \text{argmin}_{\beta^j}\{ \| \text{logit}(\mathbb{P}(Y = 1|X^j)) - X^j\beta^j \|_2^2 + \alpha(\lambda \| \beta^j \|_1 + (1-\lambda) \| \beta^j \|_2^2) \}$$

Error

LASSO penalty

Ridge penalty

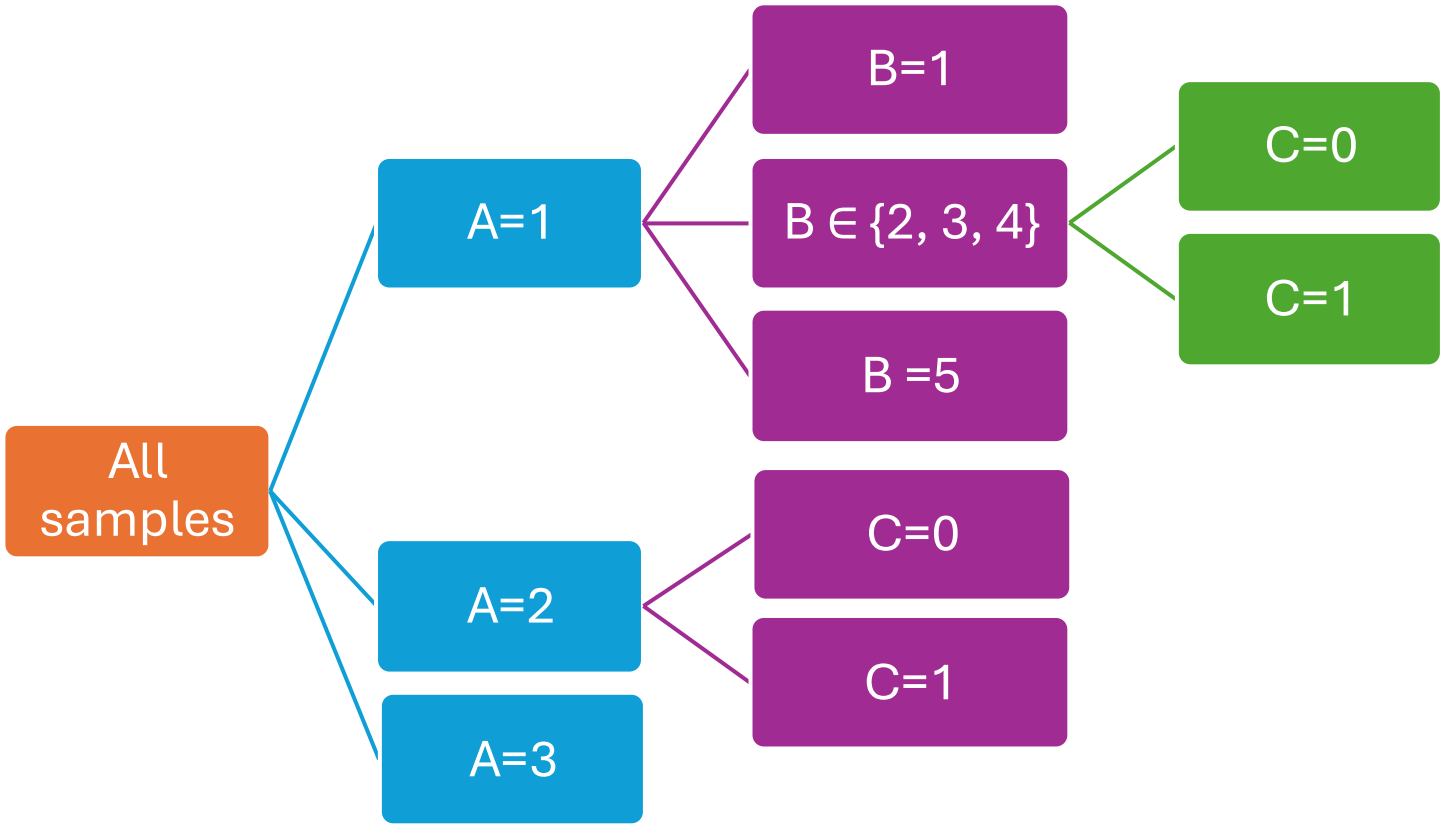
$$1 \geq \alpha \geq 0, \qquad 1 \geq \lambda \geq 0$$

