Attention Seeking in Genome-wide Association Studies

HEPING ZHANG

June 24, 2026

IMSI Workshop

Yale SCHOOL OF PUBLIC HEALTH

Disclaimer

The work presented here were not done to advance AI technologies or vice versa.

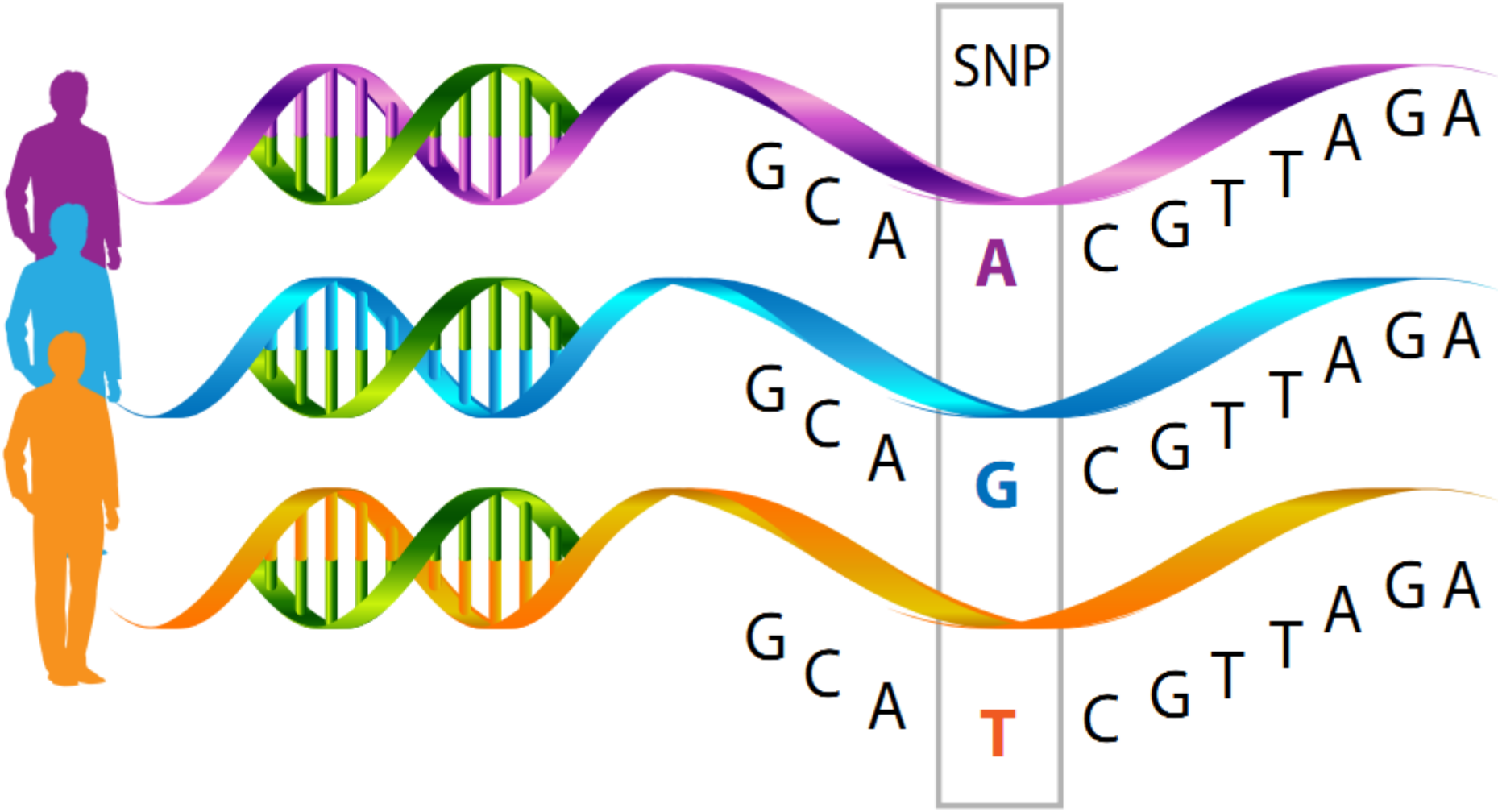
In retrospective, each may inspire the other.

My Observation

Outline

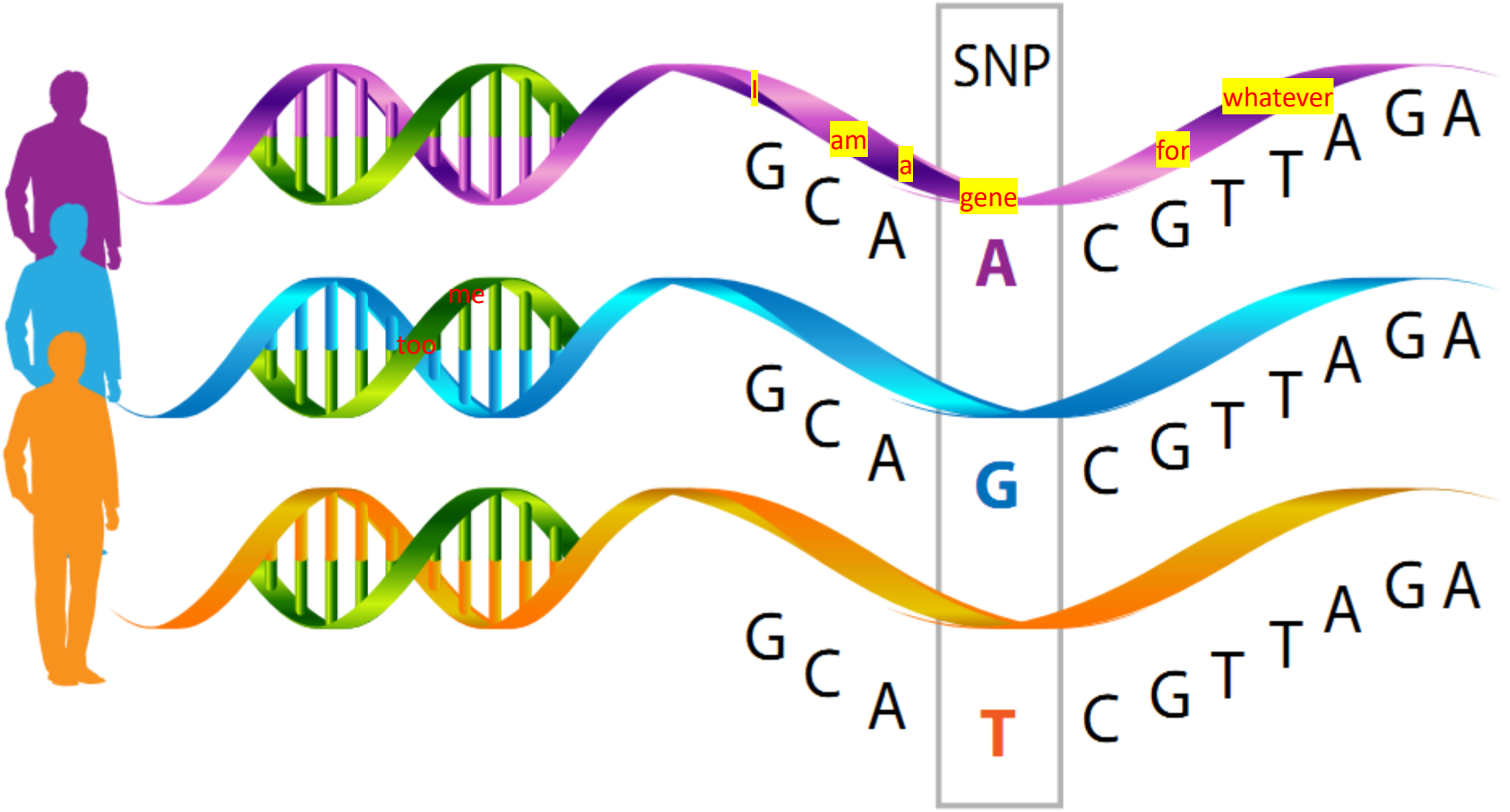
- Introduction to genome-wide association studies (GWAS)
- Connection between GWAS and language processing
  - Self-attention
  - Super-variant
  - Regional association score
- Concluding remarks

GWAS = Genomewide Association Studies



GWAS Catalog

GWAS vs Language Processing



GWAS vs Language Processing

- Looking for causal genes for complex diseases and understanding a language share a common principle but also differ in the context.
    - Both humans and machines can read not only long sentences, but also complicated books.
    - Humans' naked eyes and brain power cannot process the DNA sequences, but machines easily can.

GWAS vs Language Processing

Text data are readily available, and we understand them as we read even though we can have different understandings.

The DNA data are difficult to access, and we do not immediately see what problems they cause or what else is involved.

Looking for causal genes for complex is much more challenging than language processing

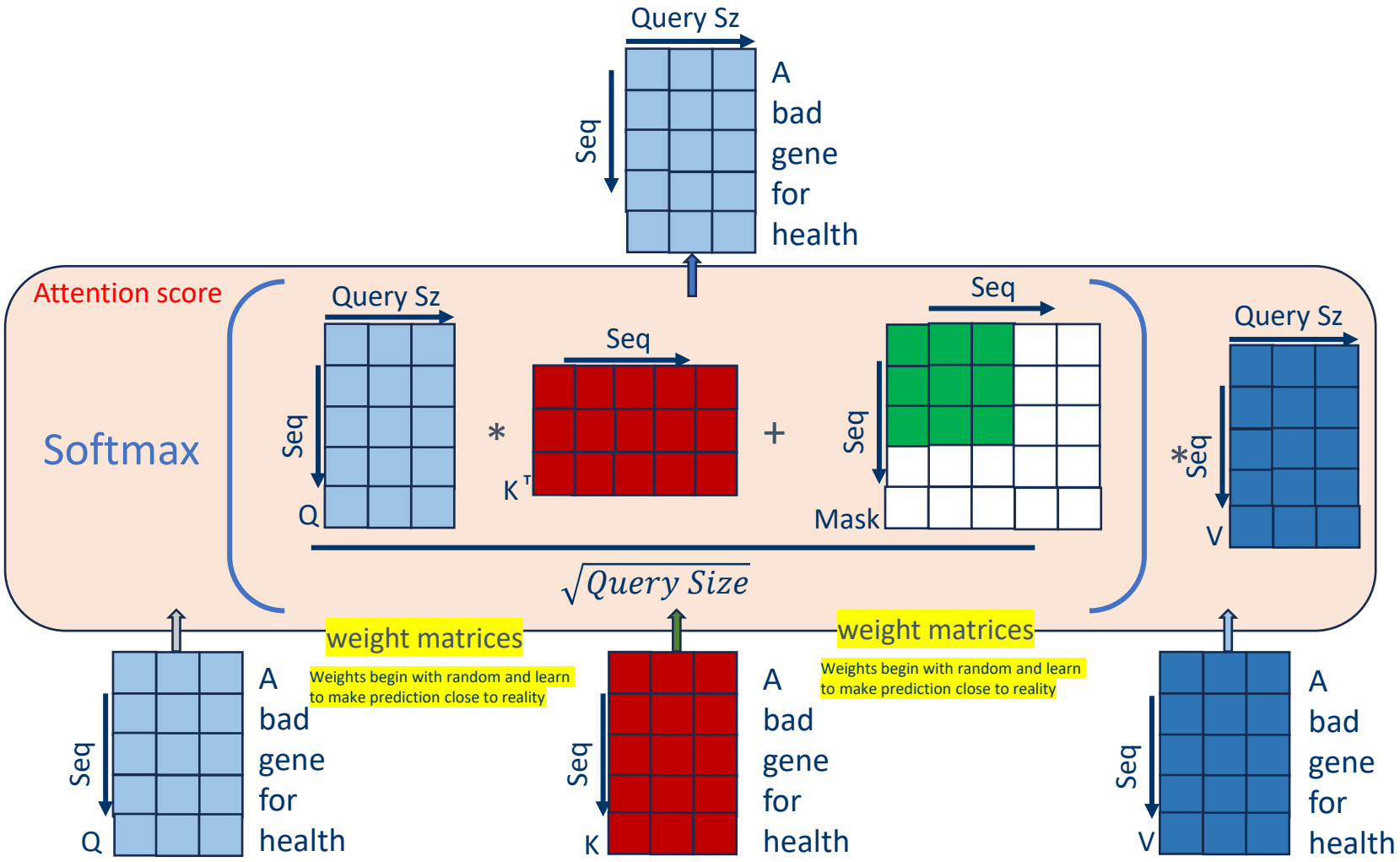
The Neural Revolution

"Attention Is All You Need" by Vaswani et al. in Neural Information Processing Systems 2017.

