Scalable Statistical Inference for Massive Health Science Data

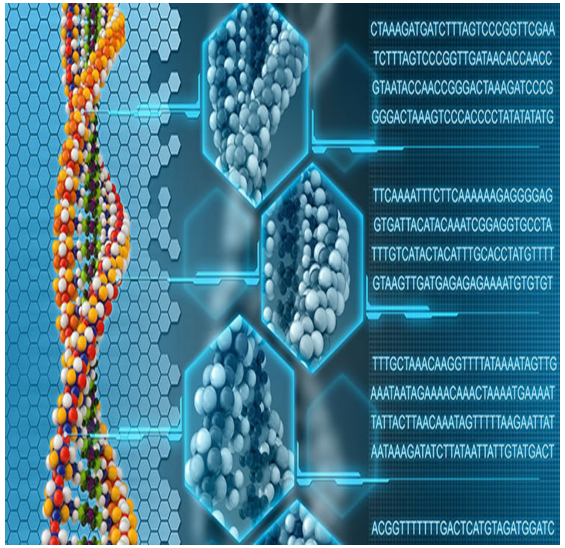
Xihong Lin

Department of Biostatistics and Department of Statistics

Harvard University

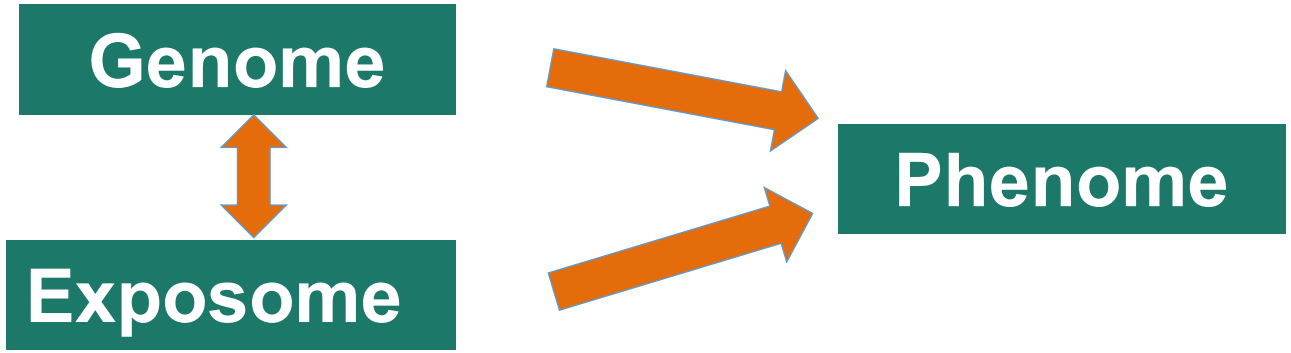
Examples of Genome, Exposome and Phenome

Whole Genome Sequencing



Smartphone Data

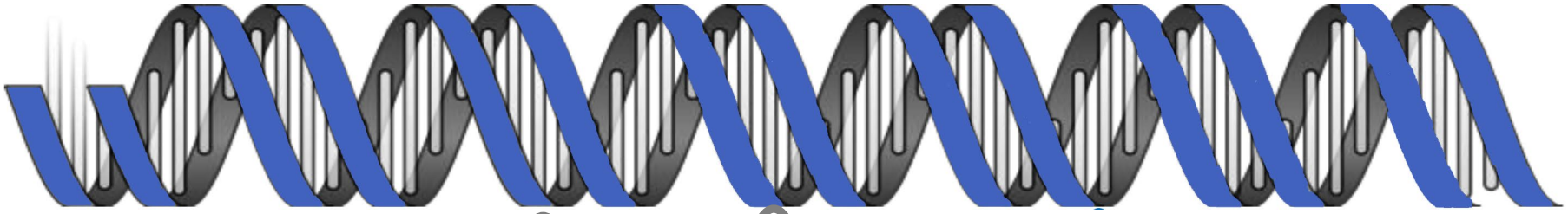
Electronic Medical Records



Our Niche in Big Data Era: Scalable Statistical Inference

Goal: To solve big problems

Whole Genome Sequencing Studies (2015-)



CCGATCCAAGTCCATATATACCGATTTAACCGAA

CCGATCCAAGTCCATATATACCAATTTAACCGAA

CCGATCCAAGTCCATACATACCGATTTAACCGAA

CCGATCCAAGTCCATACATACCGATTTAACCGAA

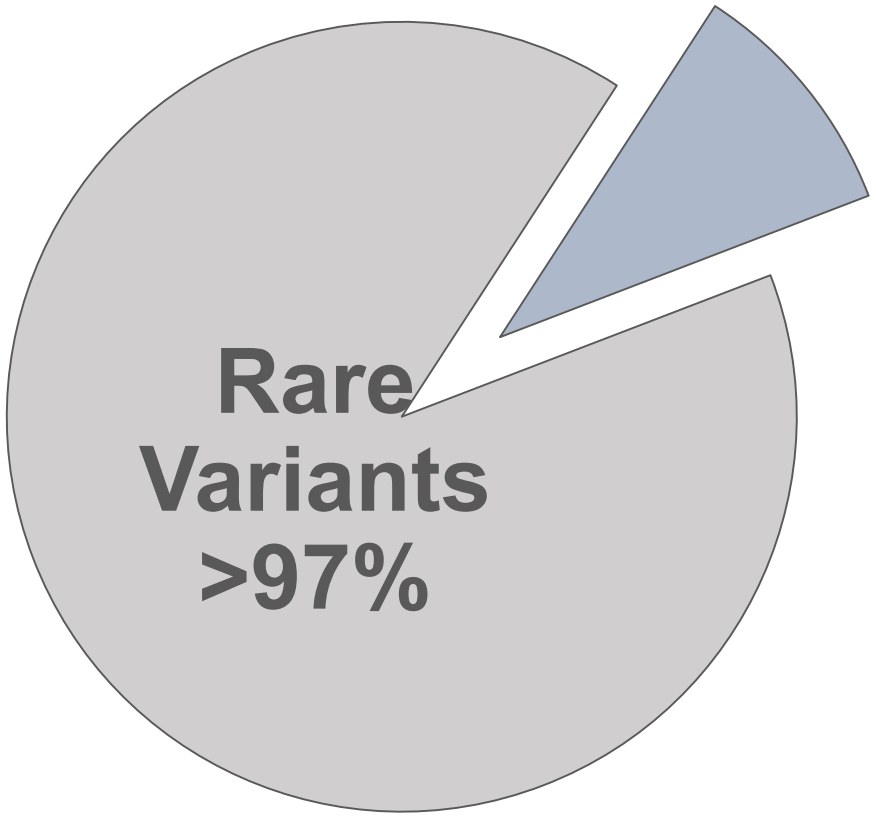
CCAATCCAAGTTCATATATACCGATTTGACCGAA

CCGATCTAAGTCCATATATACCGATTTAACCGAA

CCGATCCAAGTCCATACATACCGATTTAACCGAA

WGS Covers 100% of the Genome

TOPMed Freeze 5 (n=54,000): 430M Variants (97% are rare variants)



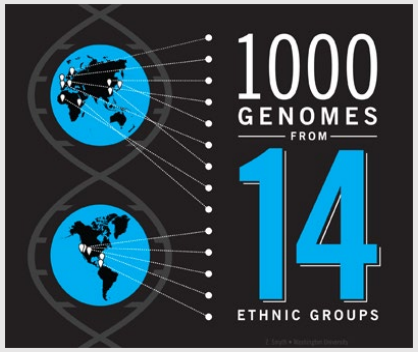
GWAS Common
Variants <3%

Rare variants are more likely to cause diseases and their coded proteins are more likely to be drug targets.

Large Scale WGS Timeline

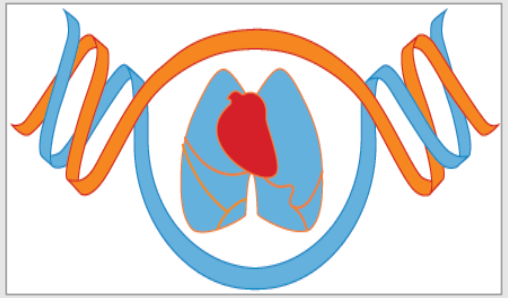
2008

1000 Genomes
N=1000



2015

TOPMed
(NHLBI)
N=150,000

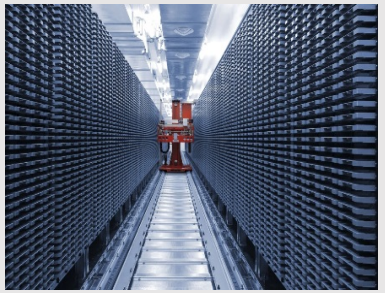


2016

GSP(NHGRI)
N=200,000

2018

Biobanks
(N=millions)



First Goal of WGS Analysis:

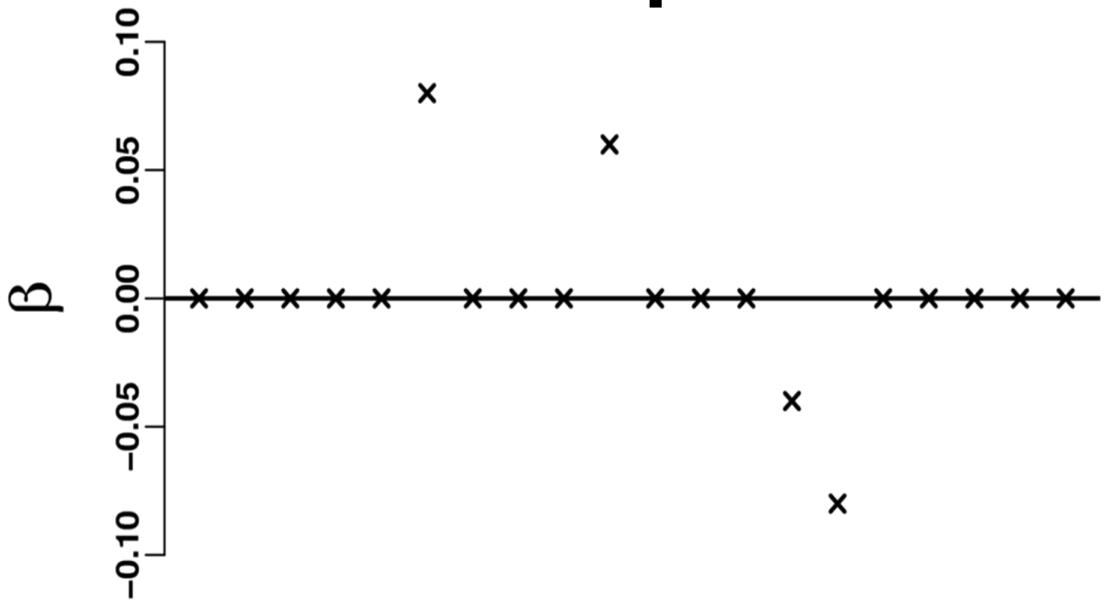
Signal Detection

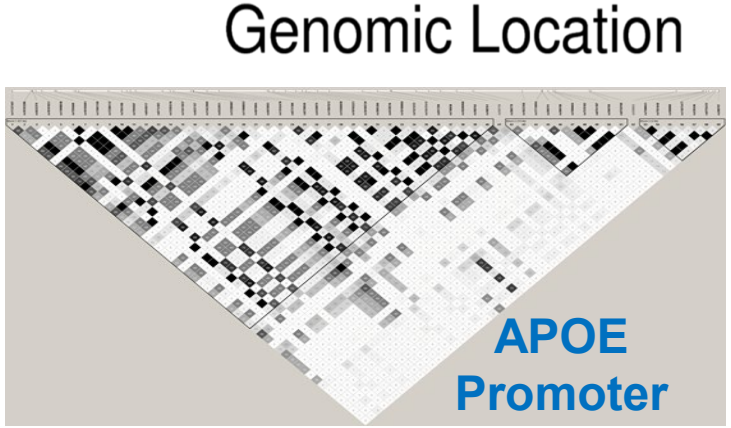
Scan the genome to identify genomic regions associated with diseases/traits