Block representatives

3rd method: Decision tree based models (XGBoost, RandomForest)

Example:

- A ∈ {1, 2, 3}
- B ∈ {1, 2, 3, 4, 5}
- C ∈ {0, 1}
- Binary target



Block representatives

3rd method: Decision tree based models (XGBoost, RandomForest)

XGBoost

- Trains a decision tree
- Iterates: training a tree on the gradient of loss of the previous trees (uses logistic loss)

RandomForest:

- Multiple trees trained on random subsets of the variables
- Predictions averaged at the end

Block representatives

3rd method: Decision tree based models (XGBoost, RandomForest)

- Decrease of impurity upon splitting node p to children $\{c_i\}$ in tree t:

$$\Delta I(t,p) = I(p) - \sum_{i=1}^{k} \frac{|c_i|}{|p|} I(c_i)$$

- Importance weight for variable X_l with set $\mathcal{T}$ of trees used:

$$MDI(I, X_j) = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \sum_{\substack{p \in t \text{ is split} \\ \text{based on } X_l}} \Delta(t,p)$$

Block-level interactions

1st method: Logistic regression

- Block representative vector for the jth block: B^j
- Regressor matrix for the ith and jth blocks: $B^{ij} = \left[1 : B^i : B^j : B^i \times B^j\right]$

- Logistic regression:

$$\text{logit}(\mathbb{P}(Y = 1|B^{ij})) = B^{ij}\beta^{ij}$$

- Least squares estimate of the coefficients:

$$\widehat{\beta^{ij}} = \text{argmin}_{\beta^{ij}}\left\{ \| \text{logit}(\mathbb{P}(Y = 1|B^{ij})) - B^{ij}\beta^{ij} \|_2^2 \right\}$$

Block-level interactions

1st method: Logistic regression

- Correction: Benjamini-Hochberg for k independent tests
- P-values of the $\mathrm{k} = \binom{m}{2}$ tests in ascending order: $p_{l_1} \leq p_{l_2} \leq \cdots \leq p_{l_k}$

$$r = \mathrm{argmax}\left\{ j : p_{l_j} \leq \frac{j}{k}\alpha \right\}$$

- The corrected significance level: $\alpha_{BH} = \frac{r}{k}\alpha$

$$\Rightarrow \mathbb{E}\left(\frac{\#(\text{false discoveries})}{\#(\text{discoveries})}\right) \leq \alpha$$

- Significant interaction: p-value below the corrected α_{BH} for $\alpha = 0.05$

Block-level interactions

2nd method: XGBoost

- Fitting XGBoost on all the block representatives
- Significant: pair's interaction importance is among the top q%

SNP-level interactions

1st method: Logistic regression

- Whithin every significant block pairs
- Fitting logistic regression for every SNP pairs
- For significant block pair B^i, B^j and SNP pairs $S^r \in B^i, S^s \in B^j$:

$$S^{rs} = [1 : S^r : S^s : S^r \times S^s]$$

$$\text{logit}(\mathbb{P}(Y = 1|S^{rs})) = S^{rs}\beta^{rs}$$

$$\widehat{\beta^{rs}} = \text{argmin}_{\beta^{rs}}\{ \| \text{logit}(\mathbb{P}(Y = 1|S^{rs})) - S^{rs}\beta^{rs} \|_2^2 \}$$