It transforms tokens into three distinct vector abstractions (query, key, and value) to calculate dynamic relationships among the tokens.

Query (Q): "Does this gene cause obesity?"
Key (K): "There is a list of bad actors for obesity."
Value (V): "Evidence isn't clear on the gene."

Self Attention



Lesson from the Self-Attention

A crucial prerequisite of self-attention is the projection of variable input representations into a common fixed-dimensional latent space.

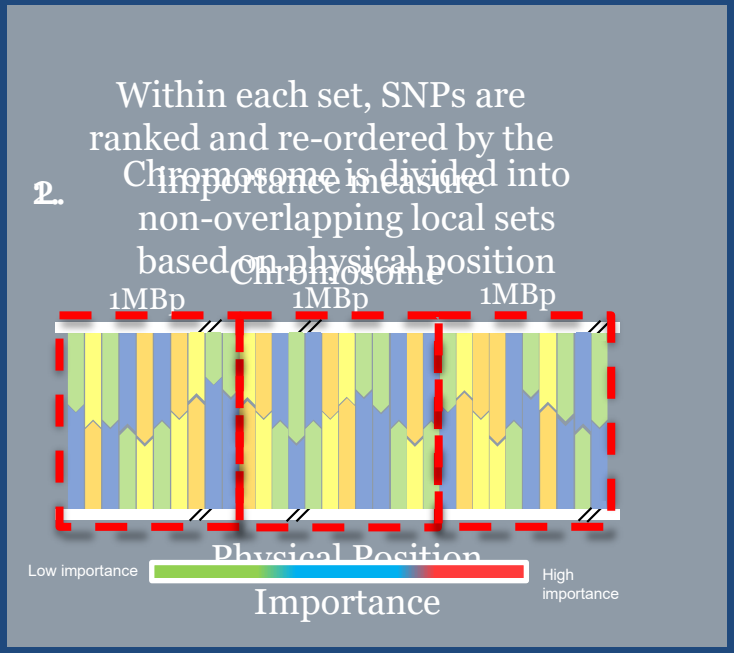
This uniform representation makes it possible to compute attention scores between tokens and to dynamically aggregate information according to their inferred relevance.

This technique is also useful in analyzing GWAS data.

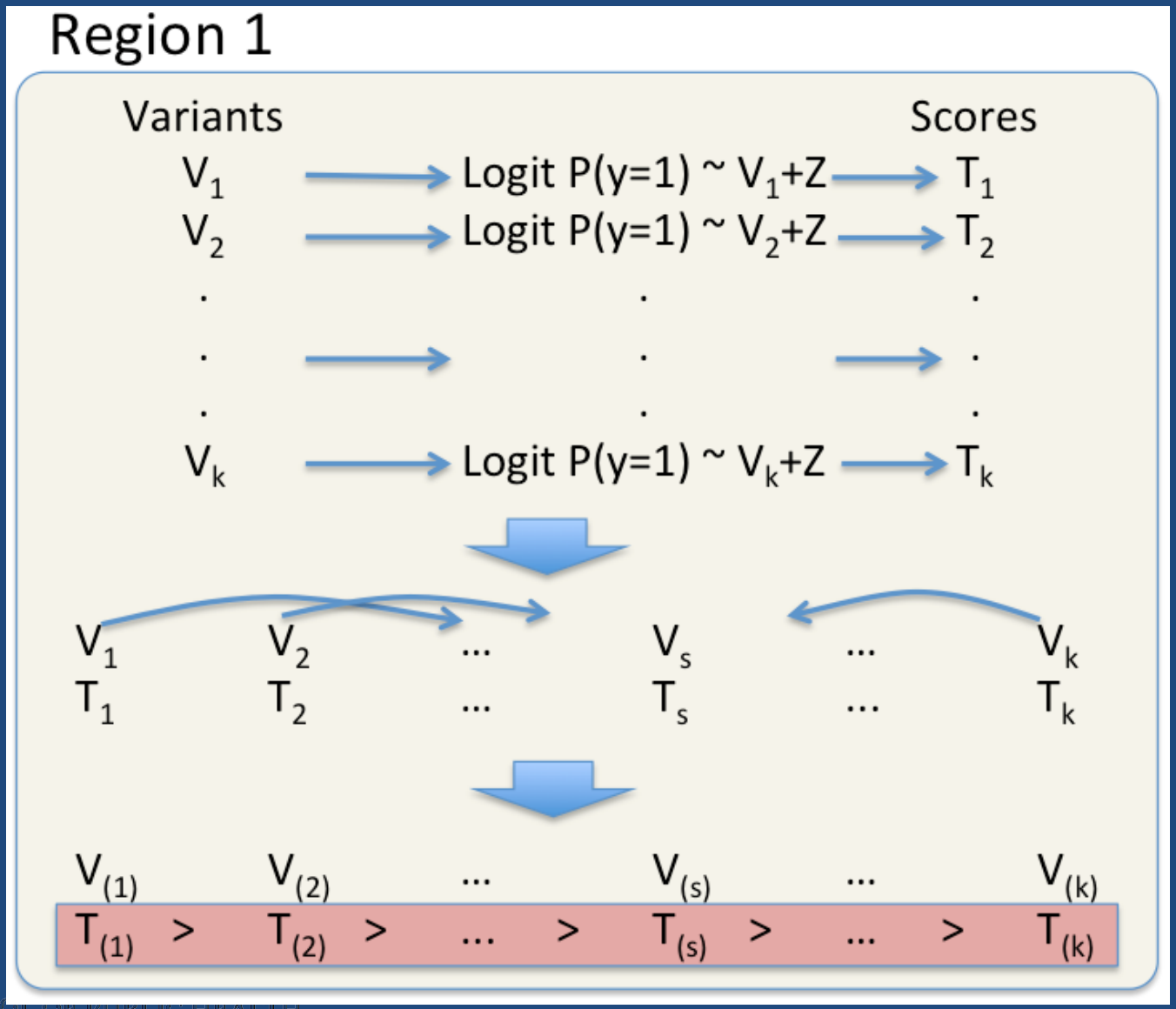
Super-variant (Song and Zhang, 2014)

The number of SNPs and windows vary like the length of a sentence.

SNPs within a window is assessed, like relationships among tokens are measured.



Data Transformation: Embedding



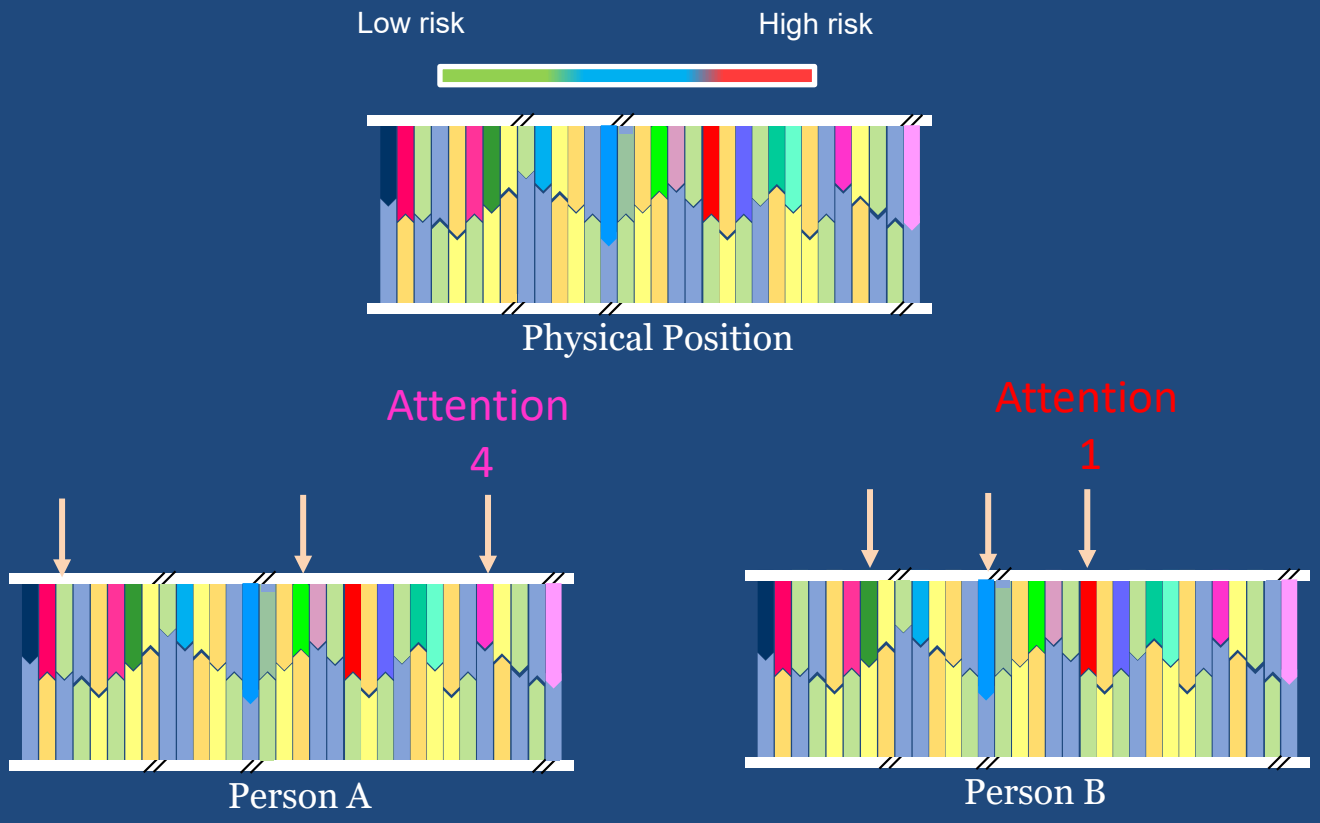
Composite Variants: Extracted Features

❖**Definition:** $X_{i1} = \min\{s: V_{i(s)} = 1\}$

❖**Interpretation:**

- After ranking the significance of all variants within a subset, sample i has its first minor allele at the s-th rare variant.
- $X_{i1} > 0$ if sample i has a minor allele in any rare variant in the subset — a generalization of the indicator variable used in existing methods.
- X_{i1} also reflects the relative strength of the association.