Challenges in Rare Variant Analysis of WGS Data

- Simple single SNP analysis does not work
- Need to perform SNP-set analysis
- Estimation is very difficult

$$LDL = G\beta + e$$

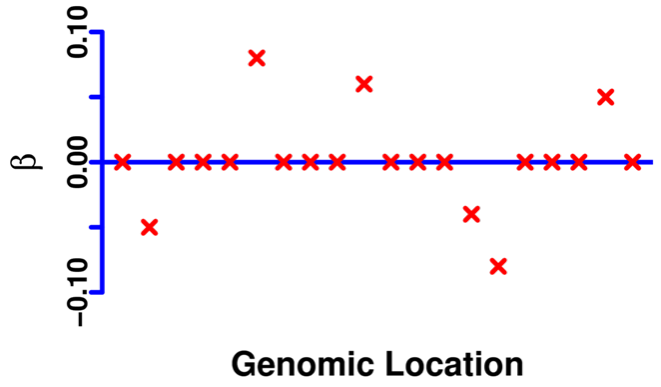




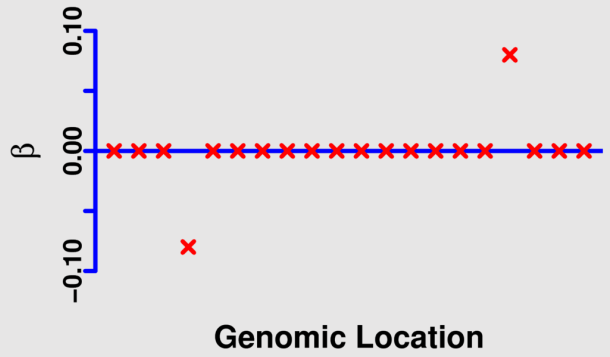
Test for Dense & Sparse High-Dimensional Alternatives

Model: $\mathbf{Y} = \mathbf{G}\boldsymbol{\beta} + \mathbf{e}$ $\quad H_0 : \boldsymbol{\beta} = \mathbf{0}$

Dense Alternative



Sparse Alternative



Sequencing Kernel Association Test (SKAT)

- Wu, et al, 2011, AJHG. (Citations=1400)
- SMMAT
- STAAR

Generalized Higher Criticism (GHC) /Generalized Berk-Jones (GBJ)/ACAT

- Murkerjee, et al, Ann. Stat, 2015
- Barnett, et al 2017 (GHC), JASA
- Sun and Lin (GBJ), 2017
- Liu, et al (ACAT), 2018

Model and Hypothesis

- Y_i is phenotype (outcome) $(i = 1, \cdots, n)$
- $\boldsymbol{X}_i$ contains q covariates
- $\boldsymbol{G}_i$ contains p SNPs (AA, AB, BB=0,1,2) in a SNV set, e.g., variants in the promoter region of APOE.
- $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ contain regression coefficients.
- $\mu_i = E(Y_i|\boldsymbol{G}_i, \boldsymbol{X}_i)$

Model

$$h(\mu_i) = \boldsymbol{X}_i^T \boldsymbol{\alpha} + \boldsymbol{G}_i^T \boldsymbol{\beta}$$

- Hypothesis of no gene/network effect (p might be large):

$$H_0 : \boldsymbol{\beta} = 0 \text{ and } H_1 : \boldsymbol{\beta} \neq 0 \text{ (weak)}.$$

Challenges Addressed in Scalable Inference for WGS Data

- $p = \dim(\boldsymbol{\beta})$ might not be small
- Full GLMs hard to fit due to rare variants
- Solution:
  - Use score statistics $Z_j = \sum_{i=1}^{n} G_{ij}(Y_i - \hat{\mu}_{i0})$
  - Scability: Fit the null same null model $g(\mu_i) = \boldsymbol{X_i'\alpha}$ only once when scanning the genome

Dense/Sparse Alternatives

- Unknown Truth: $k = p^{1-\alpha}$ of β_j's $\neq 0$
- Hypothesis

$$H_0 : \boldsymbol{\beta} = 0$$
$$H_1 : \text{Some } \beta_j \neq 0$$

- Dense alternative ($\alpha < 1/2$):

$$\text{Ex: } p = 100, \alpha = 0.4 \Rightarrow k = 16$$

- Sparse alternative ($\alpha > 1/2$):

$$\text{Ex: } p = 100, \alpha = 0.6 \Rightarrow k = 7$$

Difficulties in Testing

- No global optimal most powerful test exists.
- Test optimality depends on
  - Genotype matrix(G) : Sparsity, LD (correlation)
  - Signals β: Sparsity, strength, and sign
  - Distribution of Y

Dense Regime

- **Burden(B)** (if all variants are causal with effects (β's) in the same direction)

$$B = \left(\sum_{j}^{p} w_j Z_j \right)^2$$

- **SKAT** (if there are neural variants and/or with effects (β's) in different directions)

$$S = \sum_{j}^{p} w_j Z_j^2$$

Sparse Regime: Higher Criticism (HC) (Tukey, 1976)

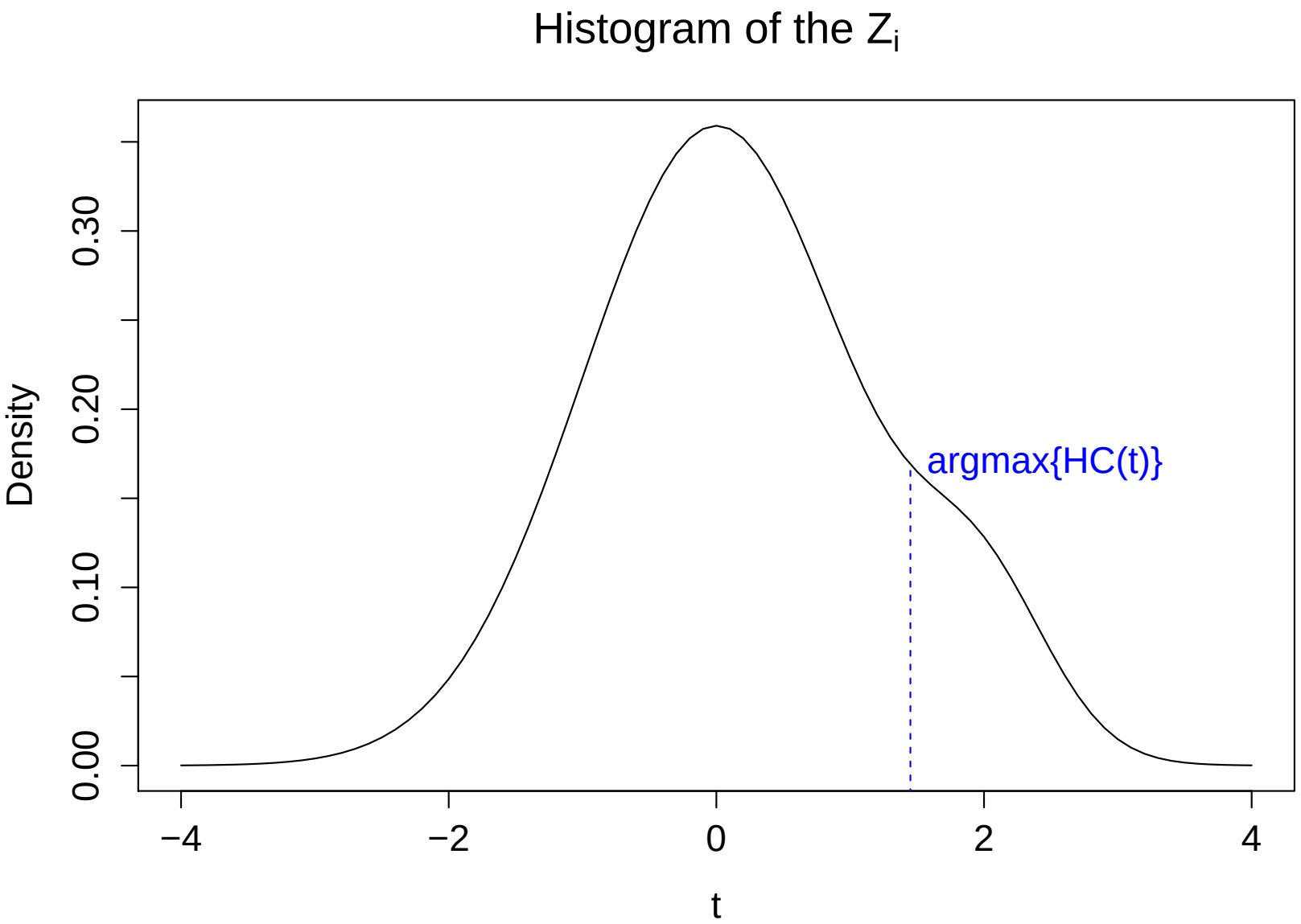
- Let

$$S(t) = \sum_{j=1}^{p} \mathbf{1}_{\{|Z_j| \geq t\}}$$

- **Assumes $\Sigma = I_p$ or sparse (G is a low coherence matrix)**
- The HC test statistic is (Ingster, 1998; Donoho and Jin, 2003; Arias-Castro, et al, 2011)

$$HC = \sup_{t>0} \left\{ \frac{S(t) - 2p\bar{\Phi}(t)}{\sqrt{2p\bar{\Phi}(t)(1 - 2\bar{\Phi}(t))}} \right\}$$

The Higher Criticism



Linear Regression: Existing Results on Detection Boundary

Dense Regime $(\alpha \leq \frac{1}{2})$	**Sparse Regime** $(\alpha > \frac{1}{2})$
$A \ll \sqrt{\frac{p^{\alpha-\frac{1}{2}}}{n}} \Rightarrow$ all tests powerless.	$A < \sqrt{\frac{2t\log p}{n}}$, $t < \rho^*_{\text{gaussian}}(\alpha) \Rightarrow$ all tests powerless.
$A \gg \sqrt{\frac{p^{\alpha-\frac{1}{2}}}{n}} \Rightarrow$ SKAT powerful	$A > \sqrt{\frac{2t\log p}{r}}$, $t > \rho^*_{\text{gaussian}}(\alpha) \Rightarrow$ HC powerful.

Setting

- Low coherence matrix **G** (sparse correlation Σ)
- A=signal strength of β.
- Sparsity index: $k = p^{1-\alpha}$

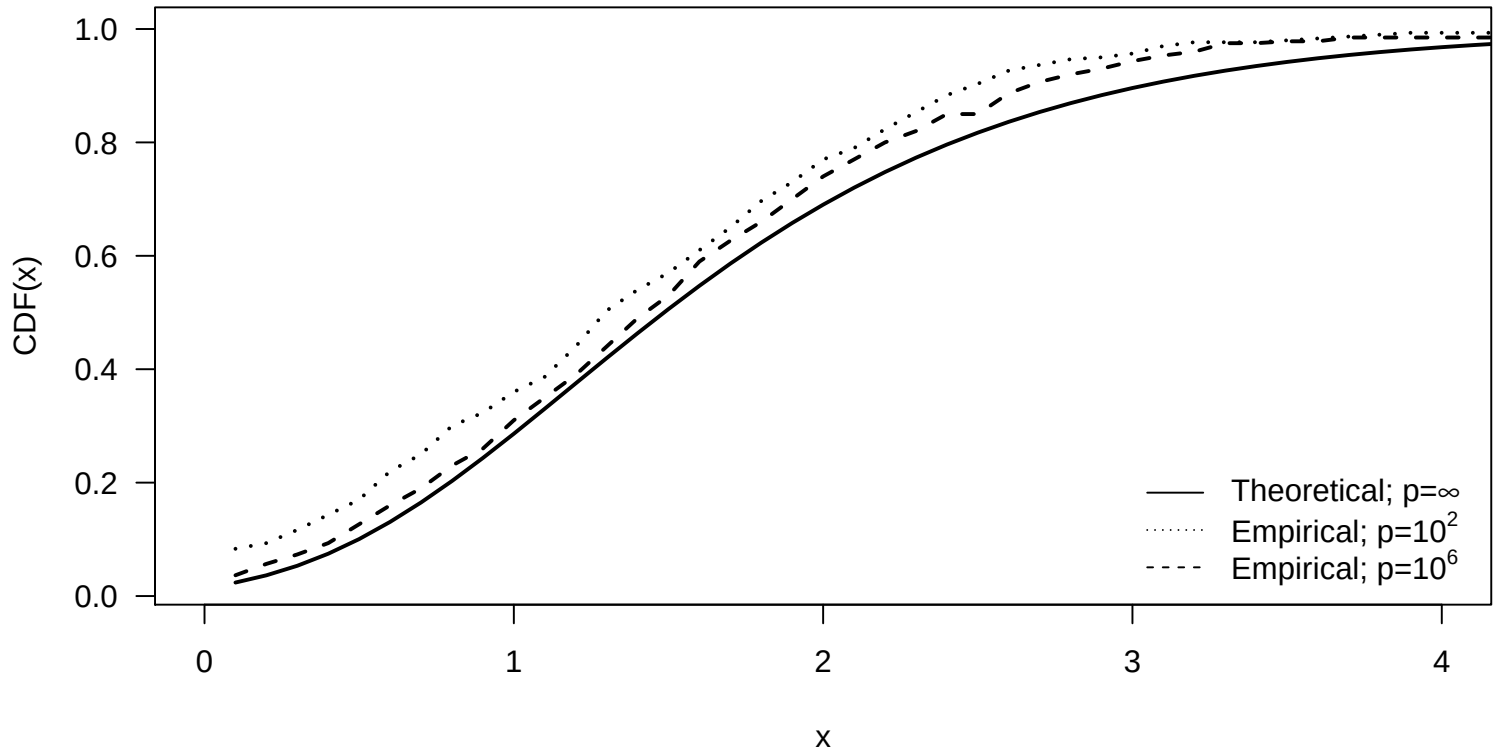
The results for binary regression are different from linear regression (Mukherjee, et al, 2015, Ann Stat)

- If design matrices are too sparse, then signal detection is impossible no matter how strong signals are.
- Two point detection boundary: Maximal Sparsity of $\mathbf{G}$ and Minimal Signal Strength $\boldsymbol{\beta}$.