SNP-level interactions

1st method: Logistic regression

- Correction: Bonferroni for k independent tests

$$\alpha_B = \frac{\alpha}{k}$$
$$\Longrightarrow \mathbb{P}(\ \#(\text{false negatives}) \geq 1\ ) \leq \alpha$$

- Estimation for number of independent tests:

$$\sum_{(B^i, B^j)\ \text{significant block pair}} Meff_i Meff_j, \qquad Meff_j = \sum_{\lambda \in (\text{eigenvalues of } B^j \text{correlation matrix})} \max(\lambda, 1)$$

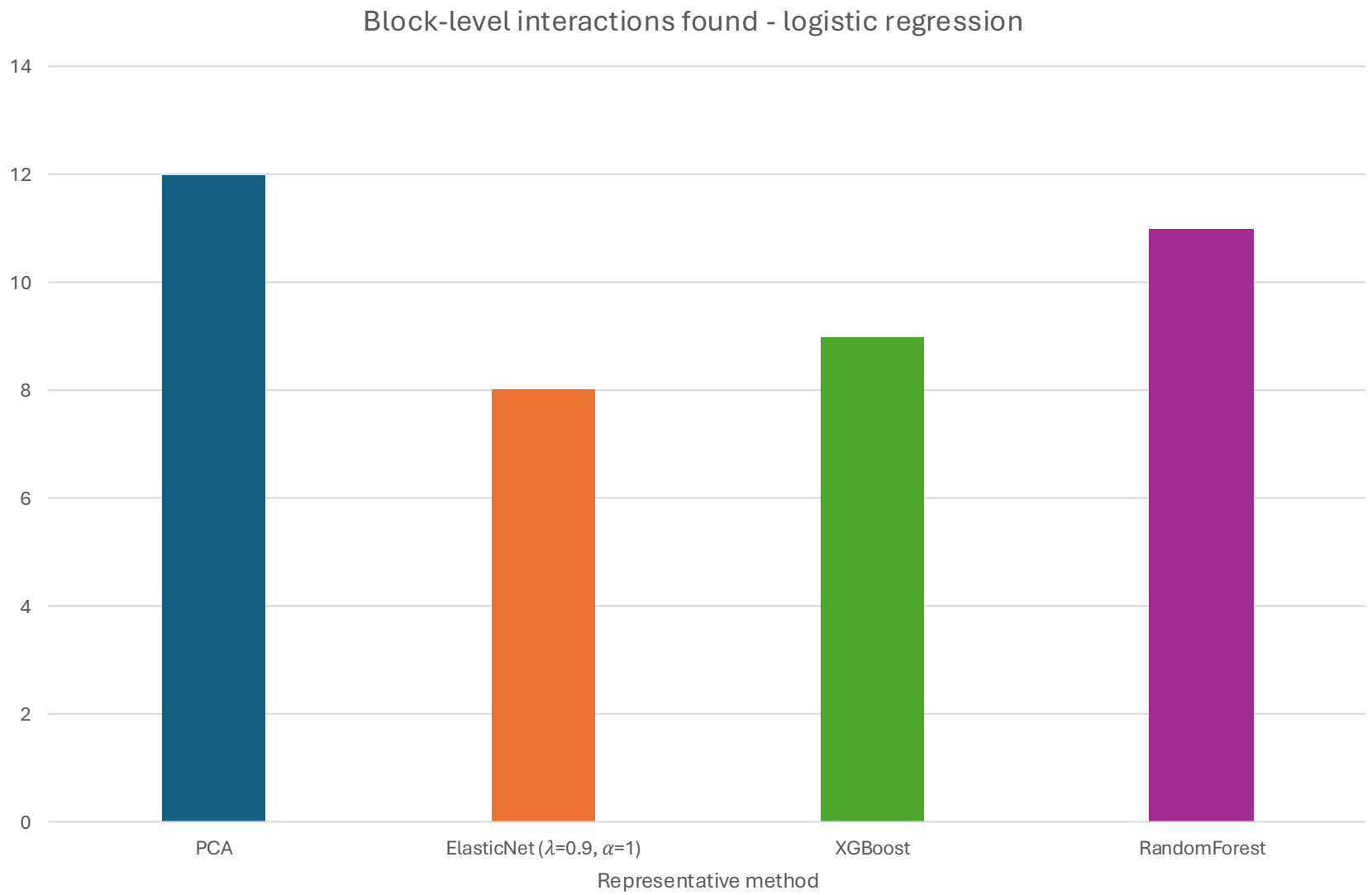
- Significant interaction: p-value below the corrected α_B for $\alpha = 0.05$