Super-variant Connection to Attention Seeking



Composite Variants: Extracted Features

- X_{i2} can be defined analogously by the ascending T values.
- X_{i1} and X_{i2} play the same role as the extracted features in self-attention.

SAGE Data and Findings

Study of Addiction: Genetics and Environment (SAGE) data.

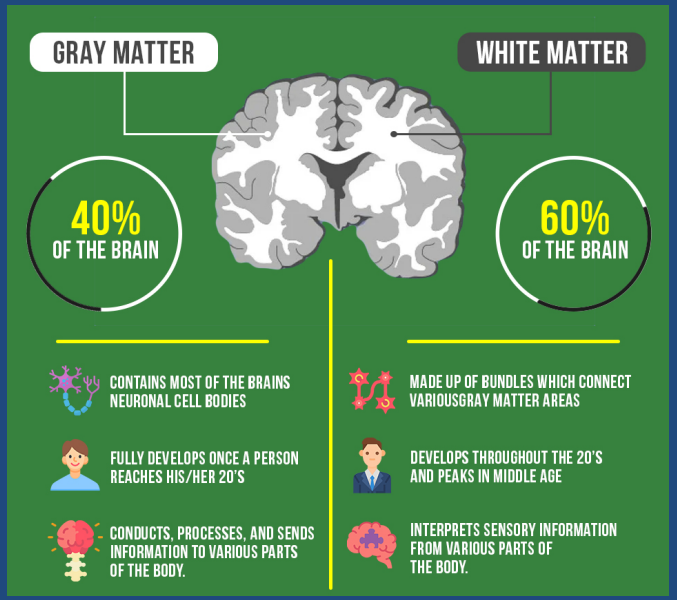
- 1151 cases of alcohol addiction, and 1336 controls.
- 43 variants in CTNNA2 are found to increase the risk of alcohol addiction in female, with an odds ratio (OR) of 1.94.
- Literature search found that CTNNA2 is very likely a true finding on the basis of both its function and existing studies.

The Variants Identified in CTNNA2

Chromosome	Position	Alteration	Frequency	rsSNP ID
2	79414140	G→C	0.0325282431	
2	79440921	C→T	0.0323334632	
2	79583819	C→T	0.0313595637	
2	79618395	G→A	0.0268796260	
2	79678697	A→G	0.0054538372	
2	79702781	T→C	0.0019477990	
2	79703081	A→T	0.0019477990	
2	79704003	A→G	0.0019477990	
2	79711967	T→G	0.0019477990	
2	79720449	A→G	0.0377873004	
2	79813928	G→A	0.0089598753	*rs11899508*
2	79814436	G→A	0.0015582392	
2	79828985	A→C	0.0009738995	
2	79854979	G→A	0.0072068563	*rs11900109*
2	...	...	...	...

Genetic Variants for Brain Connectivity

- Gray and white matters of human brain provide insights into the structural and functional abnormalities associated with disorders.
  - Gray matter (outer surface): processing and integrating information
  - White matter (deep parts): essential for the transmission of information and coordination of brain functions



Source: https://www.spinalcord.com/blog/gray-matter-vs-white-matter-in-the-brain

Functional magnetic resonance imaging

Functional magnetic resonance imaging (fMRI): brain activity through changes in blood (BOLD signals).

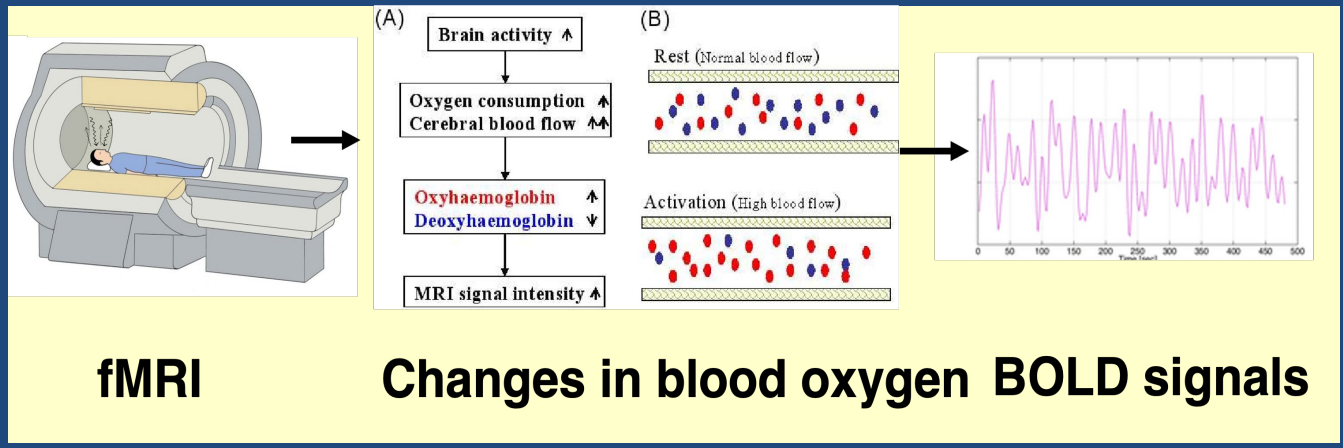


Figure source: https://www.youtube.com/watch?v=rJjHjnzmvDI

Functional Connectivity

FC: Correlations of BOLD signals between two regions of the brain.

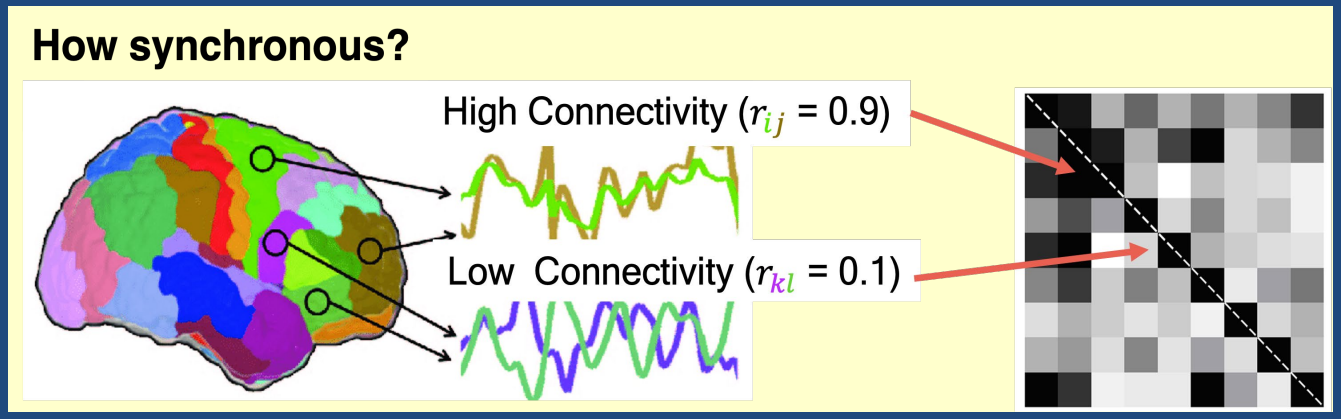
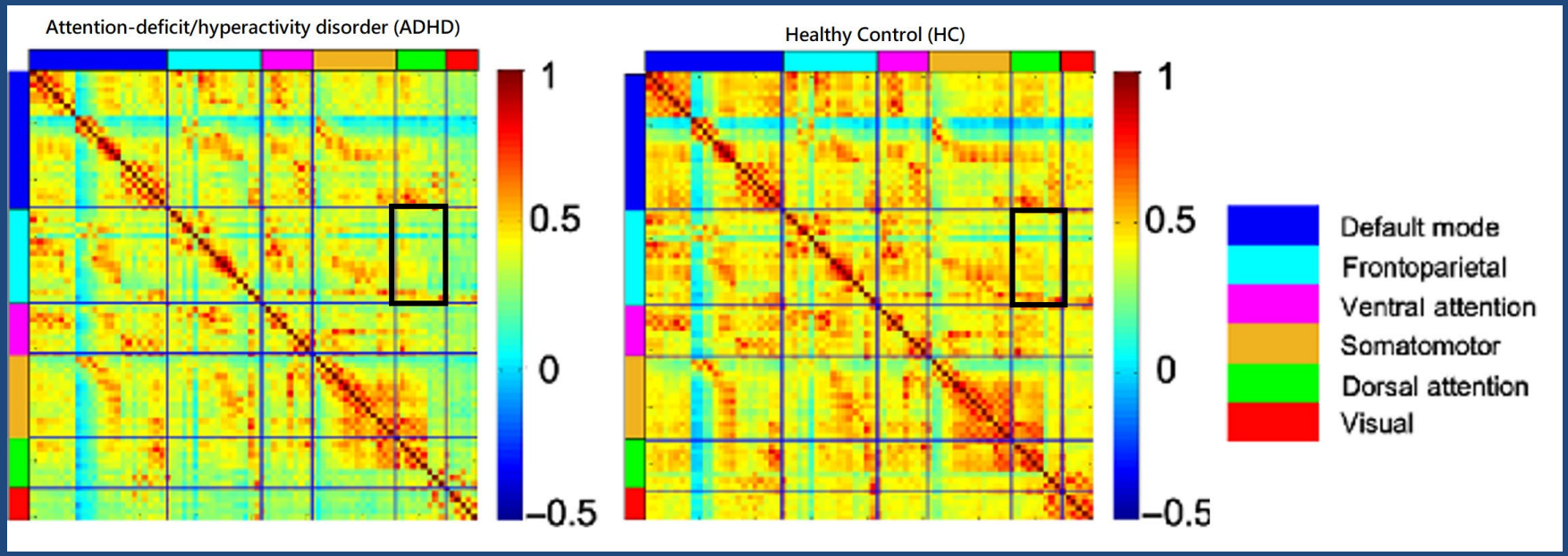


Figure: Functional connectivity, correlations of BOLD signals, Scheinost et al. (2016)

Functional Connectivity



FC alternations in patients with ADHD and healthy controls (Yadav et al. 2021)

Analysis of UK Biobank

- UK Biobank Dataset: ~40,000 UK participants with SNPs and resting-state functional connectivity with dimension of 55*55.
- **Primary aims:**
  - **Identify SNPs that are associated with resting-state functional connectivity**
- 41,502,298 SNPs divided into 2,723 local genetic SNP-sets.
- Discovery set: UKB subjects with White British ancestry ⇒ 32,262 subjects.
- Validation set: White but non-British subjects (n = 1867).
- Significance level = 1.84×10^{-5} (i.e., 0.05/2723) for discovery set; 0.05/number of significant SNP-sets for validation set

SNPs Associated with Resting-state FC

- 10 verified super-variants, including 29 SNPs.
- For super-variant chr5 23, 10 SNPs in the region 5p14.3 and gene CDH12 were found to be associated with bipolar disorder in a previous study.
- For SNPs in the super-variant chr22 39, 3 SNPs in the region 22q13.1 were found to be associated with attention deficit disorder with hyperactivity (ADHD).