Asymptotic p-values for HC Does Not Work Well for Finte p

- The supremum of this standardized empirical process follows a Gumbel distribution asymptotically.
- Jaeschke (1979) shows that this converges in distribution at an abysmal rate of $O\{(\log p)^{-1/2}\}$

Slow Convergence to Asymptotic Distribution of HC



- In genetic studies, gene and network sizes

  (p=# of SNPs=dozens to thousands)

Analytic p-values for HC for Finite p (Barnett and Lin, Biometrika, 2015)

- Letting h be the observed HC statistic:

$$\text{p-value} = pr\left(\sup_{t>0}\left\{\frac{S^*(t) - 2p\bar{\Phi}(t)}{\sqrt{2p\bar{\Phi}(t)(1-2\bar{\Phi}(t))}}\right\} \geq h\right)$$

- There exists $0 < t_1 < \cdots < t_p$, such that

$$\text{p-value} = 1 - pr\left(\bigcap_{k=1}^{p}\{S^*(t_k) \leq p-k\}\right)$$

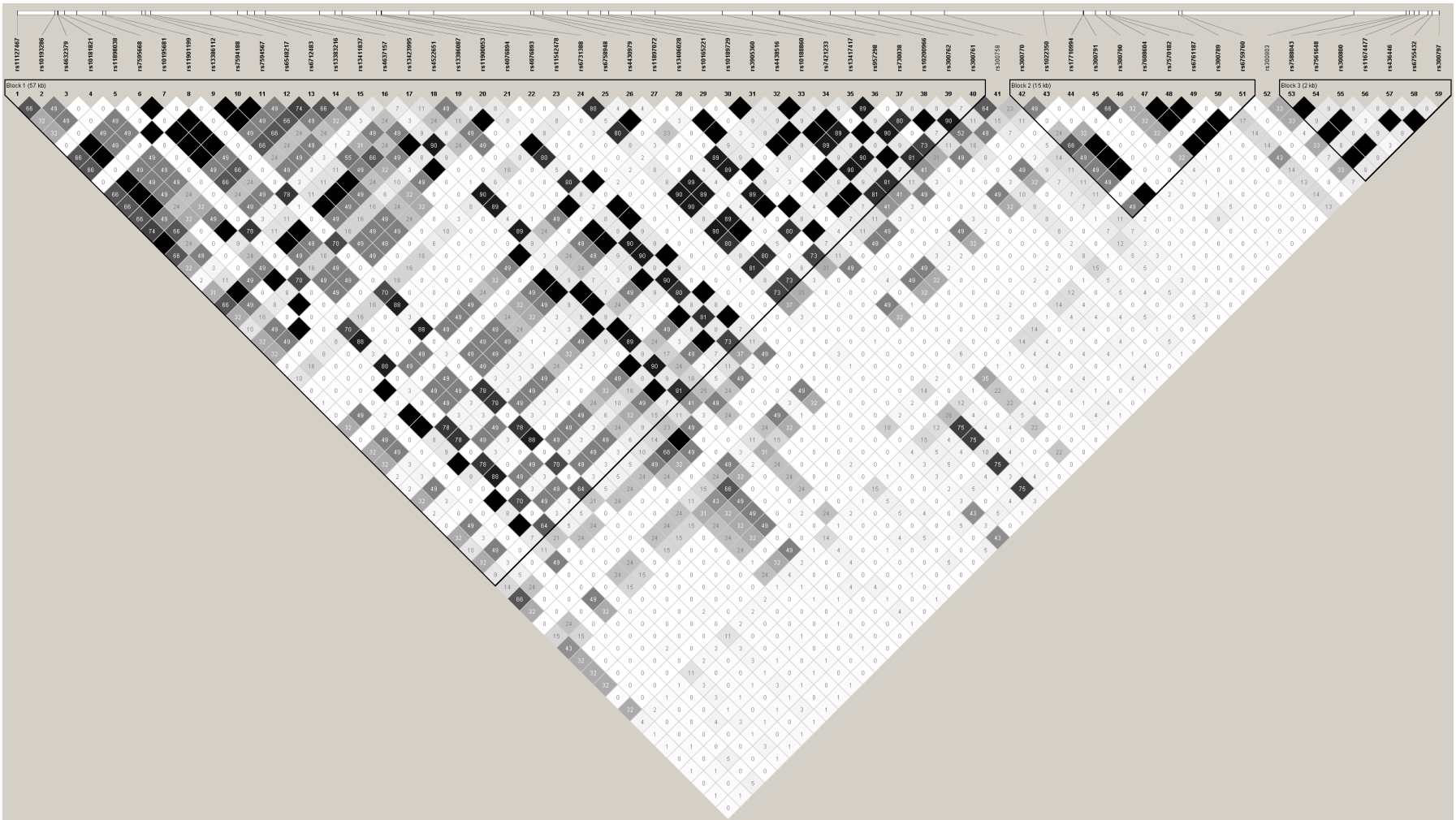
Then apply the chain rule of conditioning to get a product of binomial probabilities.

Simulation Study of Type I error rates of HC: Analytic(Exact) vs Asymptotic

	p	
α	10	50
$1{\cdot}0$	$9{\cdot}92 \times 10^{-1}(7{\cdot}31 \times 10^{-1})$	$1{\cdot}01(1{\cdot}59 \times 10^{-1})$
$1{\cdot}0 \times 10^{-1}$	$1{\cdot}01 \times 10^{-1}(6{\cdot}03 \times 10^{-2})$	$9{\cdot}75 \times 10^{-2}(4{\cdot}90 \times 10^{-3})$
$1{\cdot}0 \times 10^{-2}$	$1{\cdot}12 \times 10^{-2}(7{\cdot}30 \times 10^{-3})$	$9{\cdot}80 \times 10^{-3}(4{\cdot}00 \times 10^{-4})$

Need to account for Correlation among SNPs (LD))

- CHRNA3-5 Gene Region



Accounting for correlation: Innovated HC (iHC) (Hall and Jin, 2011)

- Letting $\boldsymbol{U}\boldsymbol{U}^{\boldsymbol{T}} = \widehat{Cov}(\boldsymbol{Z}) = \hat{\Sigma}$
- Define the transformed (decorrelated) test statistics:

$$\boldsymbol{Z}^* = \boldsymbol{U}^{-1}\boldsymbol{Z} \xrightarrow[n\to\infty]{\mathcal{L}} MVN(\boldsymbol{0}, \boldsymbol{I}_p)$$

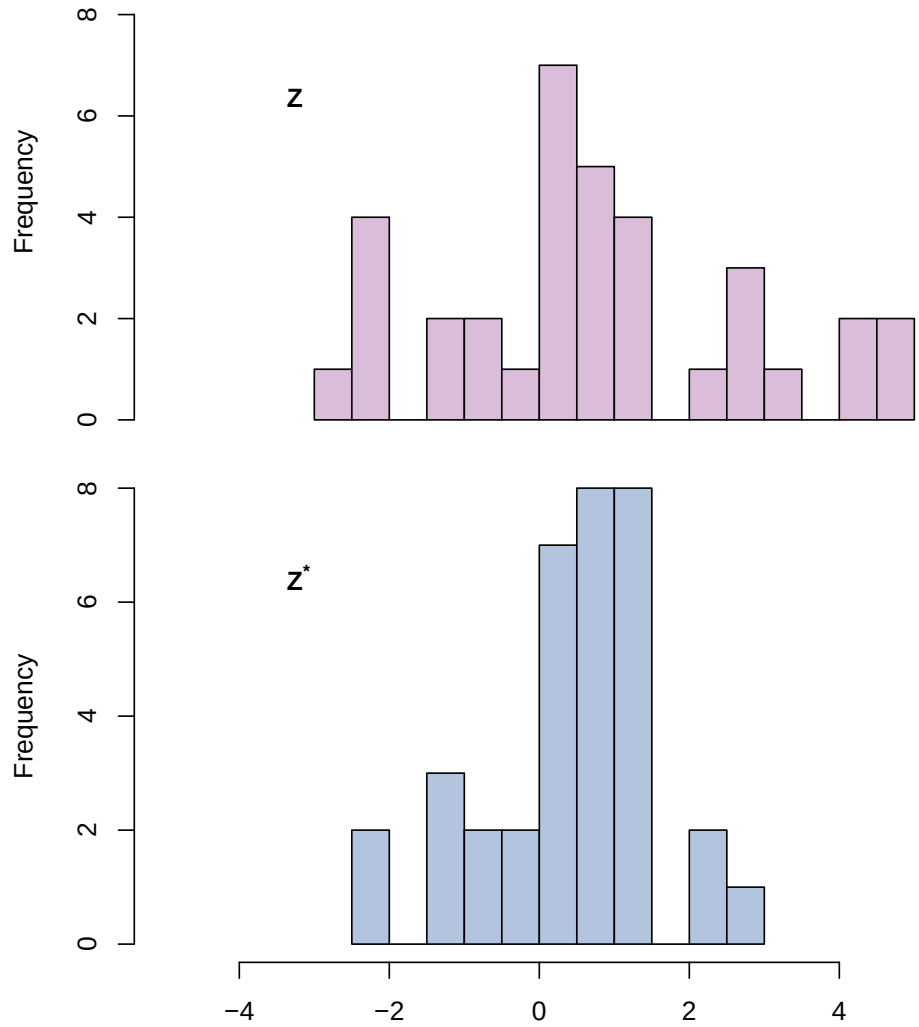
- Set

$$S^*(t) = \sum_{j=1}^{p} \mathbf{1}_{\{|Z_j^*| \geq t\}}$$

- The innovated Higher Criticism test (iHC) statistic is:

$$iHC = \sup_{t>0} \left\{ \frac{S^*(t) - 2p\bar{\Phi}(t)}{\sqrt{2p\bar{\Phi}(t)(1-2\bar{\Phi}(t))}} \right\}$$

Decorrelating using dampens true signals and causes iHC to lose power: CGEM Breast Cancer GWAS: FGFR2 gene



Generalized Higher Critcism (GHC) (Barnett, et al, 2016, JASA)

Recall

$$S(t) = \sum_{j=1}^{p} \mathbf{1}_{\{|Z_j| \geq t\}}$$

- Now **we allow Σ to have arbitrary correlation structure.**
- $S(t)$ is no longer binomial. Instead we approximate with Beta-binomial, matching on first two moments.
- The Generalized Higher Criticism (GHC) test statistic is:

$$GHC = \sup_{t>0} \left\{ \frac{S(t) - 2p\bar{\Phi}(t)}{\sqrt{\widehat{Var}(S(t))}} \right\}$$

- *GHC* achieves the same as detection boundary as *HC*.