SNP-level interactions

2nd method: XGBoost

- For every significant block pair
- Fitting XGBoost on all the SNPs in the block pair
- Significant: pair’s interaction importance is among the top q%

Results


Block-level interactions found - logistic regression
14
12
10
8
6
4
2
0
PCA
ElasticNet (λ=0.9, α=1)
XGBoost
RandomForest
Representative method

References

[1] Molly Przeworski Augustine Kong Alexander I. Young, Stefania Benonisdottir. Deconstructing the sources of genotype-phenotype associations in humans. Science, 365(6460):1396–1400, 2019.

[2] Cooper-Knock J. Stamp J. et al Balvert, M. Considerations in the search for epistasis. Genome Biol, 25(296), 2024.

[3] Yingjie Guo et al. Gene-based testing of interactions using xgboost in genome-wide association studies. Front. Cell Dev. Biol., 9(16):8011–8013, 2021.

[4] McLean G.R. Franke A. Huang, J. wenty years of genome-wide association studies: Health translation challenges and ai opportunities. Eur J Hum Genet, 33:1579–1584, 2025.

[5] Trevor Hastie Hui Zou. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society Series B: Statistical Methodology, 67(2):301–320, 2005.

[6] Kamali Z. Xie T. et al. Keaton, J.M. Genome-wide analysis in over 1 million individuals of european ancestry yields improved polygenic risk scores for blood pressure traits. Nat Genet, 56:778–791, 2024.

[7] Ji L. Li, J. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. Heredity, 95:221–227, 2005.

[8] Zeggini E. McCarthy MI. Genome-wide association studies in type 2 diabetes. Curr Diab Rep., 9(2):164–171, 2009.

[9] Hadarits O, Harreiter J, Nádasdi A, Kelemen F, Bancher-Todesca D, Komlósi Z, Németh L, Rigó J Jr, Sziller I, Somogyi A, Kautzky-Willer A, Firneisz G, Rosta K, Al-Aissa Z. Association study with 77 snps confirms the robust role for the rs10830963/g of mtnr1b variant and identifies two novel associations in gestational diabetes mellitus development. 2017.

[10] B.S. Shastry. Snps: Impact on gene function and phenotype. Single Nucleotide Polymorphisms. Methods in Molecular Biology, 578, 2009.

[11] Montgomery Slatkin. Linkage disequilibrium–understanding the evolutionary past and mappingthe medical future. Nat Rev Genet, 9(6):477–485, 2008.

[12] Holzinger-E. Dasgupta A. et al. Szymczak, S. A new variable selection method for random forestsin genome-wide association studies. BioData Mining, 9(7), 2016.

[13] Joseph K. Pickrell Tomaz Berisa. Approximately independent linkage disequilibrium blocks inhuman populations. Bioinformatics, 32(2):283–285, 2016.

[14] Robert R H Anholt Trudy F C Mackay. Gregor mendel's legacy in quantitative genetics. PLoS Biol, 20(7), 2022.

[15] Sadeesh A. Srinivasasainagendra V. et al Vejandla, S.C. Calibrating genome wide significance byminor allele frequency across three major populations. Sci Rep, 15, 2025.

Nyilatkozat

A projektmunkám során irodalomkereséséhez és a modellek futtatásához használt Python kód megírásához használtam segítségül MI-t.