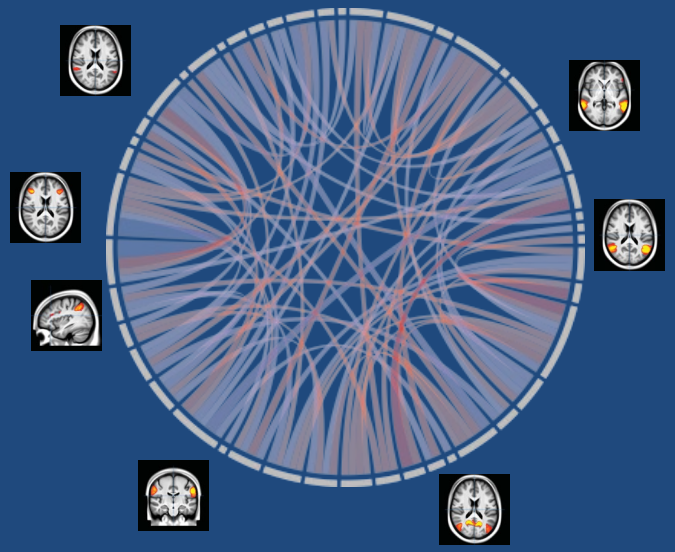


Figure: The different brain connectivity of the super-variant on chromosome 1 set 149 between major and the minor group. For visualization, differences with absolute values in top 5% are plotted. Red (blue) bands indicate the positive (negative) differences, and the widths of the bands indicate the magnitudes of the differences. We provided the axial view of the brain for several that have stronger differences in connectivity.

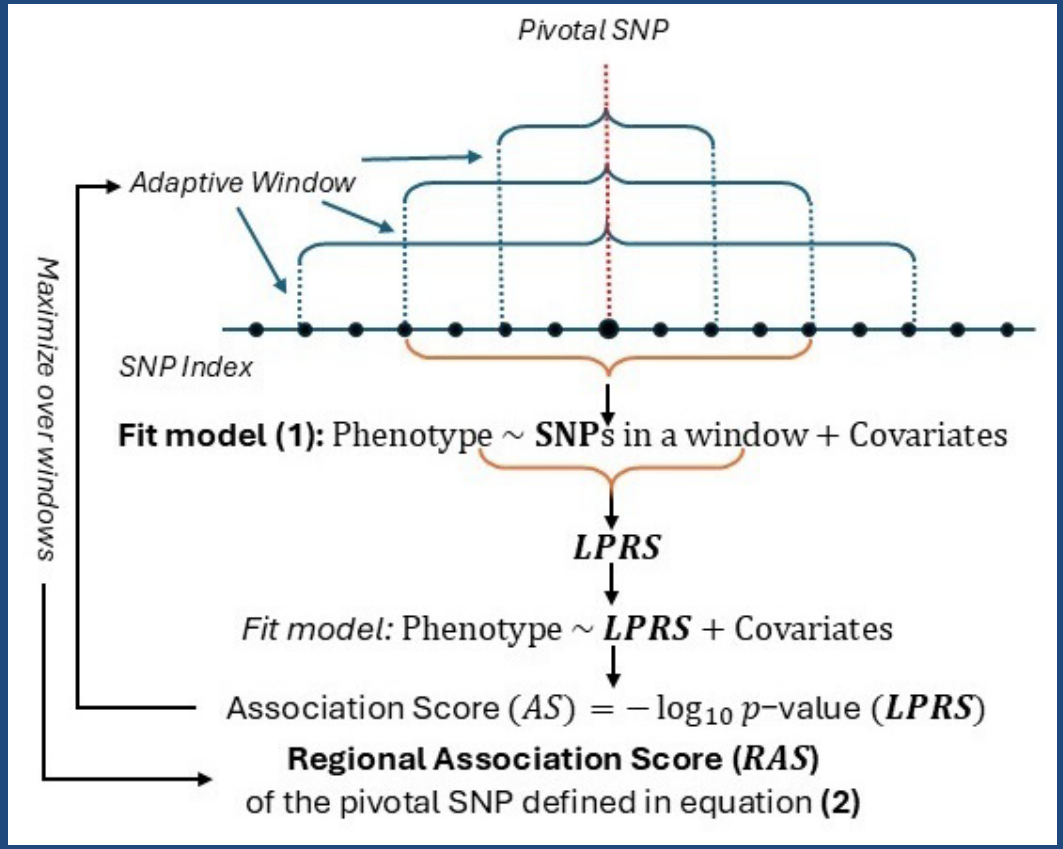
Regional Association Score (RAS)

The goal is to enhance true signals that may not be strong enough in themselves, while suppressing noise.

Jiang and Zhang, 2025, PNAS

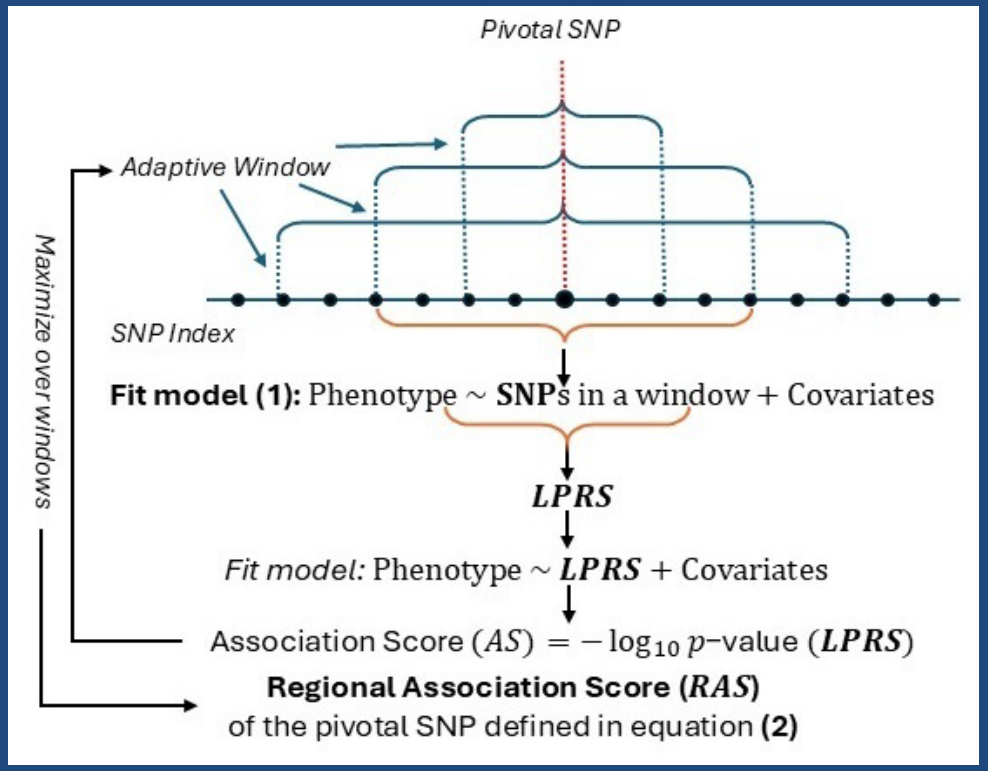
Regional Association Score (RAS)

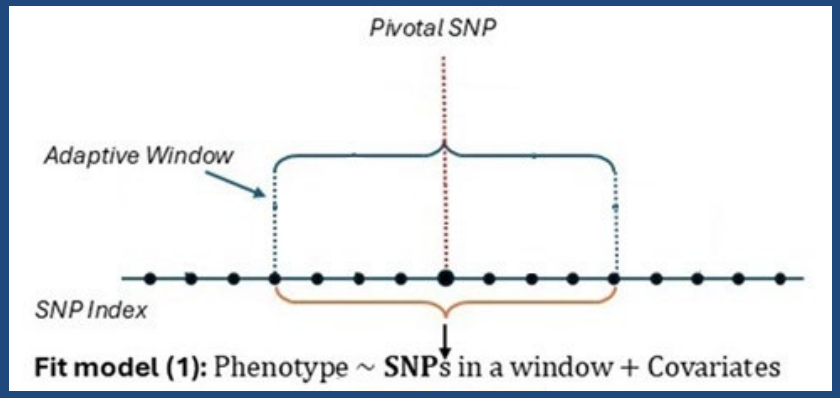
1. For each *pivotal SNP* (selected to span the whole chromosome), calculate a RAS that aggregates association signals from an adaptive window of nearby variants.