The variance estimator $\widehat{Var}(S(t))$

Let $\overline{r^n} = \frac{2}{p(1-p)} \sum_{1 \le k < l \le p} (\Sigma_{kl})^n$ and let $\mathcal{H}_i(t)$ be the Hermite polynomials: $\mathcal{H}_0(t) = 1$, $\mathcal{H}_1(t) = t$, $\mathcal{H}_2(t) = t^2 - 1$ and so on. Then

$$Cov\left(S(t_k), S(t_j) \right) = p[2\bar{\Phi}(max\{t_j, t_k\}) - 4\bar{\Phi}(t_j)\bar{\Phi}(t_k)]$$

$$+4p(p-1)\phi(t_j)\phi(t_k) \sum_{i=1}^{\infty} \frac{\mathcal{H}_{2i-1}(t_j)\mathcal{H}_{2i-1}(t_k)\overline{r^i}}{(2i)!}$$

Analytic p-values for the GHC

- Letting h be the observed GHC statistic:

$$\text{p-value} = pr\left(\sup_{t>0}\left\{\frac{S(t) - 2p\bar{\Phi}(t)}{\sqrt{\widehat{Var}(S(t))}}\right\} \geq h\right)$$

- There exists $0 < t_1 < \cdots < t_p$, such that

$$\text{p-value} = 1 - pr\left(\bigcap_{k=1}^{p}\{S(t_k) \leq p - k\}\right)$$

Generalized Berk-Jones

- **Motivation:** GHC works well in the very sparse signal case but less well in the moderately sparse signal case in finite samples.
- Let s be the realized value of $S(t)$.
- **Berk-Jones (Sup LR test):**

$$BJ = \max_{t>0} \log \left\{ \frac{\Pr\left[S(t) = s | \pi = s/p\right]}{\Pr\left[S(t) = s | \pi = \pi_0\right]} \right\} \mathbf{1}\left\{ \pi_0 < \frac{s}{p} \right\}$$

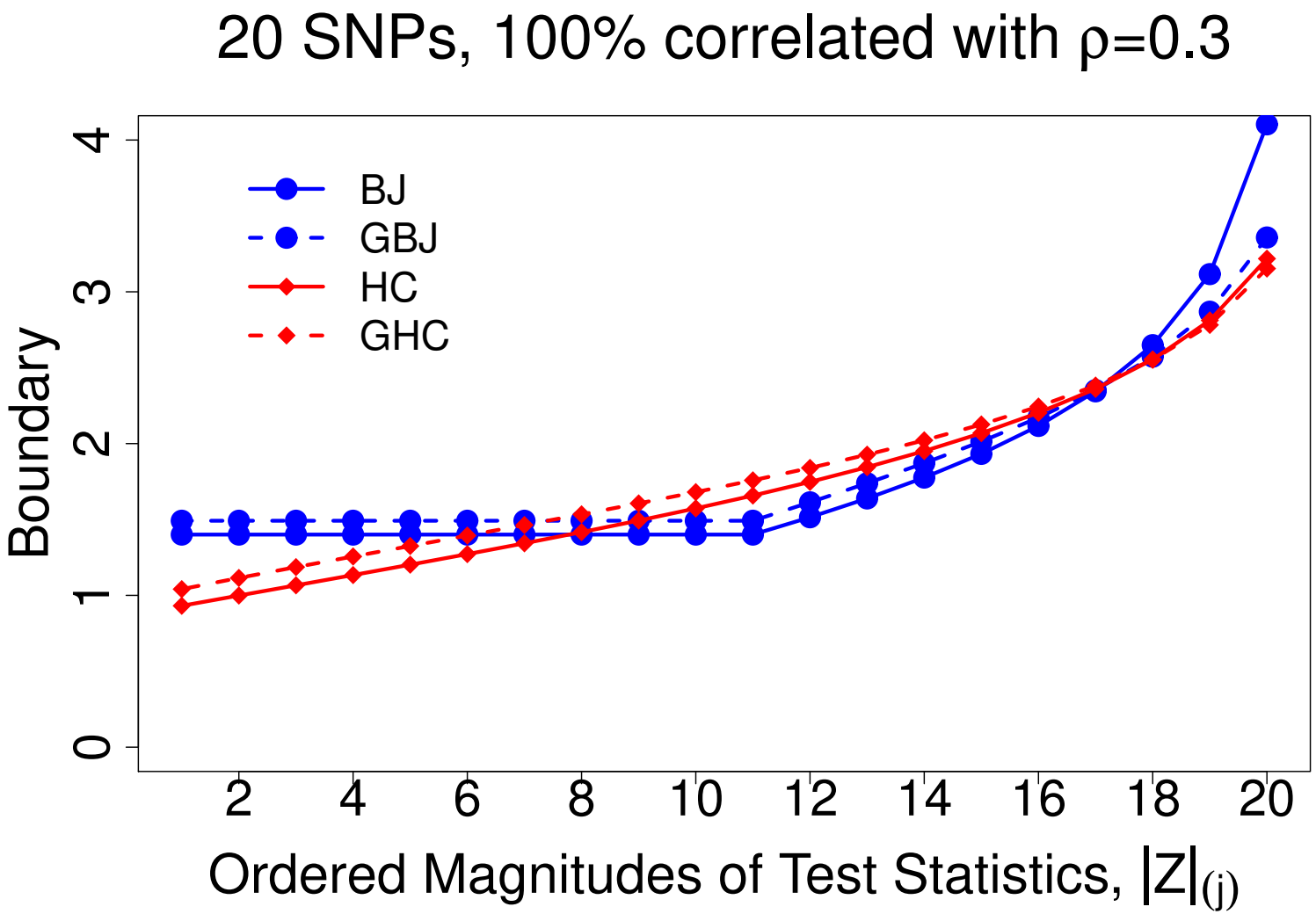
- **Generalized Berk-Jones (Account for correlation):**

$$GBJ = \max_{t>0} \log \left\{ \frac{\Pr\left[S(t) = s | \pi = s/p, \mathrm{cor}(\mathbf{Z}) = \mathbf{\Sigma}\right]}{\Pr\left[S(t) = s | \pi = \pi_0, \mathrm{cor}(\mathbf{Z}) = \mathbf{\Sigma}\right]} \right\} \mathbf{1}\left\{ \pi_0 < \frac{s}{p} \right\}$$

Inference using Generalized Higher Criticism and Generalized Berk-Jones

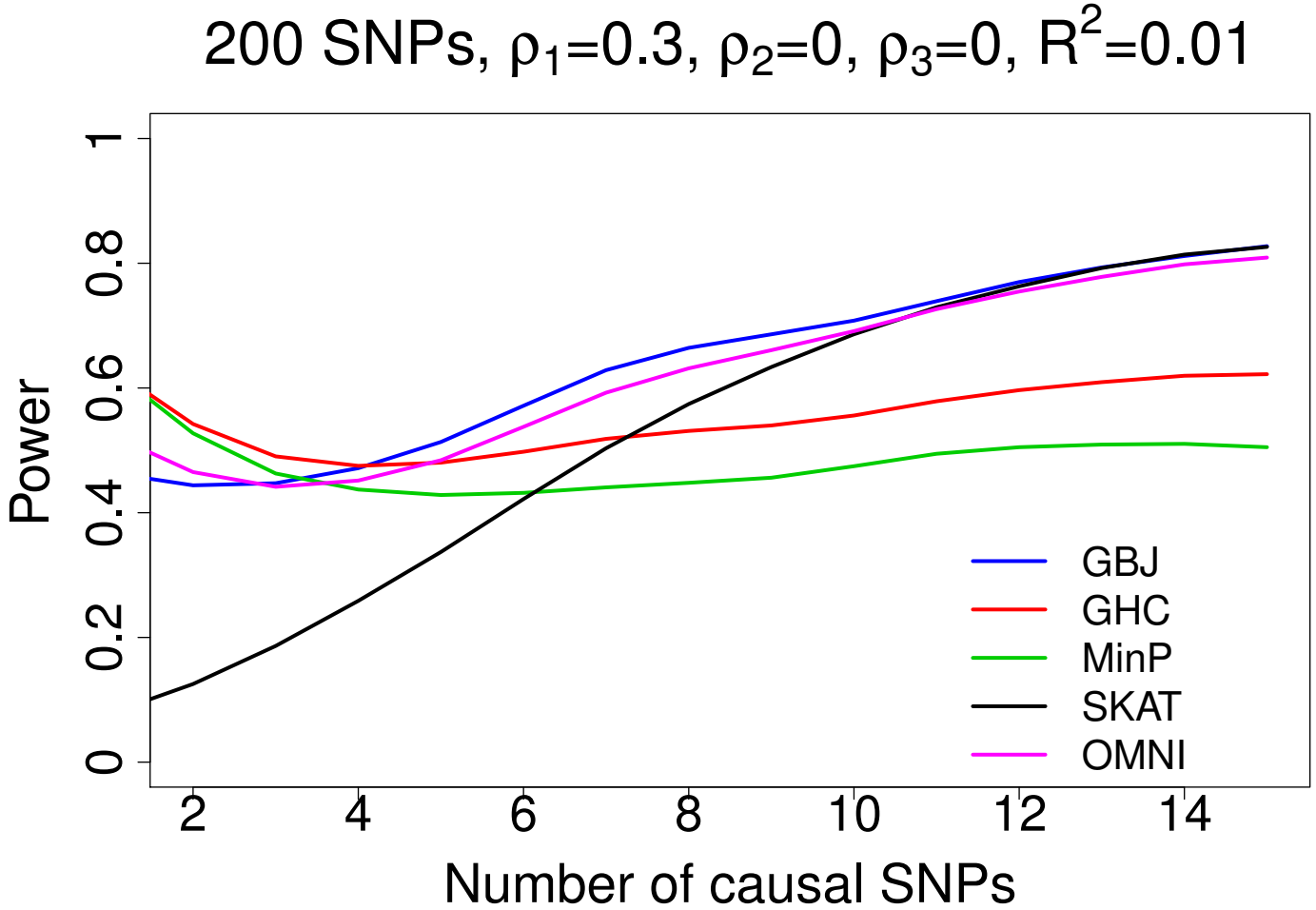
- The distribution of $S(t)$ is over-dispersed binomial and its exact distribution is hard to calculate.
- Approximate the distribution of $S(t)$ using extended beta-binomial.
- The sups in GHC and GBJ are achieved at the design points and both GHC/GBJ and their distributions are calculated analytically using approximations.

Rejection Boundary Comparisons: GHC vs GBJ



Note how we gain 'volume' in the rejection region near the expected signals.

Simulation (Main advantage of GBJ: Power gain in finite sample for moderate sparsity)



Extremely sparse regime: 1-3 causal. Moderately sparse regime: 4-13 causal. Dense regime: 14+ causal.

Sparse Regime: ACAT: Aggregated Cauchy Association Test

Yaowu Liu, et al (JASA 2018, AJHG, 2019)

Key features:

- A general method for combining p-values.
- Super fast computation under arbitrary correlation and robust to correlation.
- Powerful when signals are sparse.
- Can be used for constructing robust test.