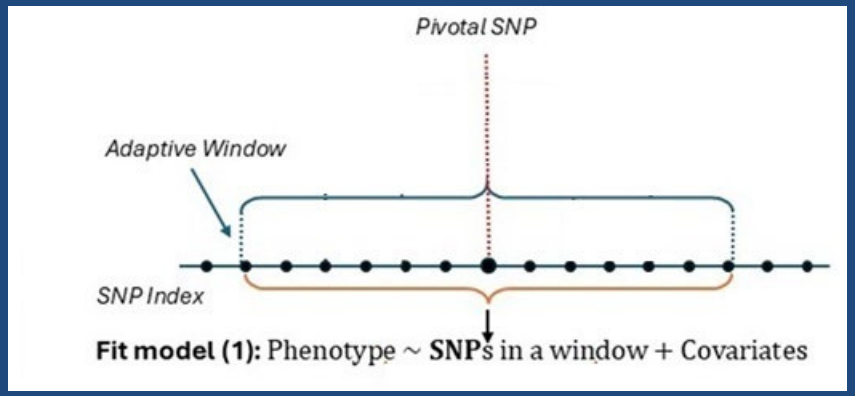


Jiang and Zhang, 2025, PNAS

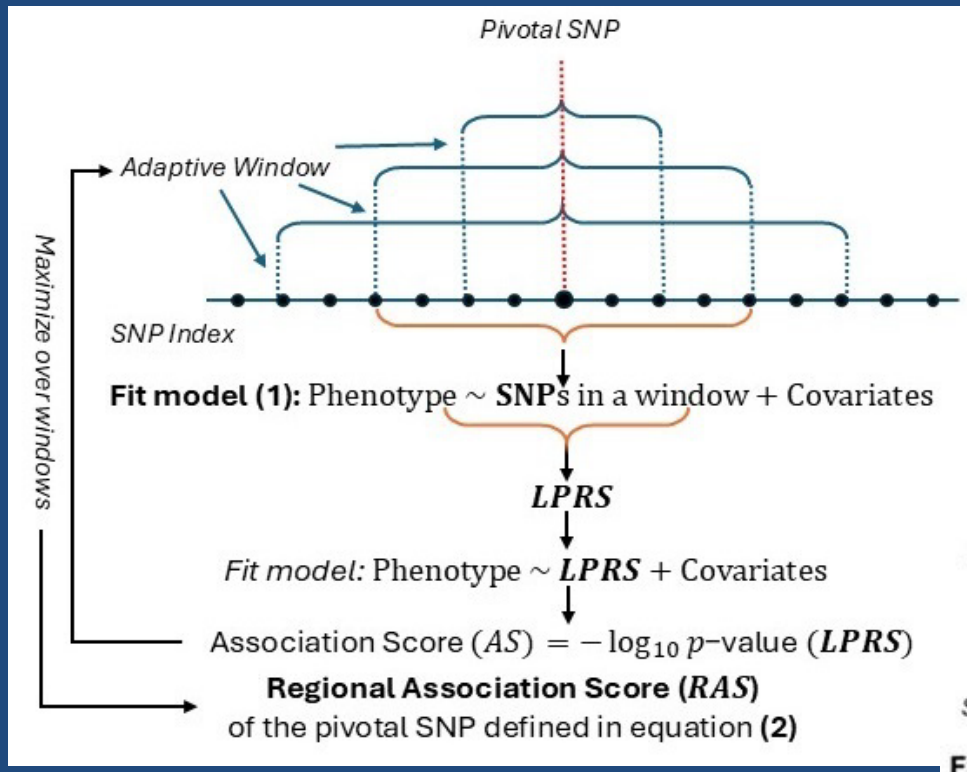
RAS

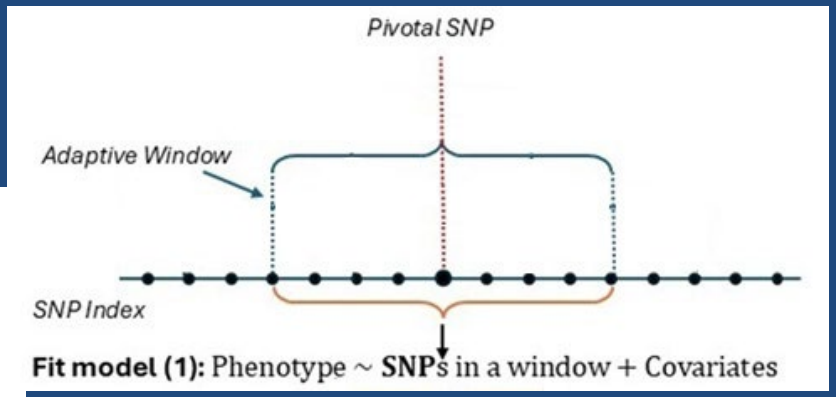

Pivotal SNP
Adaptive Window
SNP Index
Maximize over windows
Fit model (1): Phenotype ~ SNPs in a window + Covariates
LPRS
Fit model: Phenotype ~ LPRS + Covariates
Association Score (AS) = −log10 p-value (LPRS)
Regional Association Score (RAS)
of the pivotal SNP defined in equation (2)

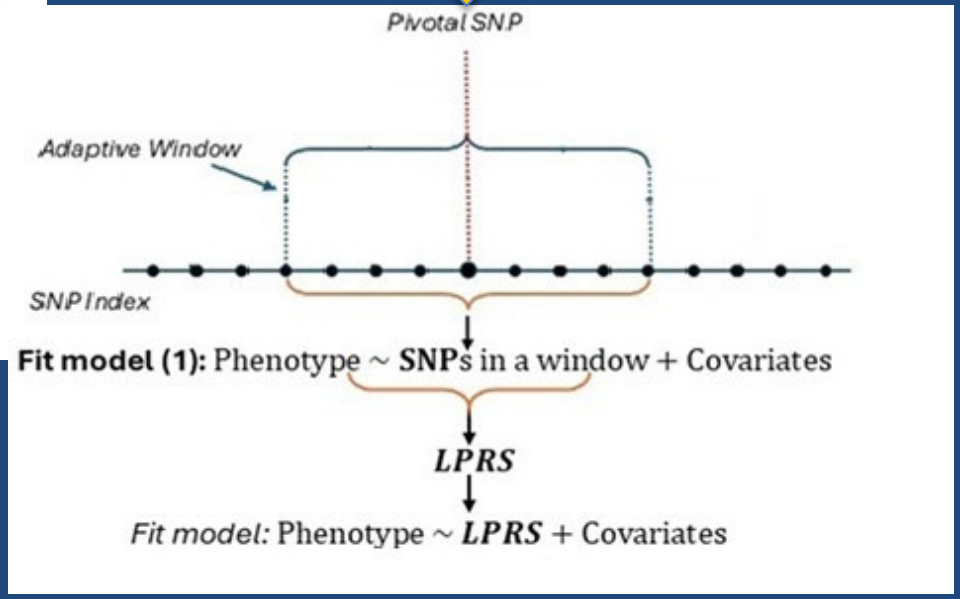

Pivotal SNP
Adaptive Window
SNP Index
Fit model (1): Phenotype ~ SNPs in a window + Covariates


Pivotal SNP
Adaptive Window
SNP Index
Fit model (1): Phenotype ~ SNPs in a window + Covariates

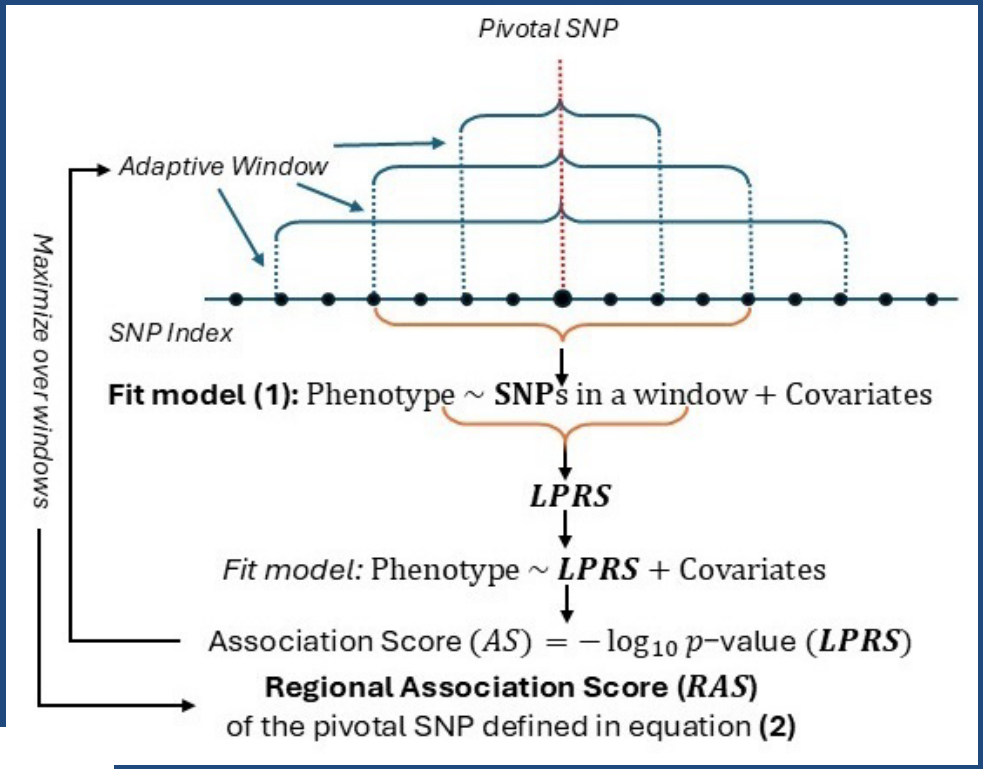
RAS

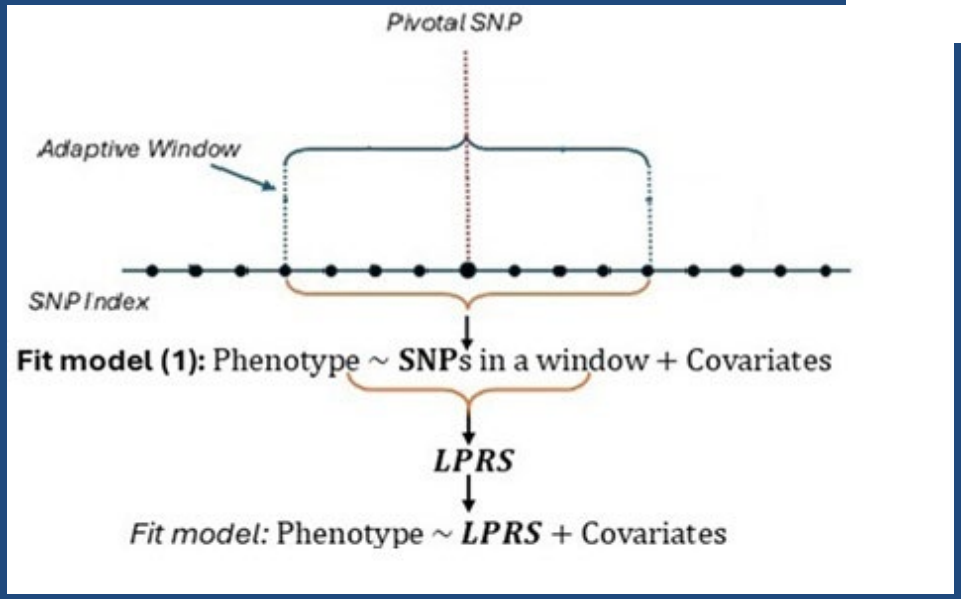

Pivotal SNP
Adaptive Window
SNP Index
Fit model (1): Phenotype ~ SNPs in a window + Covariates
LPRS
Fit model: Phenotype ~ LPRS + Covariates
Association Score (AS) = −log10 p−value (LPRS)
Regional Association Score (RAS)
of the pivotal SNP defined in equation (2)
Maximize over windows


Pivotal SNP
Adaptive Window
SNP Index
Fit model (1): Phenotype ~ SNPs in a window + Covariates


Pivotal SNP
Adaptive Window
SNP Index
Fit model (1): Phenotype ~ SNPs in a window + Covariates
LPRS
Fit model: Phenotype ~ LPRS + Covariates

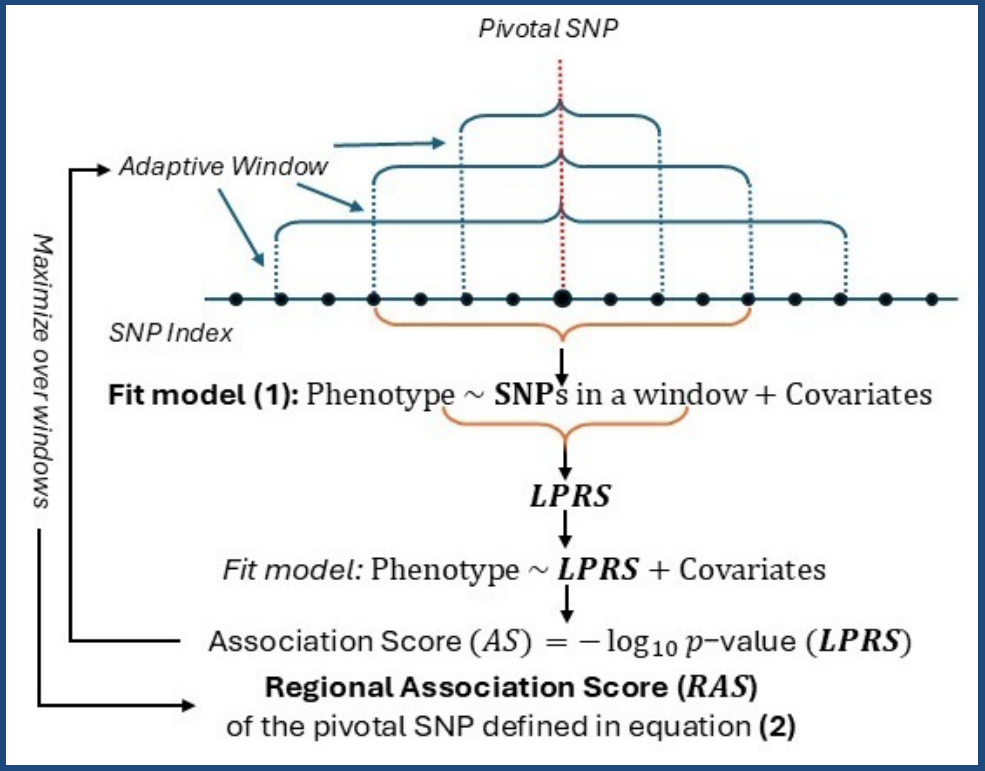
RAS

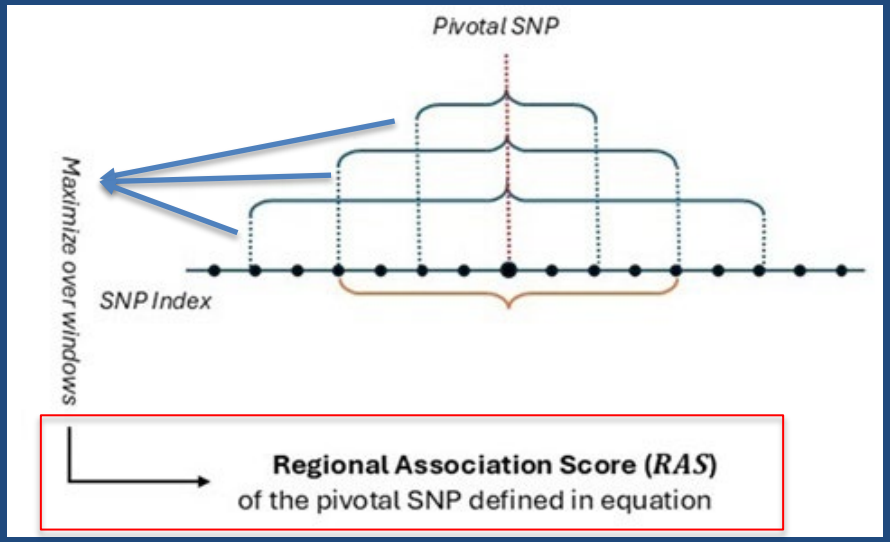




Association Score (AS) is calculated as $-\log_{10}$ p-value(LPRS)

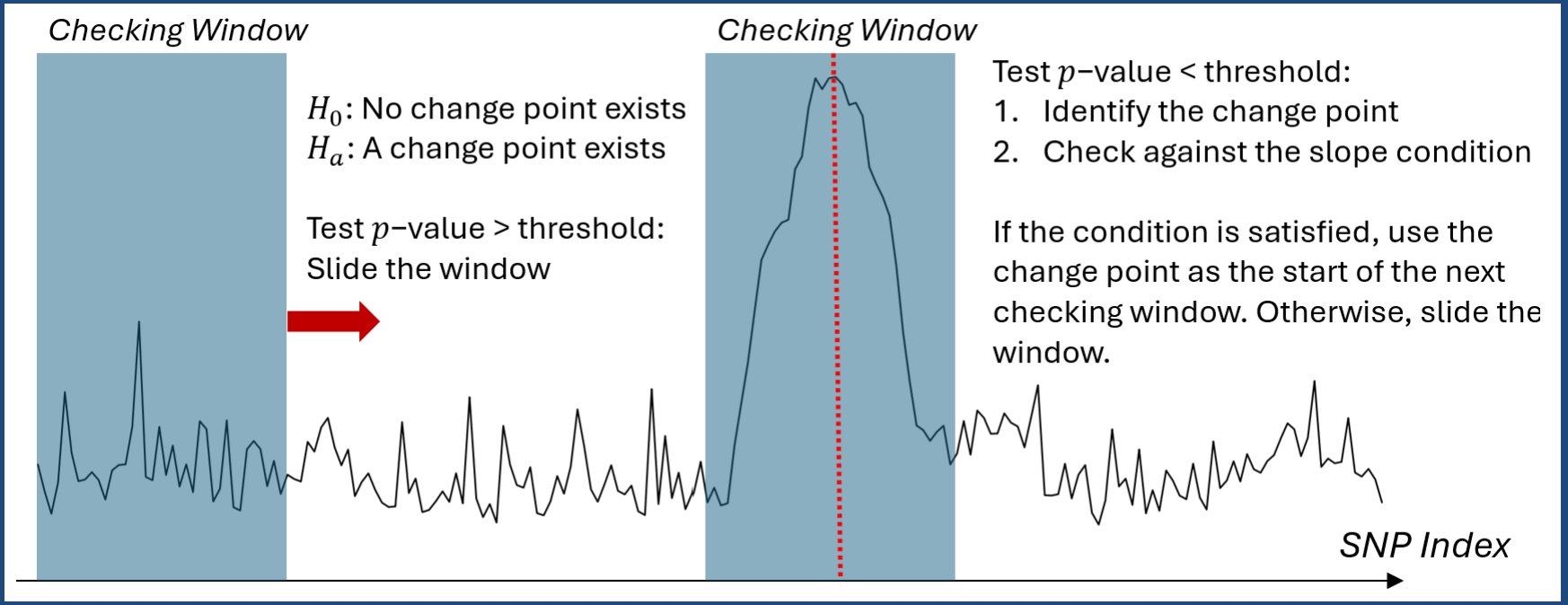
RAS


Pivotal SNP
Adaptive Window
Maximize over windows
SNP Index
Fit model (1): Phenotype ~ SNPs in a window + Covariates
LPRS
Fit model: Phenotype ~ LPRS + Covariates
Association Score (AS) = −log10 p−value (LPRS)
Regional Association Score (RAS)
of the pivotal SNP defined in equation (2)

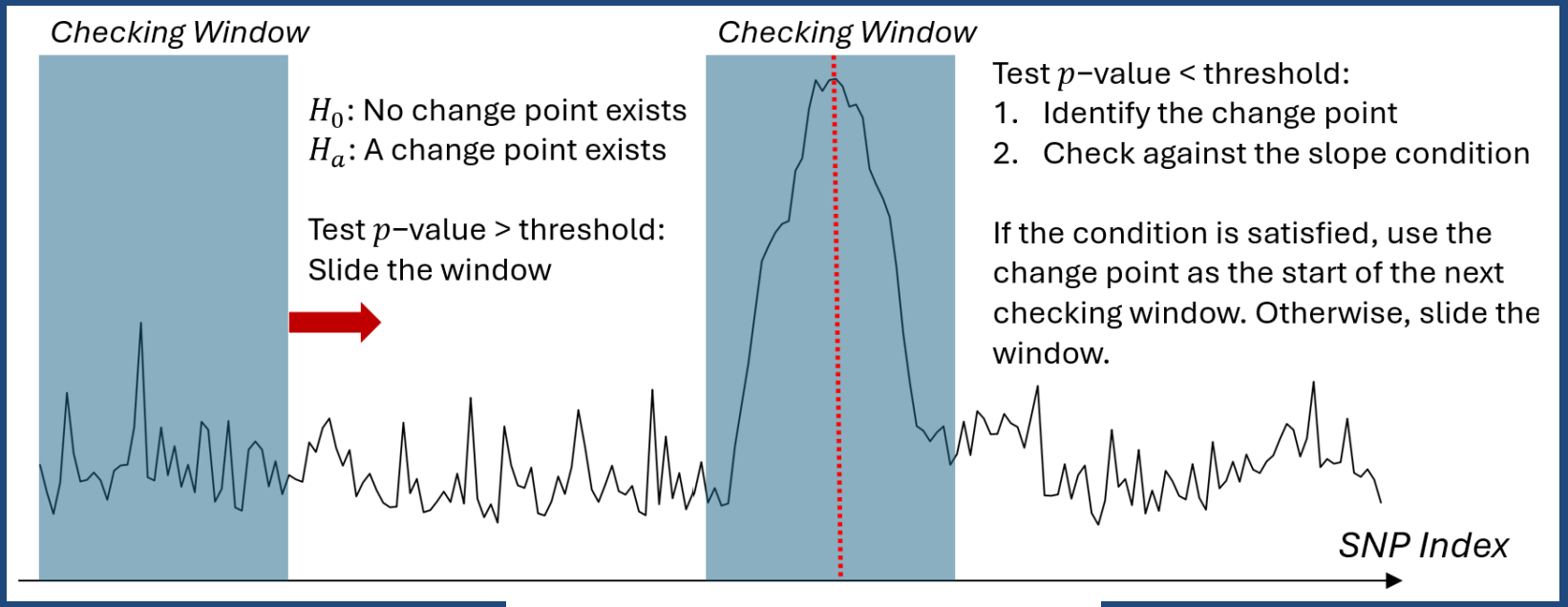

Pivotal SNP
Maximize over windows
SNP Index
Regional Association Score (RAS)
of the pivotal SNP defined in equation

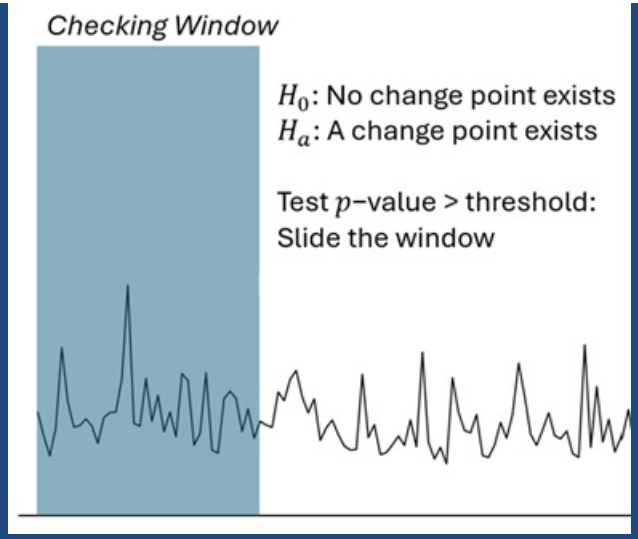
RAS

2. Formulate the RAS into time series data by ordering the RAS values by genomic position along the whole chromosome.
3. Apply a *changepoint detection algorithm* to scan this RAS time-series to identify “peaks.”