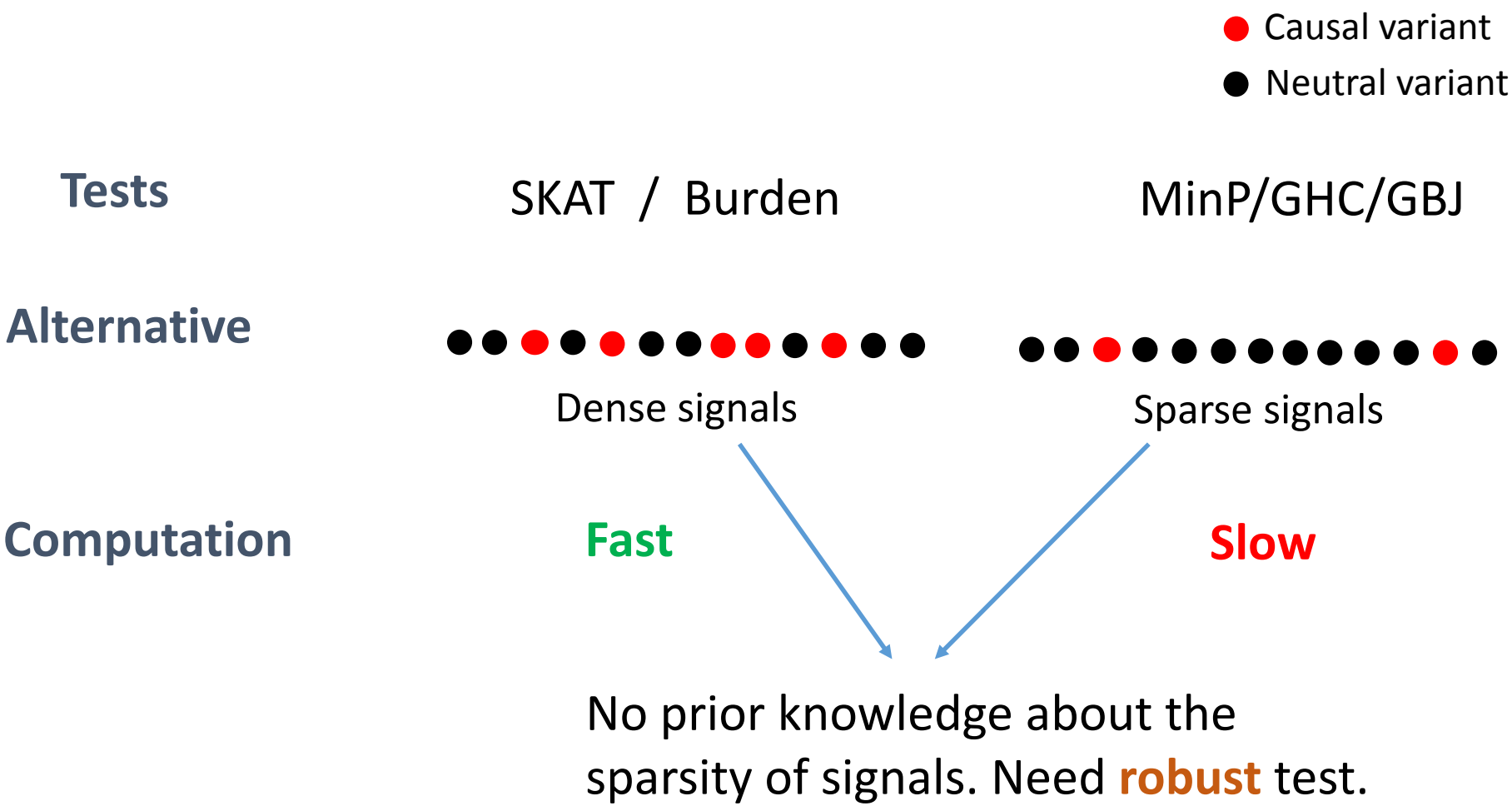
Aggregated Cauchy Association Test (ACAT)

Transform p-value to Cauchy

$$T_{ACAT} = \sum_{i=1}^{d} w_i \tan\{(0.5 - p_i)\pi\}$$

Weights

Existing SNV-set tests



Theory about ACAT

Assumptions: $I.\ p_i \longleftrightarrow |Z_i|$ (z-score) $\quad II.\ \forall i,j,(Z_i,Z_j) \sim N_2(0,\Theta_{ij})$

Theorem: For any $\Sigma \geq 0$, we have

$$\lim_{t\to+\infty} \frac{P\{T_{ACAT} > t\}}{P\{\text{Cauchy}(0,1) > t\}} = 1.$$

Tail is Cauchy

P-value calculation:

$$\text{p} - \text{value} \approx 1/2 - \{arctan(T_{ACAT})\}/\pi$$

Correlation of p-values
Not required

→ **Super fast**

Some insights

Sample mean $(\bar{X} = \frac{1}{d}\sum_{i=1}^{d} X_i)$	Perfectly dependent	Independent	General Dependency
$X_i \sim$ Cauchy(0,1)	$\bar{X} \sim$ Cauchy(0,1)	$\bar{X} \sim$ Cauchy(0,1)	$\approx$Cauchy(0,1)
$X_i \sim$ Normal(0,1)	$\bar{X} \sim$ N(0,1)	$\bar{X} \sim$ N(0,1/d)	

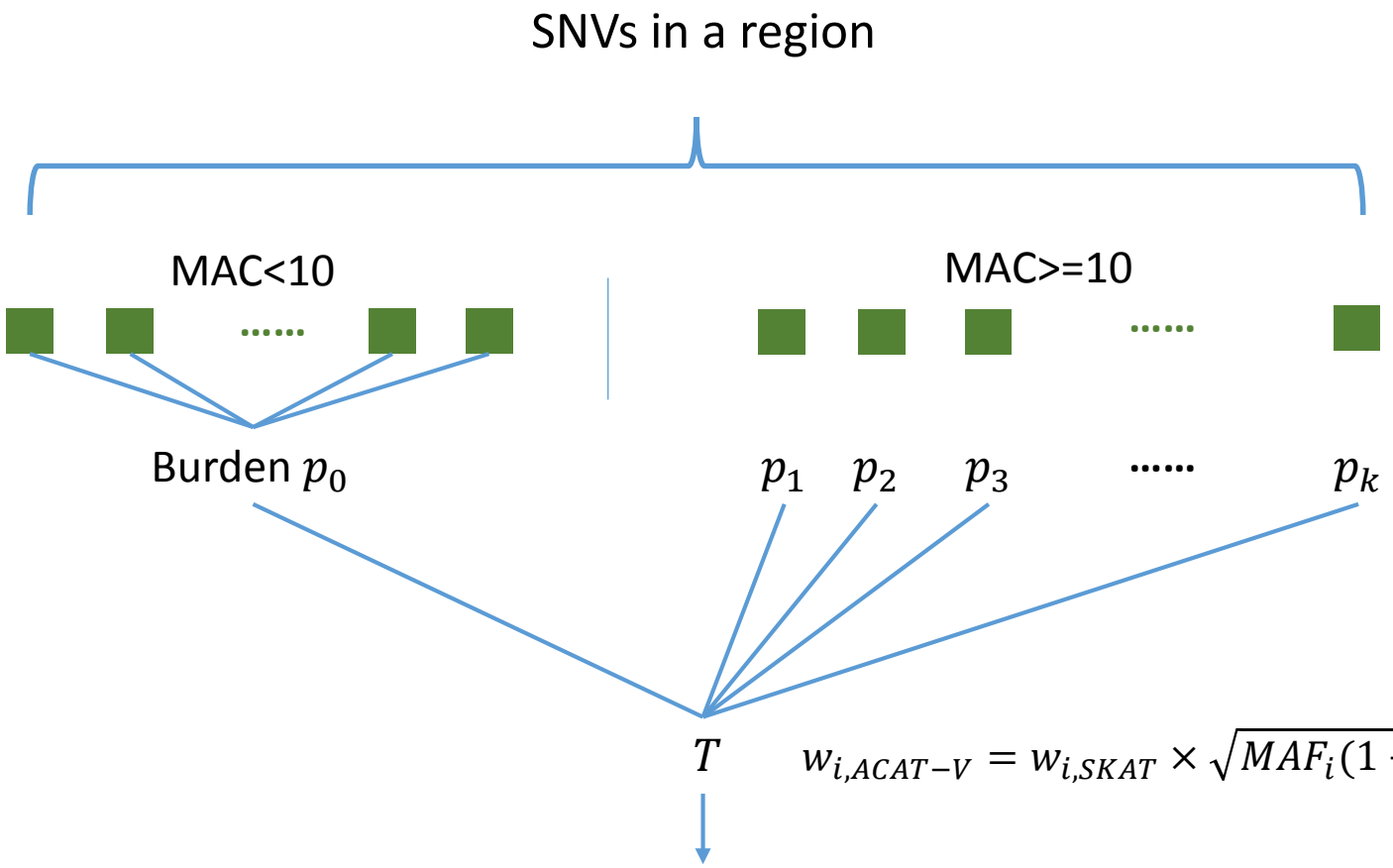
Heavy tail makes Cauchy distribution insensitive to correlation

ACAT is powerful against sparse alternatives

P-values	Cauchy values	
0.45	0.16	
0.35	0.51	
0.25	1.00	
0.15	1.96	**233**
0.05	6.31	
5e-03	63.7	
2e-03	159	

ACAT uses ***a few smallest p-values*** to represent the significance.

ACAT-V for testing a SNV-set



Saddlepoint method (Dey, et al, 2017)

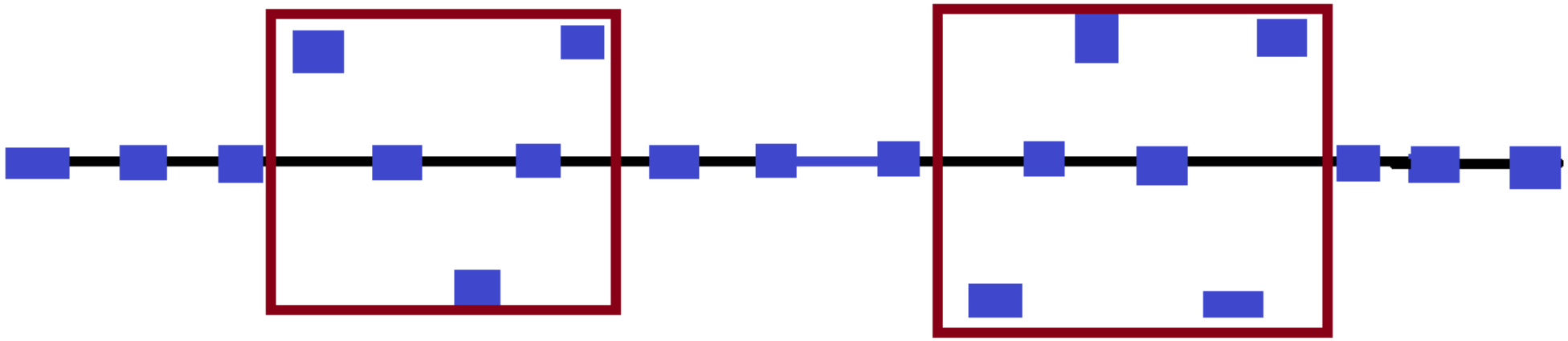
STAAR: variant-Set Test for Association using Annotation infoRmation

Xihao Li and Zilin Li

Key features:

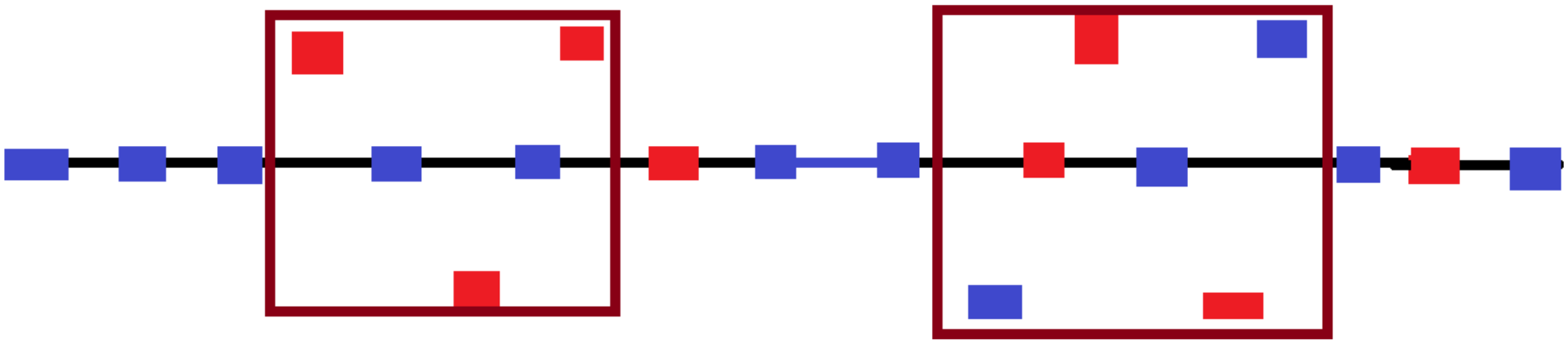
- Boost RV analysis power by optimally combining statistical evidence of MAFs (default in SKAT), functional annotations, and phenotypic information
- Computationally scalable
- Applicable to any given variant-set

Signal Regions (Effect Sizes (β)) in the Genome



Optimal weighting: True effect sizes (unknown)