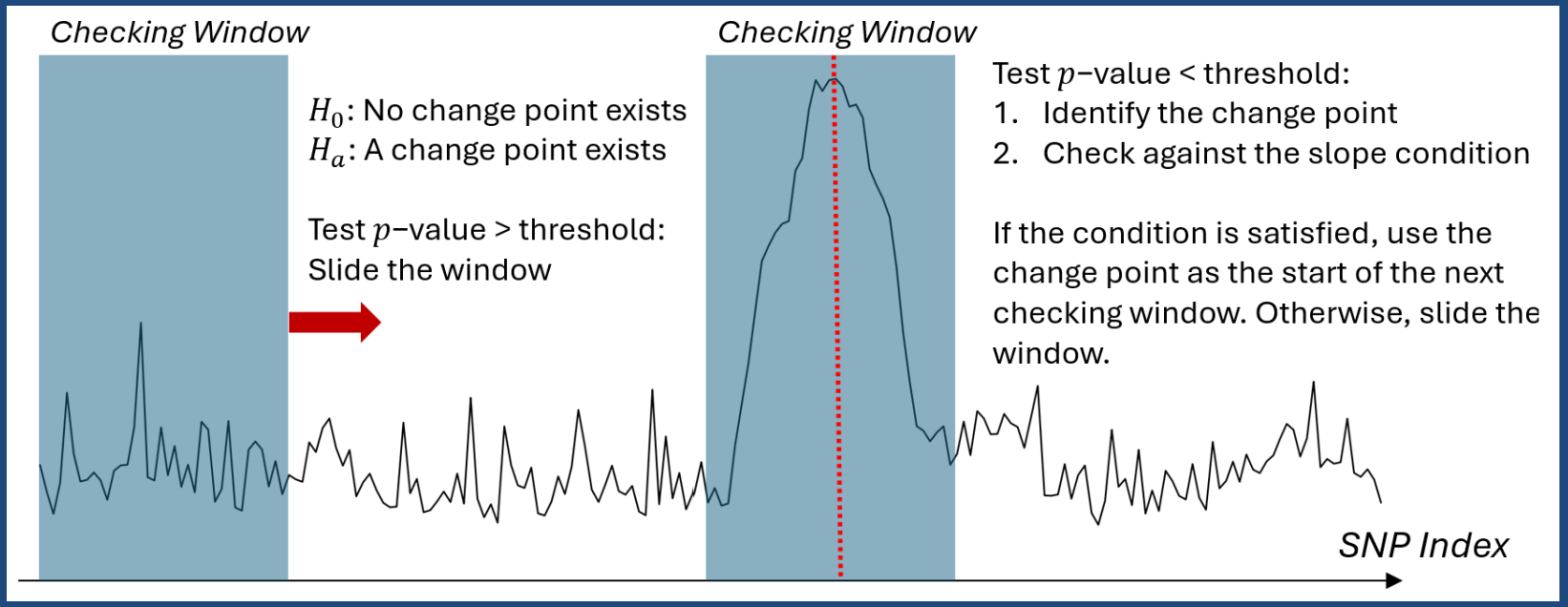


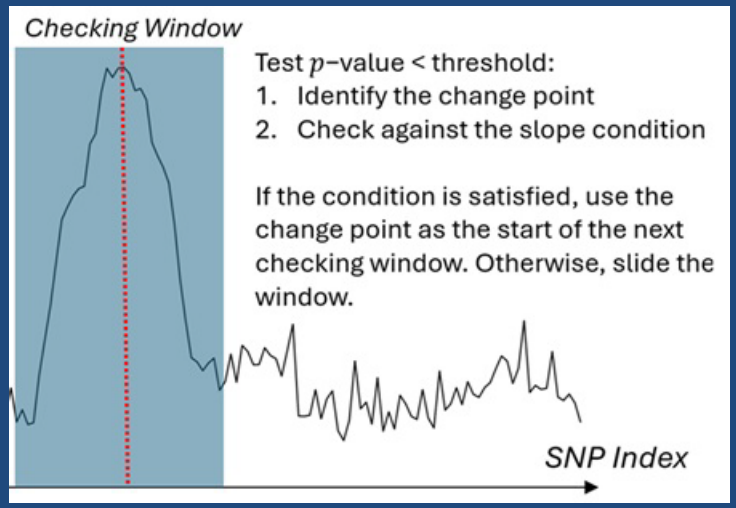
RAS


Checking Window
Checking Window
H_0: No change point exists
H_a: A change point exists
Test p–value > threshold:
Slide the window
Test p–value < threshold:
1. Identify the change point
2. Check against the slope condition
If the condition is satisfied, use the change point as the start of the next checking window. Otherwise, slide the window.
SNP Index

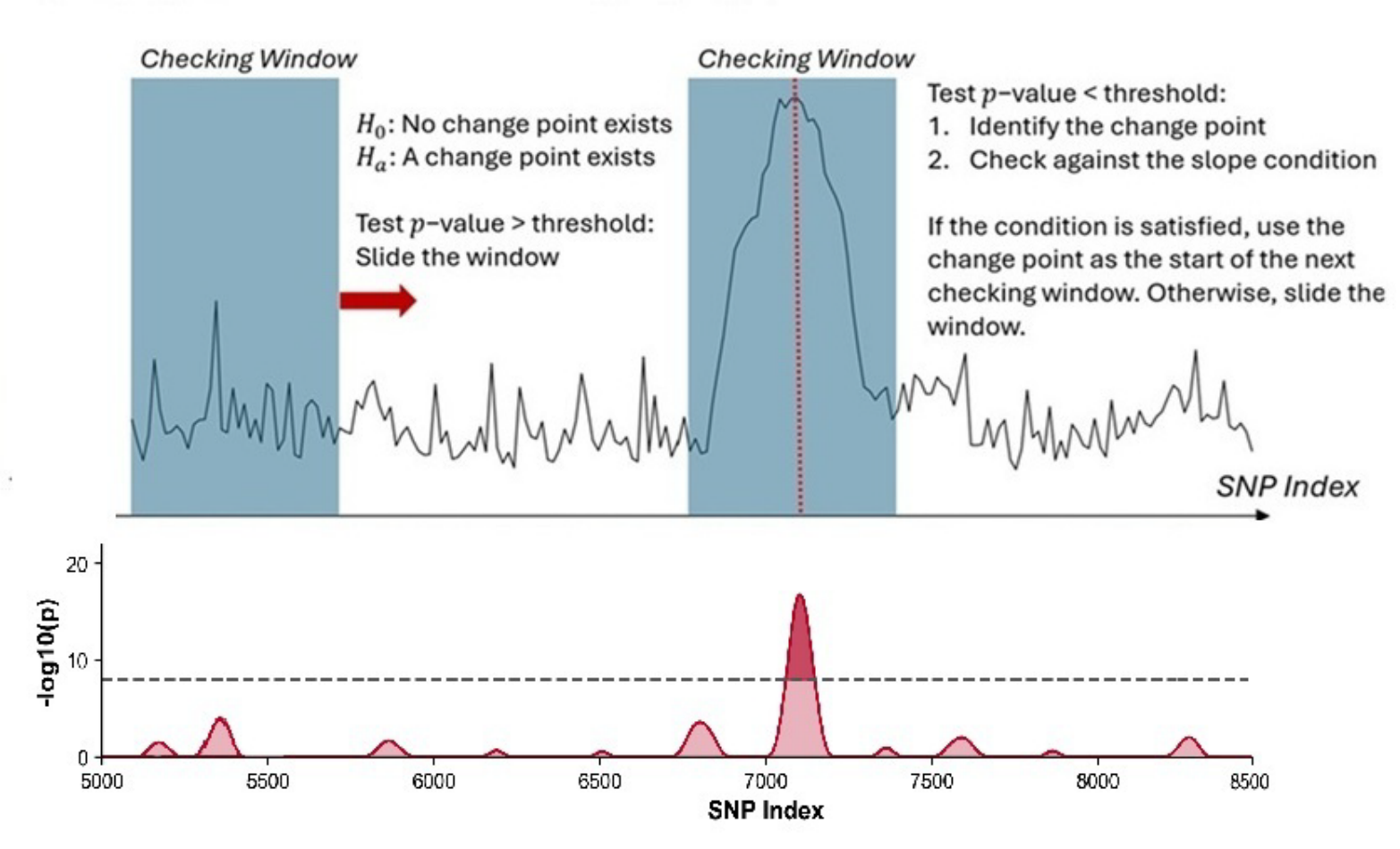

Checking Window
H_0: No change point exists
H_a: A change point exists
Test p–value > threshold:
Slide the window

RAS


Checking Window
Checking Window
H_0: No change point exists
H_a: A change point exists
Test p–value > threshold:
Slide the window
Test p–value < threshold:
1. Identify the change point
2. Check against the slope condition
If the condition is satisfied, use the change point as the start of the next checking window. Otherwise, slide the window.
SNP Index


Checking Window
Test p–value < threshold:
1. Identify the change point
2. Check against the slope condition
If the condition is satisfied, use the change point as the start of the next checking window. Otherwise, slide the window.
SNP Index

RAS



(Illustrative figures. Do not reflect real result)

RAS: Changepoint Detection

To account for an unknown number of changepoints, we apply the sliding window approach proposed by Zhang (1995), combined with the Davies test (Davies et al., 1995).

Single Changepoint Detection

Given a model

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 (X_i - \psi)_+ + \epsilon_i,\ \epsilon \sim N(0, \sigma^2),$$

Davies test tests for the existence of a single break point:

- H_0: No such a ψ exists or $\beta_2 = 0$,
- H_a: ψ exists or $\beta_2 \neq 0$.

Changepoint Detection

- Assume K candidate break points, $\psi_1 < \psi_2 < \ \ldots < \psi_K$.
- For each candidate break point, a likelihood ratio form statistics $S(\psi_k)$ can be calculated.
- The final test statistics is calculated as the supreme of the K likelihood ratio statistics.

Davies provided an approximation of the p-value of the final statistic

$$p \approx \Phi(-M) + V\exp\left(-\frac{M^2}{2}\right)(8\pi)^{-1/2}$$

where $M = \ \max\left(S(\psi_k)\right)$ and $V = \Sigma_k(|S(\psi_k) \ - S(\psi_{k-1})|)$ is the total variation of $S(\psi_k)$. $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution.