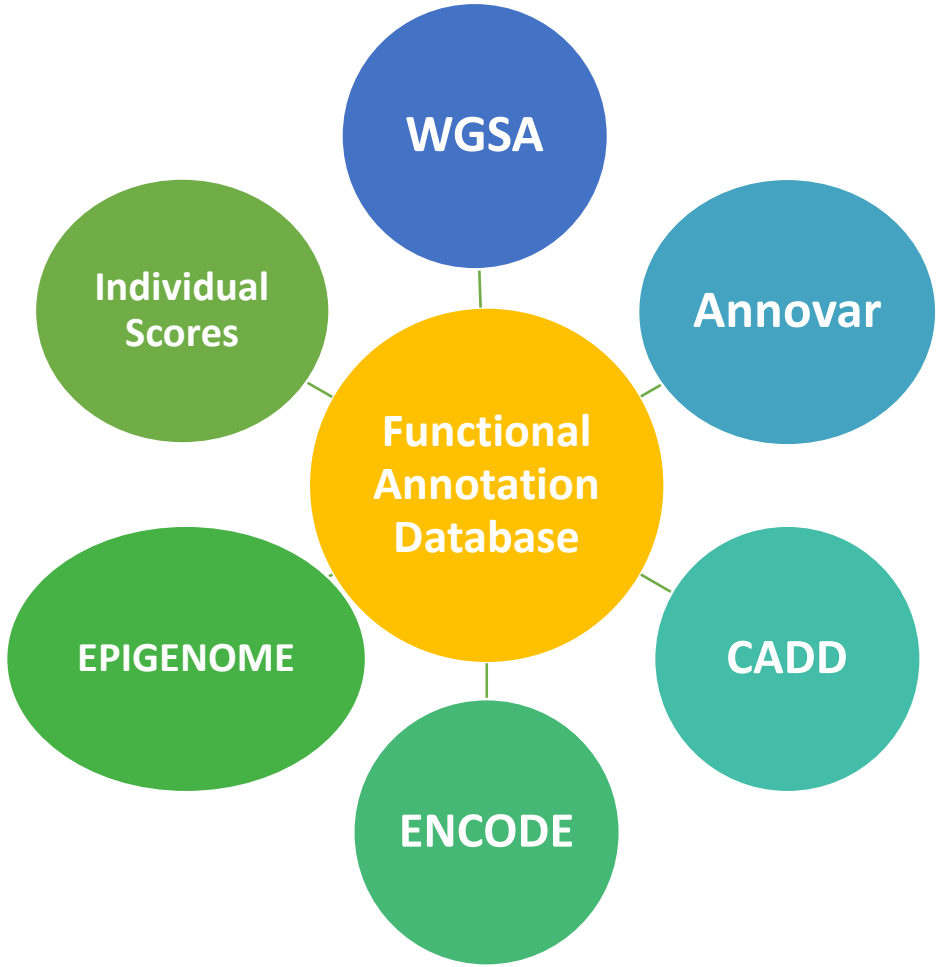
Use Functional Annotations to Prioritize Variants in a Variant-Set



Question: Which functional scores to use boost power of RV association analysis in a variant-Set

Choosing Weights w_j to Empower WGS Association Analysis

Genome Functional Variant Annotations (GSP+TOPMED) Hufeng Zhou)



>260 annotations

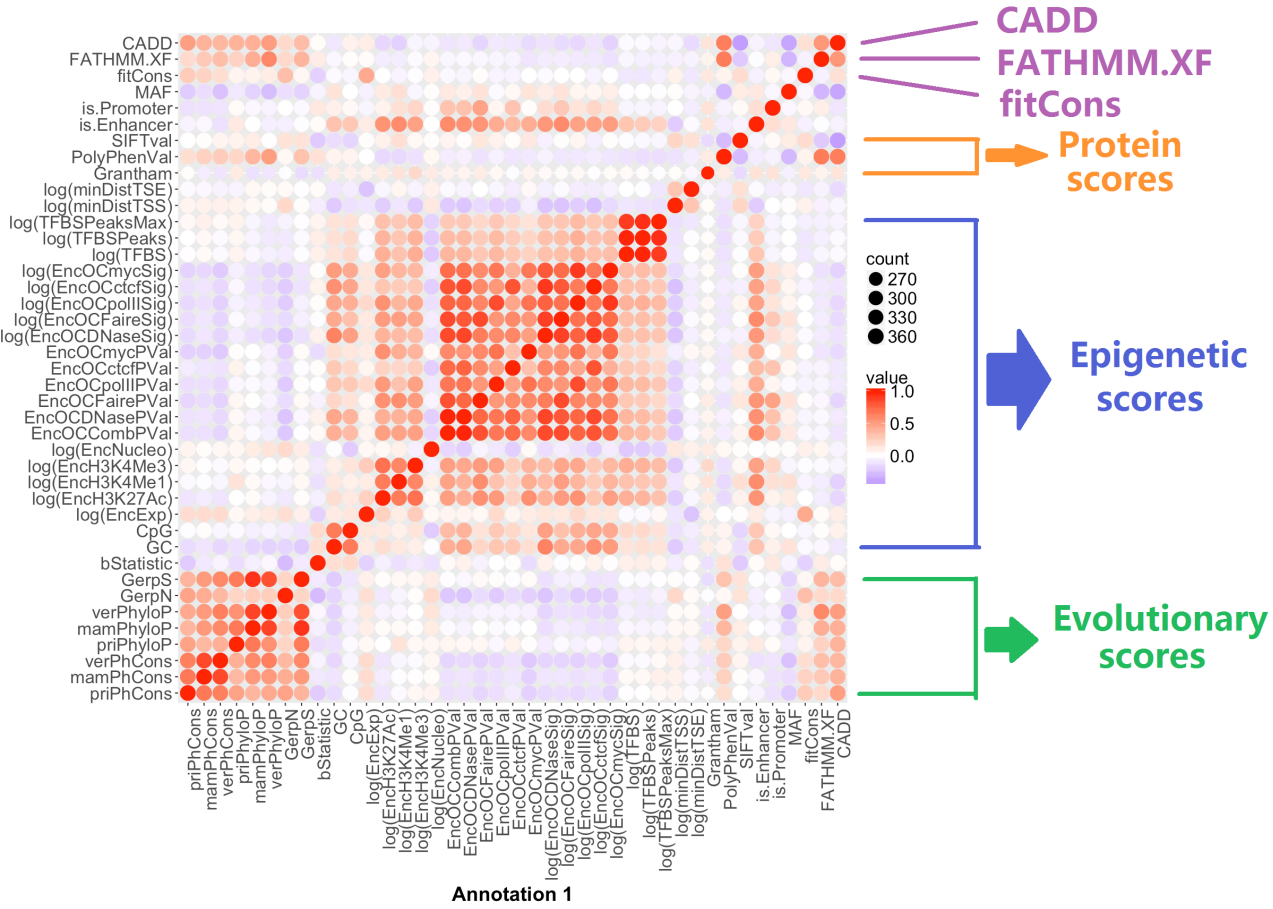
80% built on hg38

- Allele Frequency
- Conservation
- Protein Function
- Epigenetics
- Variant Effect Predictor
- MapAbility
- microRNA
- Molecular
- Local DNA Structure
- SNP database
- Clinical Variants
- 3D genomics
- eQTL
- Integrative
- Tissue-specific

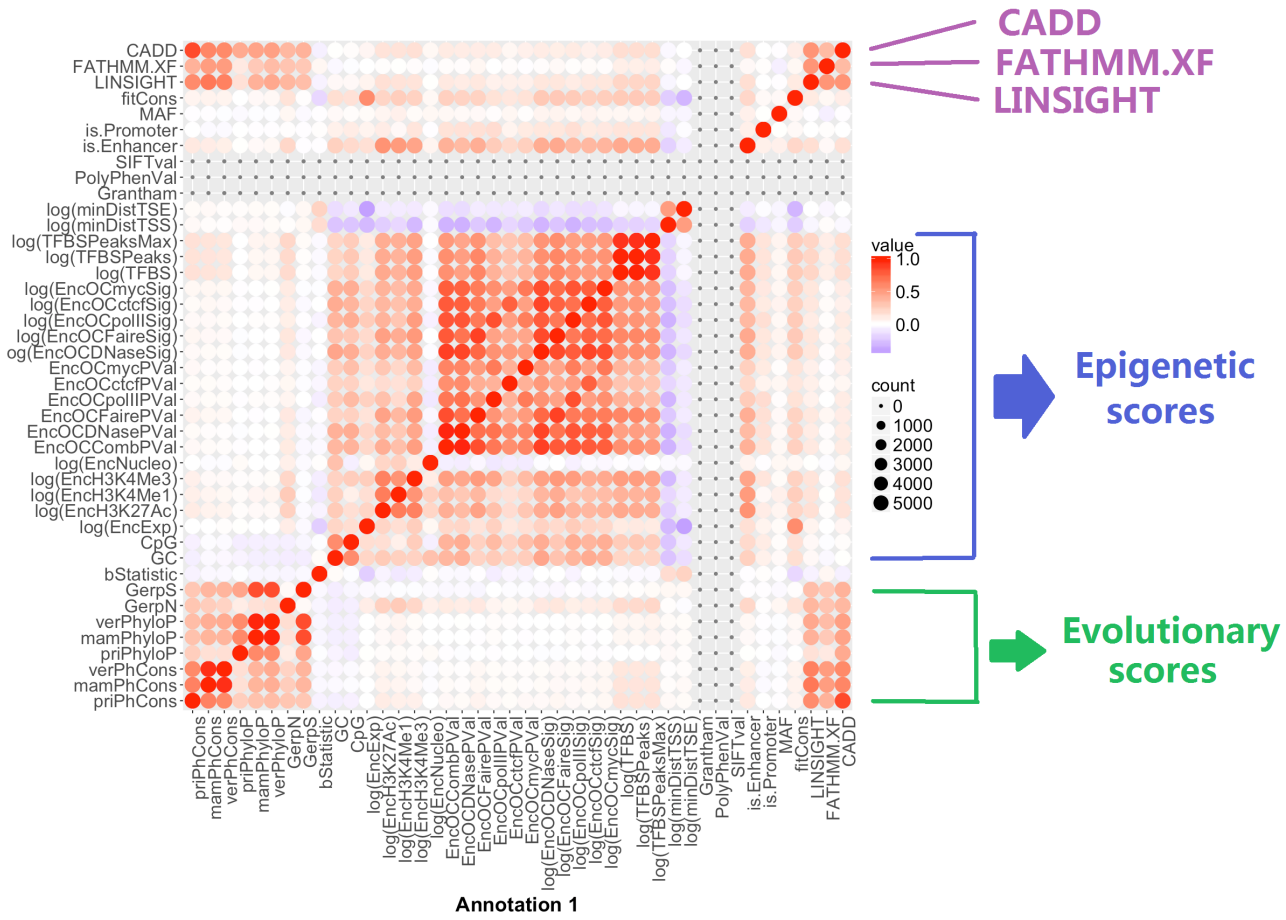
15 Types of Annotations

Dynamically incorporate multiple annotation weights in RV Tests

Existing Integrative Annotation Scores are Mainly Driven by Protein and Conservation Scores with Little Correlation with Epigenetic Scores



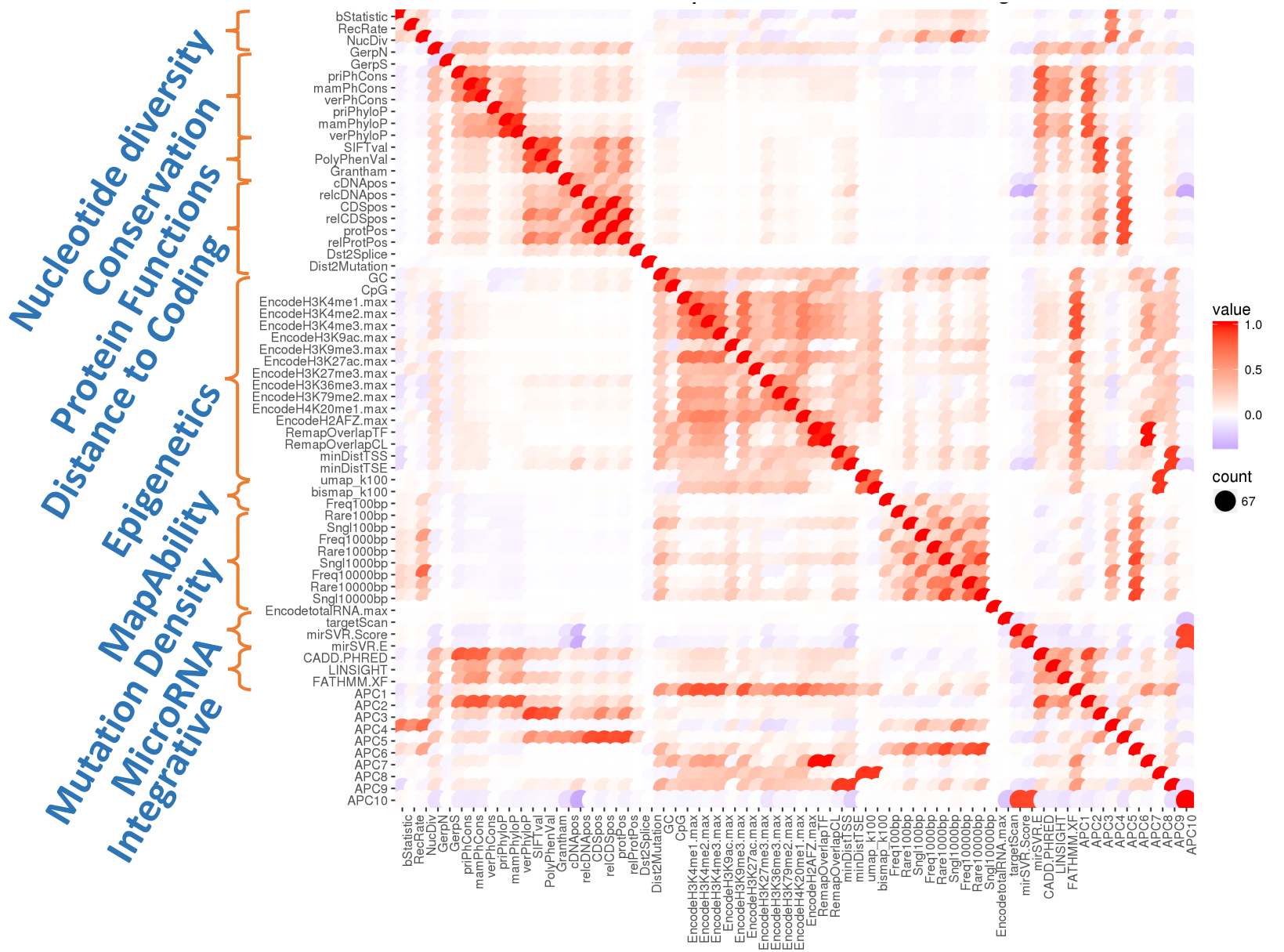
Coding Variants



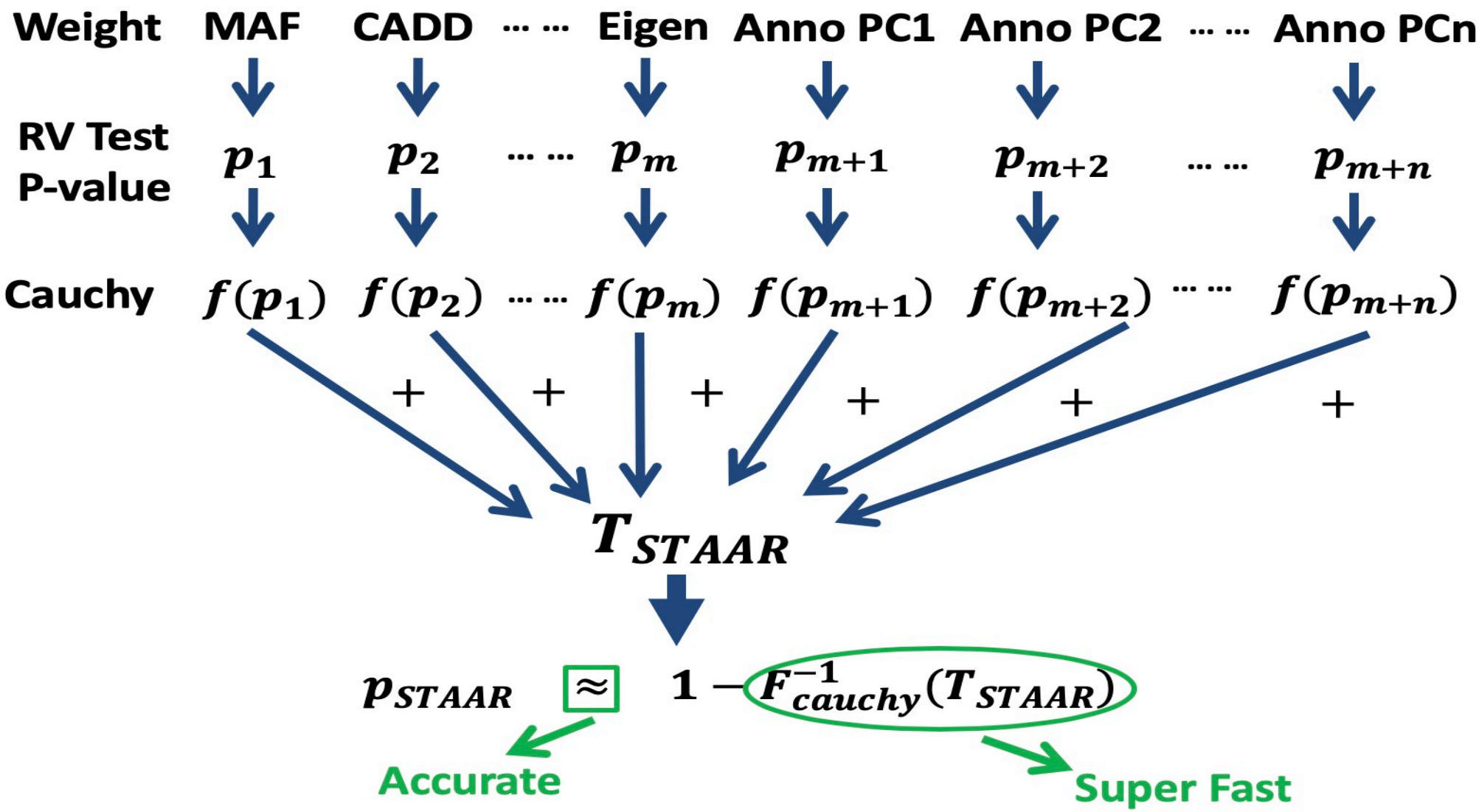
Non-coding Variants

Correlation Heatmap with Annotation PCs (GSP Freeze 1, hg38)

- APC1: Epigenetics
- APC2: Conservation
- APC3: Protein Function
- APC4: Negative Selection
- APC5: Distance to Coding
- APC6: Mutation Density
- APC7: Transcription Factor
- APC8: MapAbility
- APC9: Distance to TEE/TSE
- APC10: MicroRNA



STAAR: Incorporate Multiple Functional Scores to Boost Power of RV Association Analysis Using ACAT

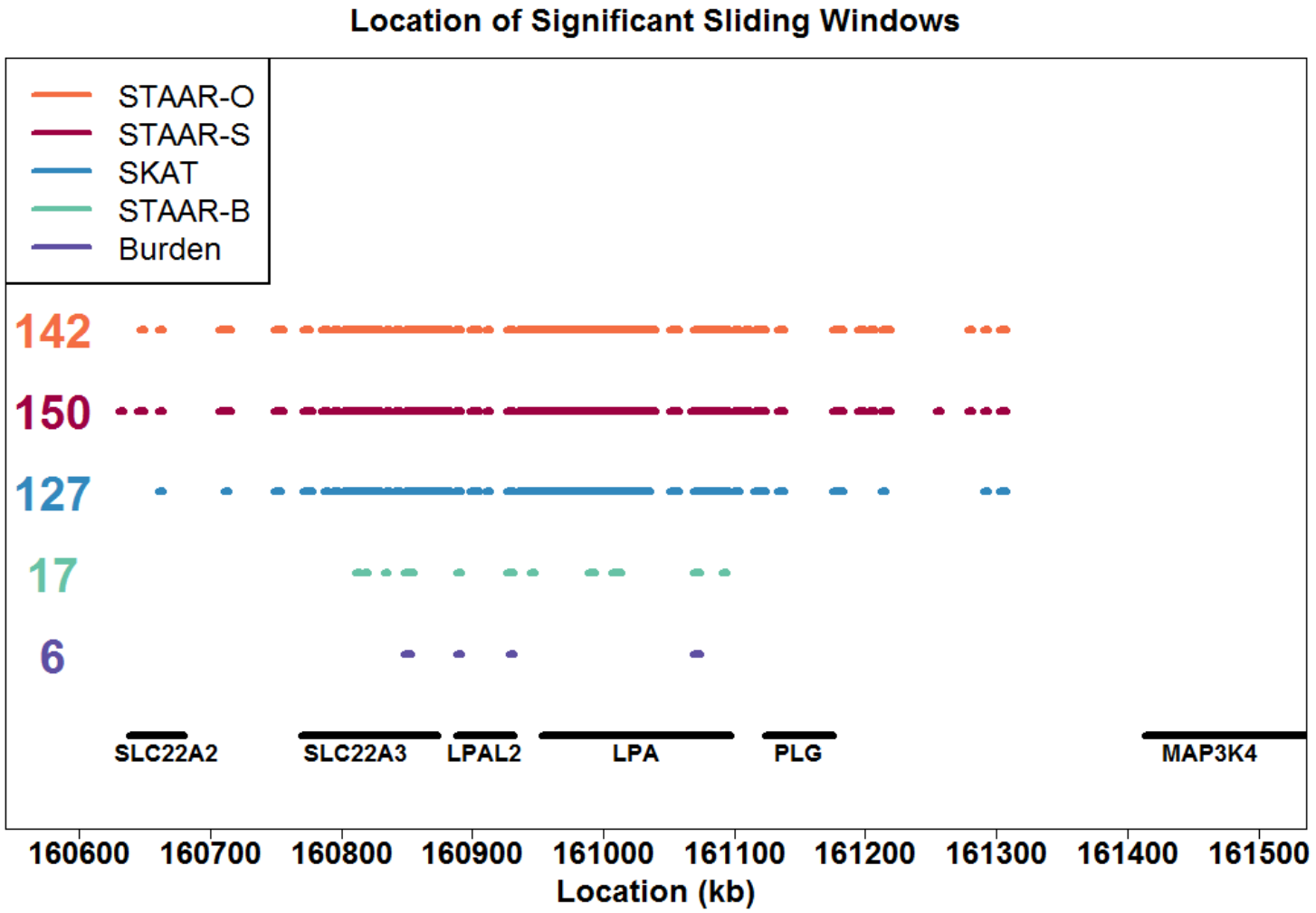


Type I Error Rates Using STAAR are Protected: Simulated WGS data Using COSI (n = 10,000)