Type I Error Rates

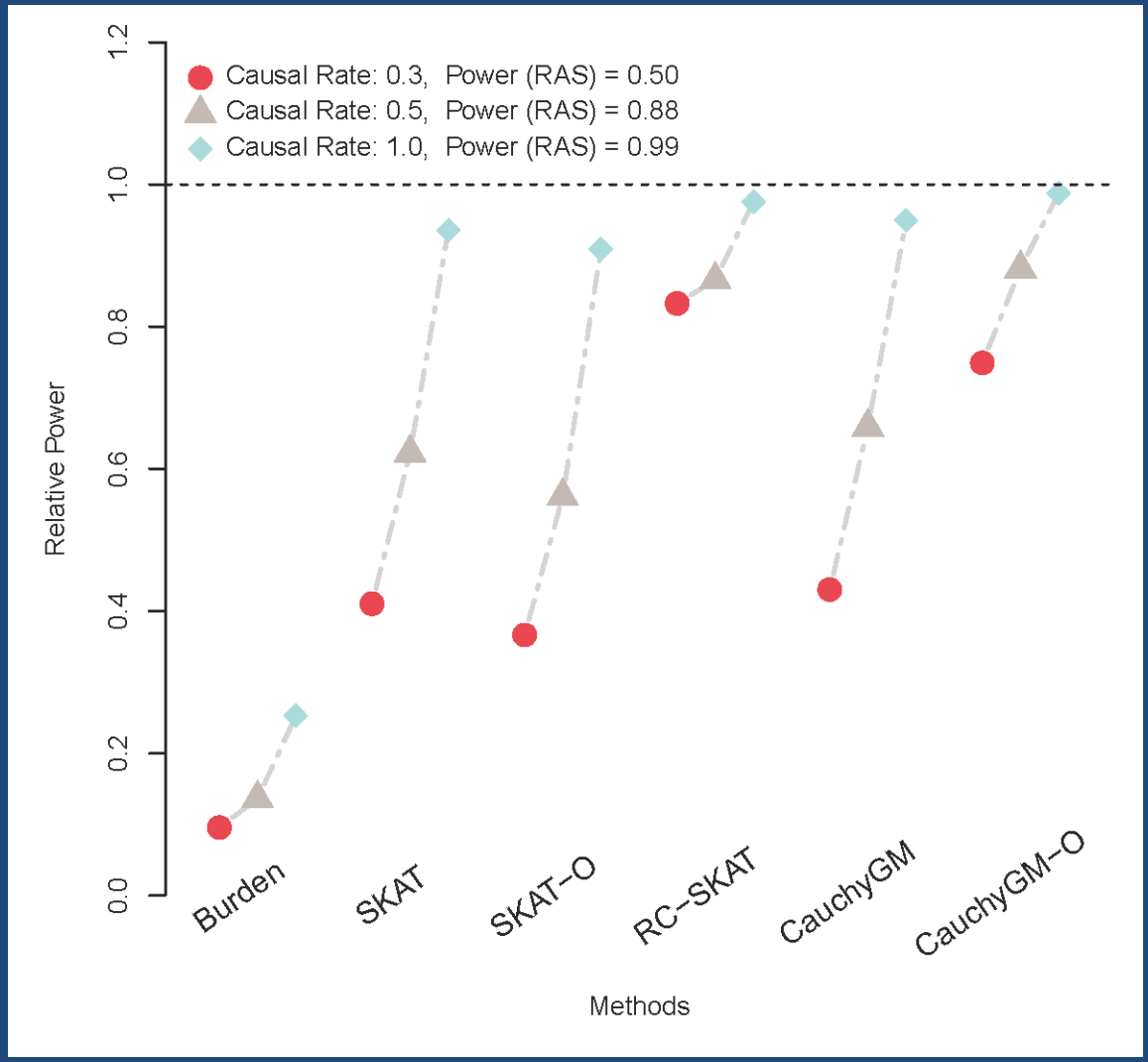
	RAS	Burden	SKAT	SKAT-O	RC-SKAT	CauchyGM	CauchyGM-O
Continuous	0.042 (0.009)	0.046 (0.009)	0.050 (0.010)	0.050 (0.010)	0.050 (0.010)	0.050 (0.010)	0.048 (0.010)
Dichotomous	0.044 (0.009)	0.050 (0.010)	0.050 (0.010)	0.048 (0.010)	0.050 (0.010)	0.050 (0.010)	0.050 (0.010)

Estimated with 500 runs. The standard deviations are given in parentheses. The experiment is conducted on chromosome 21 of the ABCD data, containing 5,718 SNPs.

Power Comparison

The power of RAS is used as the reference, and the y-axis is the ratio of the power of each alternative method to that of RAS.

Causal SNPs are located in three separated regions (100 SNPs each) with the effect sizes correlated to the MAFs of the causal SNPs.

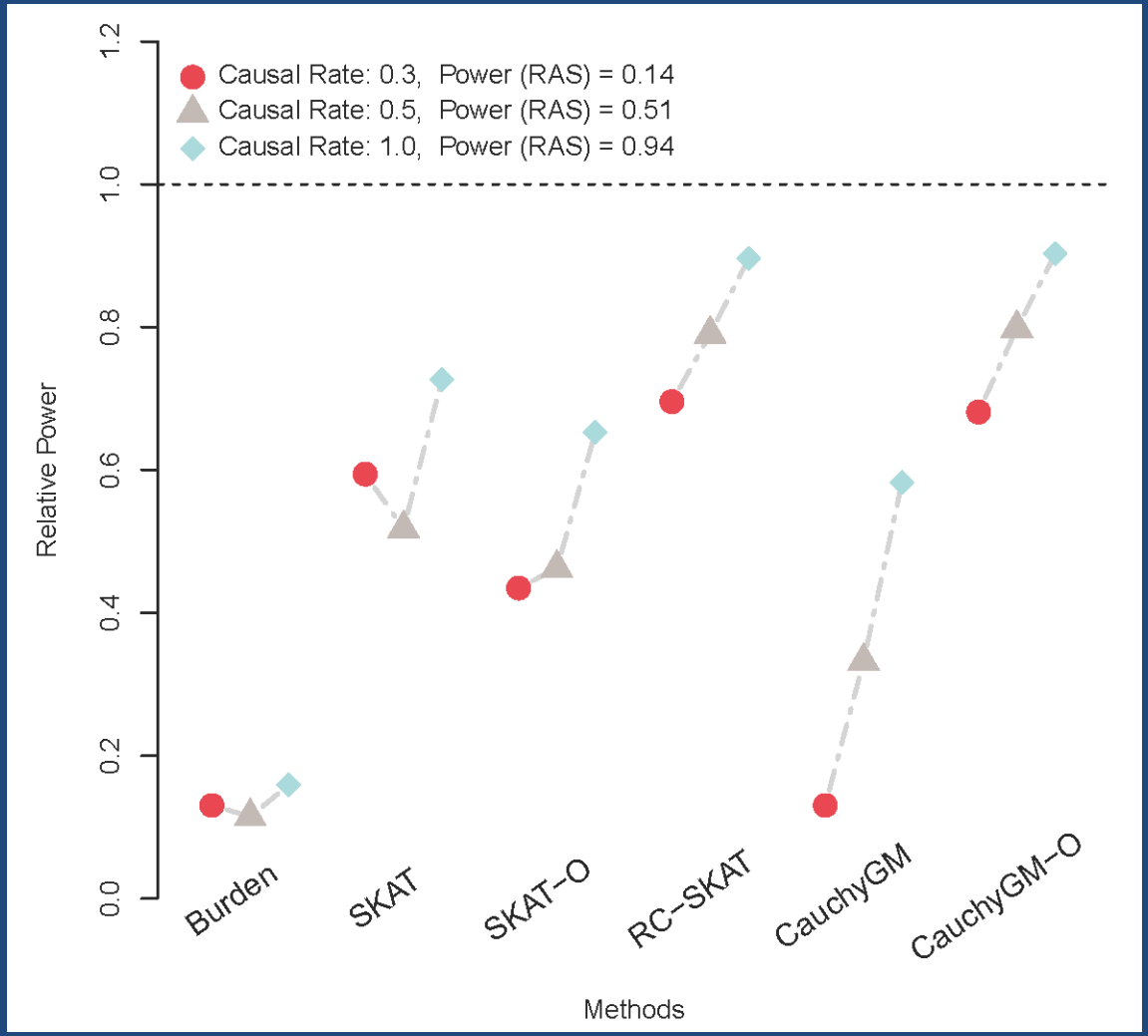


Continuous Trait

Power Comparison

The power of RAS is used as the reference, and the y-axis is the ratio of the power of each alternative method to that of RAS.

Causal SNPs are located in three separated regions (100 SNPs each) with the effect sizes correlated to the MAFs of the causal SNPs.



Binary Trait

Body Mass Index

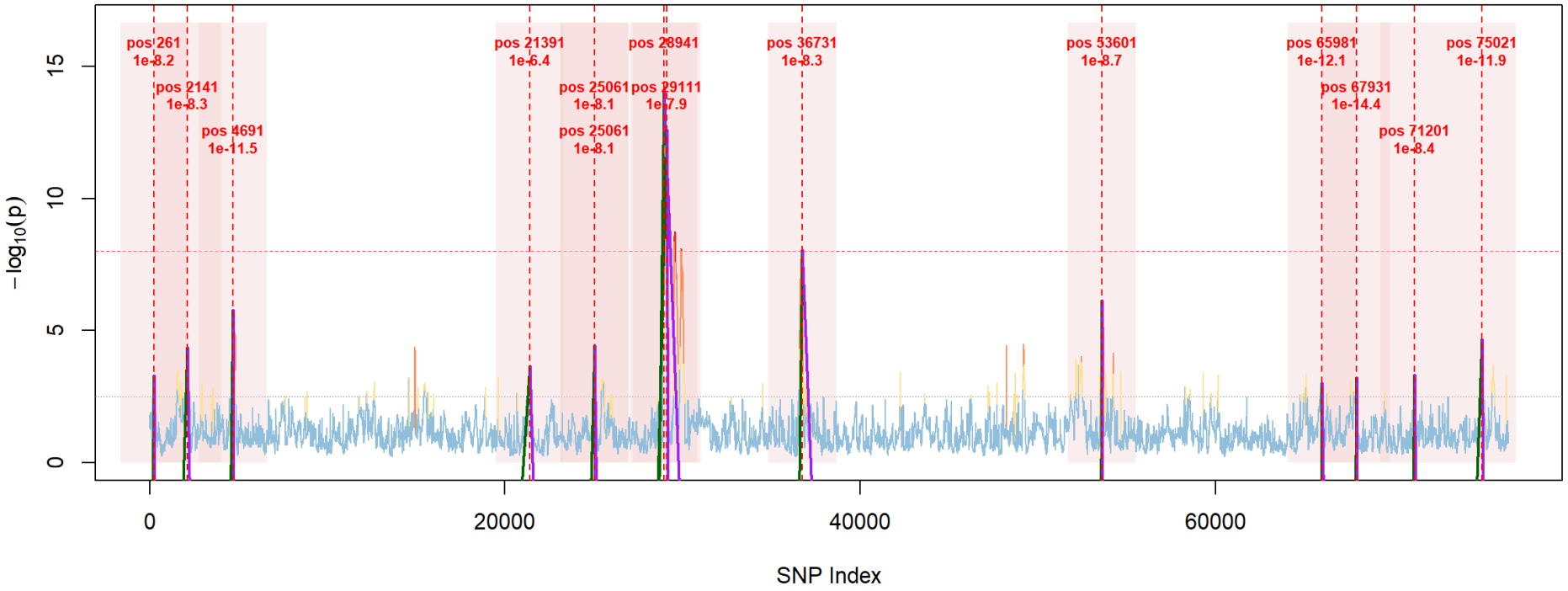
Sample size: 895,285 individuals from All of Us Program

BMI: Min: 10.0; 1st Q: 24.5; Median: 28.4; Mean: 29.8; 3rd Q: 33.6; Max: 121.3

Covariates: Sex, age, and the first 10 genetic principal components

Genotype data: Chromosome 16, 76,486 SNPs after quality control

Body Mass Index



@261: RAB11FIP3, related to metabolism and vesicle transport, but direct BMI evidence is weak
@2,141: lncRNA region, weak BMI evidence
@4,691: SMIM22, little BMI evidence
@36,731: FTO, one of the most established BMI loci
@67,931: CDH13, related to obesity, insulin resistance, and adiponectin signaling

Conclusion

- With high dimensional features, the use of domain knowledge and “attention seeking” ideas can be effective in both dimension reduction and identification of relationship.
- In statistics, we emphasize on the training of using mechanistic techniques such as regularization. But this is insufficient in studying more challenging data structures.

Collaborators

Yiran Jiang

Yue Hu

Jiahe Jin

Acknowledgements

- This research has been conducted using the UK Biobank Resource under application number 42009.
- The Yale Center for Research Computing provided the research computing infrastructure.