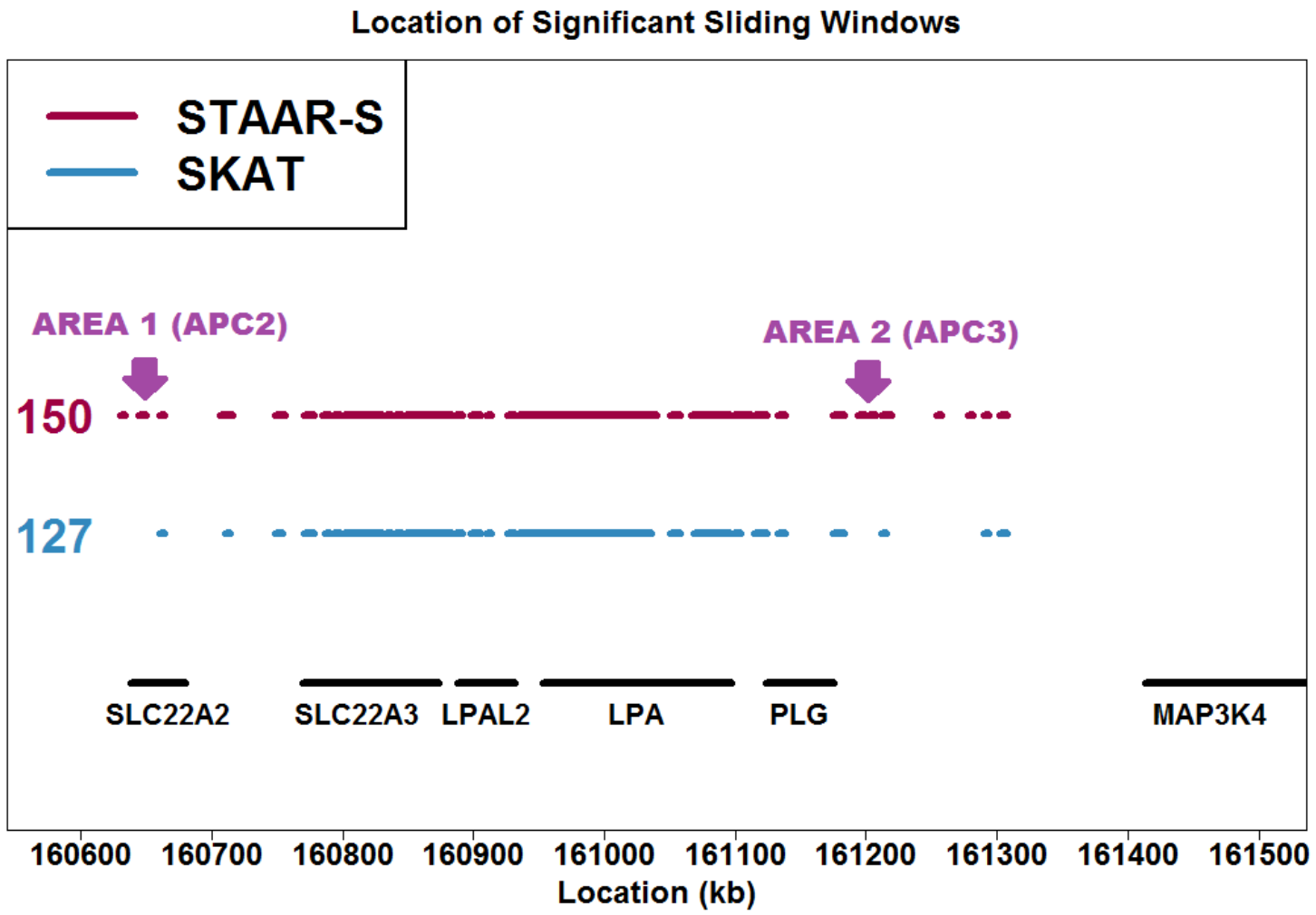
$\alpha = 10^{-6}$	Continuous Traits	Dichotomous Traits
STAAR-B	1.1×10^{-6}	1.0×10^{-6}
STAAR-S	9.9×10^{-7}	7.8×10^{-7}
STAAR-O	9.3×10^{-7}	1.0×10^{-6}

STAAR-O uses ACAT to combine STAAR-B and STAAR-O

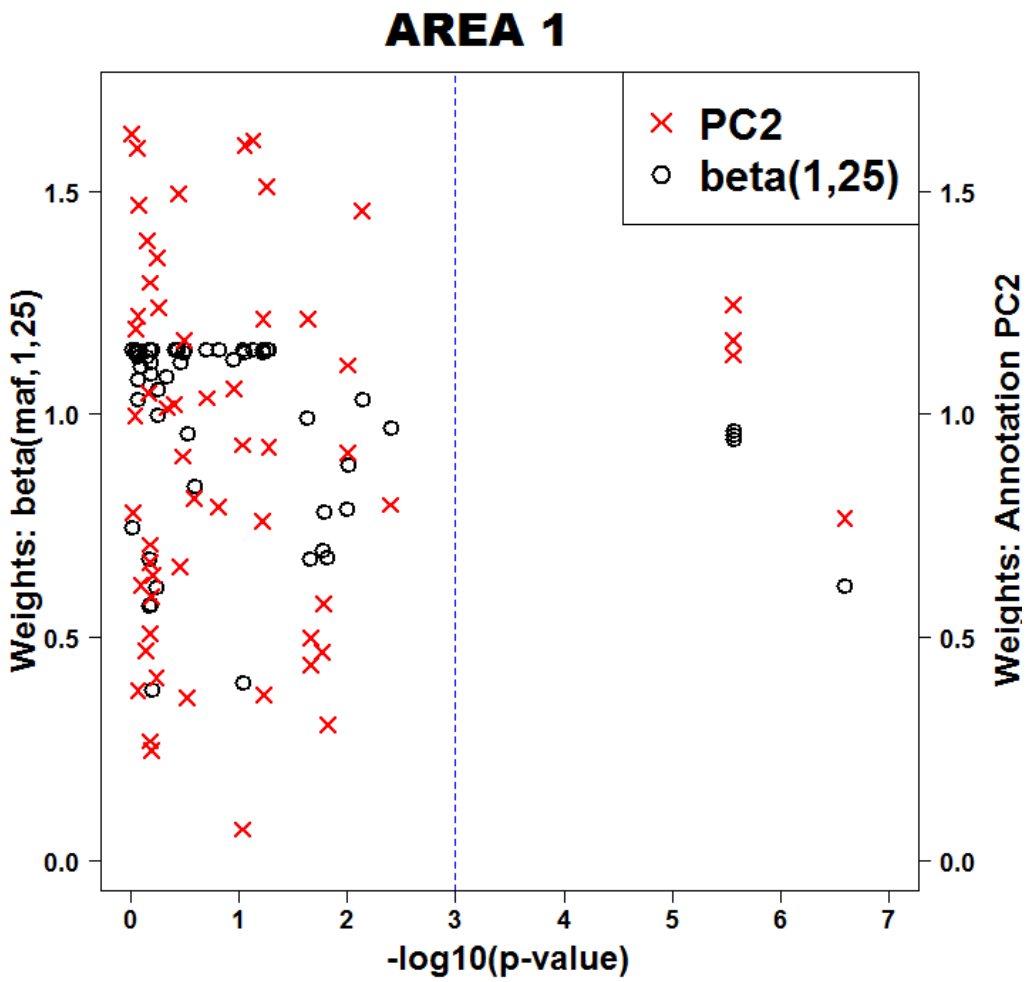
ARIC WGS data of LPA (AA, n=1800): Significant 4KB Sliding Windows in Chr 6

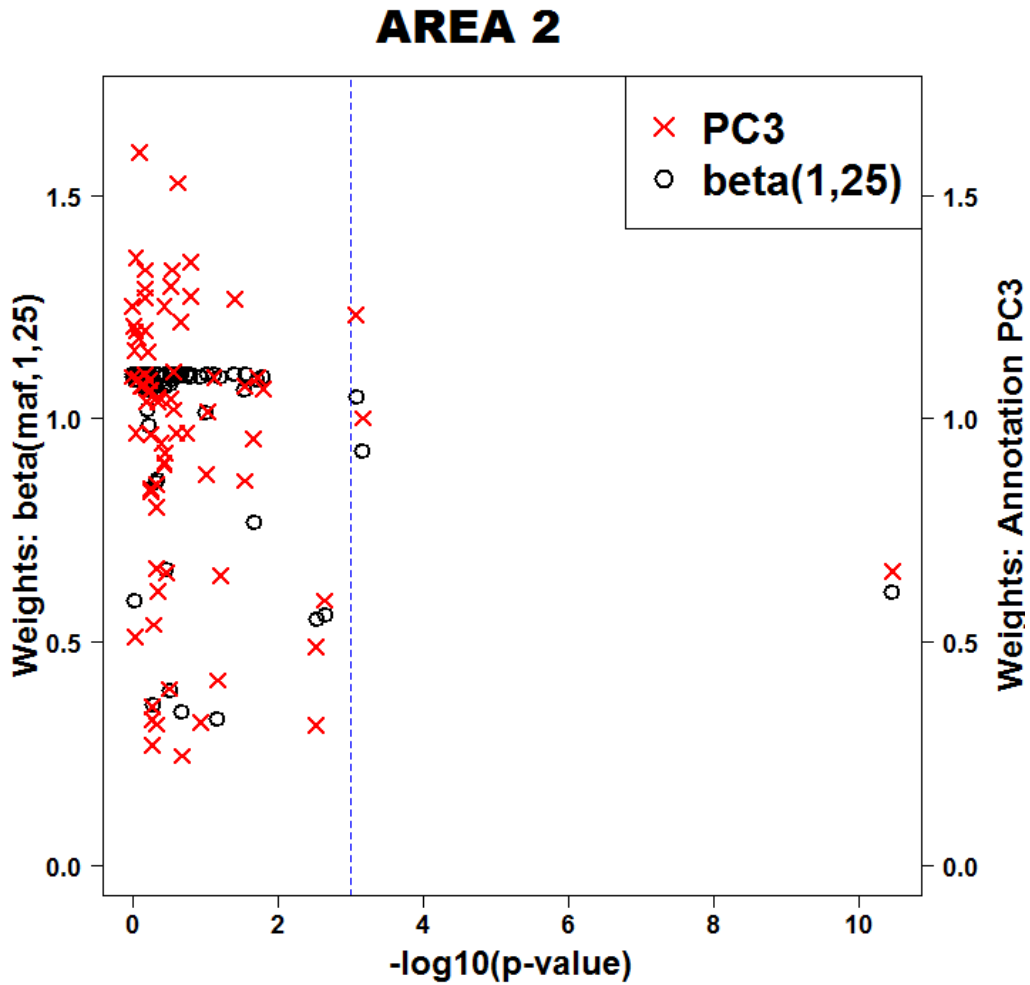


LPA (AA): Significant 4KB Sliding Windows in Chr 6

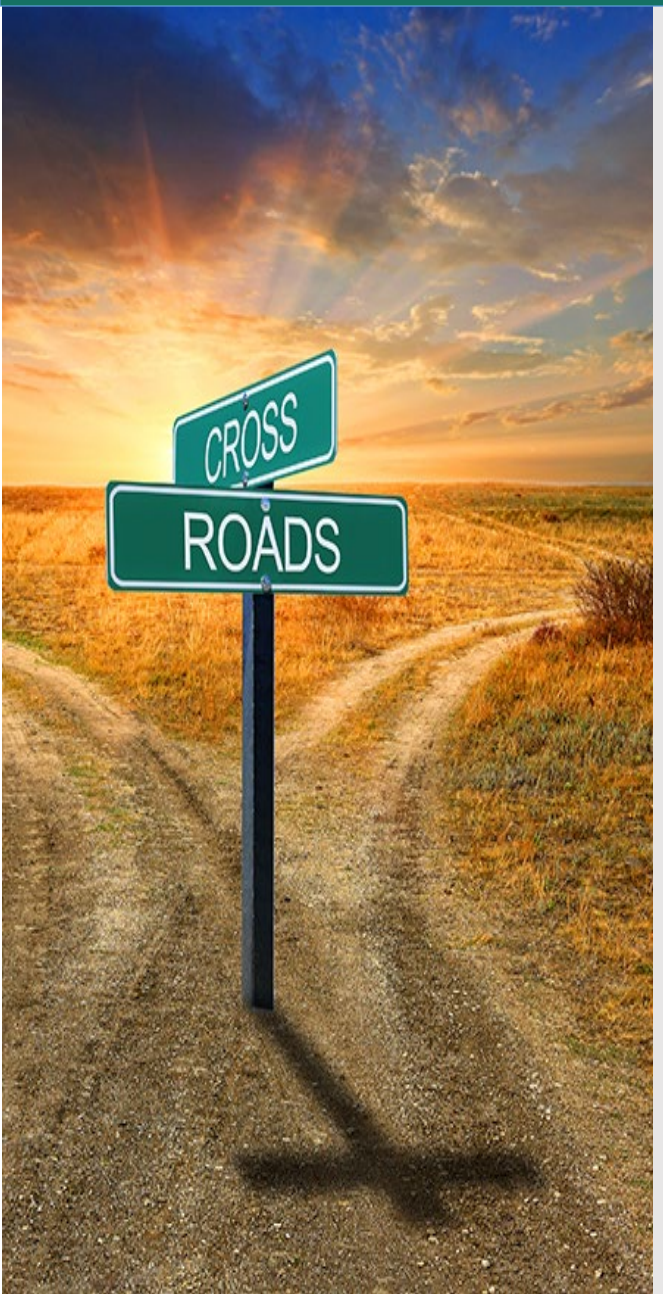


Area 1 and Area 2: Weights





Final Remarks



- Scalable statistical inference is a critical niche for analysis of big data.
- It is important to integrate domain science and computational science in scalable statistical inference to accelerate statistical science and scientific discovery.
- “Optimal” statistical inference needs to context-specific, e.g., dense and sparse regimes for high-dimensional hypothesis testing
- Asymptotic and finite sample results are both important.